

Wir evaluieren uns zu Tode
Möglichkeiten und Grenzen der
Bewertung von Online-Lernen.
Eine Meta-Evaluation.

Dissertation zur Erlangung des akademischen Grades einer

Doktorin der Philosophie
(Dr.in phil.)

an der Fakultät für Kultur- und Sozialwissenschaften
der FernUniversität in Hagen

vorgelegt von
Annabell Preußler

Betreuer
Prof. Dr. Peter Baumgartner

im Promotionsfach
Erziehungswissenschaft

Mit herzlichem Dank an meine Familie und meine Freundinnen und Freunde, ohne deren Unterstützung diese Arbeit nur unter erschwerten Bedingungen möglich gewesen wäre – insbesondere an meine Frau Melanie Preußler, die mich sowohl motiviert als auch, wenn nötig, abgelenkt hat, die mir mit ihrer liebevollen Geduld immer zur Seite stand und von der ich für jeden fertigen Arbeitsschritt ein Geschenk bekommen habe; an meine Eltern Sabine & Klaus Preußler und meine Schwester Esther Preußler, die nie gezweifelt haben; an meine Großeltern, die mich immer mit Doktorhut sehen wollten.

Ich bedanke mich darüber hinaus herzlich bei Anne Ott und Kirsten Willeke für ihre großartige Unterstützung und viele Jahre wunderbarer Freundschaft selbst in schwierigen Zeiten und bei Verena Schmidt, von der ich so vieles gelernt habe und die zu kennen mir viel bedeutet. Danke auch an Kathrin Fußangel, Dr. Martin Bensen und Andreas Büchter – die mir viel Zuversicht und Ermutigung zugesprochen haben und die mich immer wieder zum Lachen bringen können. Diana Köhne danke ich überdies für den herrlichen Zeitvertreib im Web 2.0!

Ein herzlicher Dank richtet sich auch an meine lieben (teilweise ehemaligen) Kolleginnen und Kollegen an der FernUniversität in Hagen: PD Dr. Thomas Brüsemeister, Dr. Klaus-Dieter Eubel, Nicole Barban, Sven Becker, Sarah Beerheide und Saliha Saou-Dreisewerd für die große Hilfe und Entlastung am Lehrgebiet sowie Leif Pullich, Marco Kalz, Viola Naust-Schulz, Susanne Neumann, Christian Glameyer und Marc Jelitto für die Diskussionen zwischen Tür und Angel. Besonders danken möchte ich Prof. Dr. Peter Baumgartner für die vortreffliche Betreuung und Beratung, Prof. Dr. Claudia de Witt und Prof. Dr. Theo Bastiaens für ihre freundliche Unterstützung sowie Prof. Dr. Georg Hansen für seinen unermüdlichen Einsatz. Weiterer Dank geht an meine MitstreiterInnen Katja Caspari, Katrin Högåve, Silke Kleindienst, Antje Neuhoff, Ute von Oertzen Becker, Heidi Weber und Bernhard Zech für anregende Debatten und hilfreiche Kritik.

Nicht zuletzt möchte ich mich bei Marco Kalz, Viola Naust-Schulz, Anne Ott, Melanie Preußler, Dieter Seeger und Kirsten Willeke für die große Hilfe bei der Korrektur bedanken.

Ich danke Euch allen dafür, dass Ihr mir vertraut habt und – Gott sei Dank – kann auch ich vertrauen.

„The only way [...] [we] could be fairly awarded is to have everyone in a given category play the same part in front of the same judges.“

(Humphrey Bogart über die Academy Awards, zit. nach Scarvia Anderson & Samuel Ball 1978:190)

Inhaltsübersicht

1	Einleitung	9
1.1	Problemaufriss und Zielsetzung	10
2	State of the Art – Online-Lernen.....	12
2.1	Mediengestütztes Lernen: ein kurzer historischer Abriss.....	12
2.2	Lernen in der Wissensgesellschaft.....	17
2.3	Die Kozma-Clark-Debatte.....	21
2.4	Metaanalyse im Bereich des mediengestützten Lernens	23
3	Evaluation – eine Annäherung.....	26
3.1	Begriffsklärung	26
3.2	Typisierung und Ablauf von Evaluationen	29
3.3	Vergleichendes Evaluationsdesign	32
4	Lernqualität und Lernerfolg	34
4.1	Das Konstrukt Qualität	34
4.2	Lernerfolg als Indikator für (Lern-)Qualität	35
4.3	Lernerfolg – Definitionsversuche.....	37
4.4	Dimensionen von Lernerfolg.....	40
4.5	Lernziele als Faktor.....	44
5	Lernerfolg und Leistungsmessung in Evaluationen und Vergleichsstudien	46
6	Taxonomie der Lernziele nach Lorin W. Anderson und David R. Krathwohl (2001).....	51
6.1	Struktur der Taxonomie.....	52
6.2	Praktische Relevanz der Taxonomie zur Erfassung von Lernzielen und Lernerfolg	55
6.3	Aufgabentypen und Lernziele.....	58
6.4	Zusammenfassung	62
7	Meta-Evaluation und Metaanalyse.....	65
7.1	Metaanalyse	65
7.2	Meta-Evaluation.....	68
8	Meta-evaluative Untersuchung	75
8.1	Fragestellung und Hypothesen	75
8.2	Methodisches Vorgehen.....	77
8.3	Bewertung der Primärstudien	96
8.4	Vergleich der Primärstudien und Integration der Ergebnisse	165
8.5	Darstellung der Ergebnisse in Bezug auf die Forschungsfragen.....	171
9	Fazit.....	177
9.1	Schlussfolgerungen	181
9.2	Empfehlungen	182
10	Literatur	185
	Anhang	I

Inhaltsverzeichnis

1	Einleitung	9
1.1	Problemaufriss und Zielsetzung	10
2	State of the Art – Online-Lernen.....	12
2.1	Mediengestütztes Lernen: ein kurzer historischer Abriss.....	12
2.2	Lernen in der Wissensgesellschaft.....	17
2.3	Die Kozma-Clark-Debatte.....	21
2.4	Metaanalyse im Bereich des mediengestützten Lernens	23
3	Evaluation – eine Annäherung.....	26
3.1	Begriffsklärung	26
3.2	Typisierung und Ablauf von Evaluationen	29
3.3	Vergleichendes Evaluationsdesign	32
4	Lernqualität und Lernerfolg	34
4.1	Das Konstrukt Qualität	34
4.2	Lernerfolg als Indikator für (Lern-)Qualität	35
4.3	Lernerfolg – Definitionsversuche.....	37
4.4	Dimensionen von Lernerfolg.....	40
4.4.1	Modell der lernrelevanten Kontexte und Komponenten nach Sigmar-Olaf Tergan (2004)	40
4.4.2	(ökonomische) Effizienz mediengestützten Lernens	44
4.5	Lernziele als Faktor.....	44
5	Lernerfolg und Leistungsmessung in Evaluationen und Vergleichsstudien	46
6	Taxonomie der Lernziele nach Lorin W. Anderson und David R. Krathwohl (2001).....	51
6.1	Struktur der Taxonomie.....	52
6.2	Praktische Relevanz der Taxonomie zur Erfassung von Lernzielen und Lernerfolg.....	55
6.3	Aufgabentypen und Lernziele.....	58
6.3.1	Zuordnung von Aufgabentypen zu Lernzielen.....	60
6.4	Zusammenfassung	62
7	Meta-Evaluation und Metaanalyse.....	65
7.1	Metaanalyse	65
7.1.1	Probleme der Metaanalyse	67
7.2	Meta-Evaluation.....	68
7.2.1	Evaluationsmodelle	69
7.2.2	Evaluationstandards	71
8	Meta-evaluative Untersuchung	75
8.1	Fragestellung und Hypothesen	75
8.2	Methodisches Vorgehen.....	77
8.2.1	Modelle der Meta-Evaluation	77
8.2.2	Auswahl geeigneter Primärstudien	80
8.2.3	Auswahl und Festlegung geeigneter Kriterien.....	85
8.2.4	Erläuterung der einzelnen Kriterien	89
8.2.5	Operationalisierung der Kriterien	91
8.2.6	Codierung der Primärstudien.....	93

8.3	Bewertung der Primärstudien	96
8.3.1	P 1: Edward R. Jones et al. (1999)	96
8.3.2	P 2: Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr. (2001).....	102
8.3.3	P 3: Rob Martens, Stanley J. Portier & Martin Valcke (1995)	107
8.3.4	P 4: Josef Smolle, Reinhard Staber, Heide Neges, Gilbert Reibnegger & Helmut Kerl (2005)	115
8.3.5	P 5: Randall S. Davies & Robert Mendenhall (1998).....	120
8.3.6	P 6: Vassilis Hartzoulakis (2002).....	128
8.3.7	P 7: William J. Ryan (2002)	134
8.3.8	P 8: Thommy Eriksson, Alois Goller & Sergei Muchin (2001).....	140
8.3.9	P 9: Ruth Green & Karen Gentemann (2001)	146
8.3.10	P 10: David J. Pucel & Thomas F. Stertz (2005).....	152
8.3.11	P11: Larry S. Tinnerman (2006)	158
8.4	Vergleich der Primärstudien und Integration der Ergebnisse	165
8.4.1	Auswertung der Einzelkriterien.....	165
8.4.2	Analyse der bestbewerteten Primärstudien	170
8.5	Darstellung der Ergebnisse in Bezug auf die Forschungsfragen.....	171
9	Fazit.....	177
9.1	Schlussfolgerungen	181
9.2	Empfehlungen	182
10	Literatur	185
	Anhang	I
I	Beschreibung der Primärstudien	II
II	Bewertung der Primärstudien nach den übrigen Kriterien.....	XL
III	Übersicht über die Operationalisierung der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001).....	LXXXVIII

Abbildungsverzeichnis

Abbildung 1:	Formen von Evaluationen (angepasst nach Martin Bensen & Andreas Büchter 2005:5)	30
Abbildung 2:	Evaluationskreislauf (angepasst nach Hans-Günther Rolff 2001:93)	31
Abbildung 3:	Beispiel für ein Testsiegel auf einer Software (ibis acam GmbH 2004).	36
Abbildung 4:	Bildungsmedien als Angebote zur Anregung von Lernprozessen (nach Michael Kerres 2001).....	39
Abbildung 5:	Dimensionen von Lernerfolg (erstellt nach Sigmar-Olaf Tergan 2004:17)	41
Abbildung 6:	Lernrelevante Kontexte und Komponenten technologiebasierter Lernszenarien (Sigmar-Olaf Tergan 2004:18).....	41
Abbildung 7:	Karikatur „Zum Ziele einer gerechten Auslese lautet die Prüfungsaufgabe für Sie alle gleich: Klettern Sie auf den Baum!“ (Cartoon von Hans Traxler 1999:36/37)	48
Abbildung 8:	Ablauf der Taxierung einer Prüfungsaufgabe (Franziska Keller 2004:17).....	57
Abbildung 9:	Mindmap Aufgaben (nach Norbert Meder & Andrea Frick 2006:75)	59
Abbildung 10:	Kognitive Prozesse (nach David R. Krathwohl 2002:212)	60
Abbildung 11:	Kognitive Prozesse und Aufgabentypen	61

Tabellenverzeichnis

Tabelle 1:	Taxonomiematrix nach Lorin W. Anderson & David R. Krathwohl (übersetzt nach Kurt Schobel & Ulrike von Holdt, 2004)	53
Tabelle 2:	Modelle der Meta-Evaluation nach Thomas D. Cook & Charles L. Gruder (1978:17, nach der Übersetzung von Werner Wittmann 1985:404).....	70
Tabelle 3:	Übersicht über die Primärstudien	84
Tabelle 4:	Raster zur Bewertung der Kriterien in den Primärstudien in Anlehnung an Thomas Widmer (1996).....	95
Tabelle 5:	Einordnung von P 1 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)	100
Tabelle 6:	Bewertungsraster P1	101
Tabelle 7:	Bewertungsraster P2	106
Tabelle 8:	Einordnung von P 3 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)	113
Tabelle 9:	Bewertungsraster P3	114
Tabelle 10:	Einordnung von P 4 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)	118
Tabelle 11:	Bewertungsraster P4	119
Tabelle 12:	Einordnung von P 5 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)	127
Tabelle 13:	Bewertungsraster P5	127
Tabelle 14:	Einordnung von P 6 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)	132
Tabelle 15:	Bewertungsraster P6	133
Tabelle 16:	Einordnung von P 7 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)	138
Tabelle 17:	Bewertungsraster P7	139
Tabelle 18:	Einordnung von P 8 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)	144
Tabelle 19:	Bewertungsraster P8	145

Tabelle 20:	Einordnung von P 9 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)	150
Tabelle 21:	Bewertungsraster P9	151
Tabelle 22:	Einordnung von P 10 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)	156
Tabelle 23:	Bewertungsraster P10	157
Tabelle 24:	Einordnung von P 11 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)	163
Tabelle 25:	Bewertungsraster P11	164
Tabelle 26:	Zusammenfassung der Einzelkriterien	165
Tabelle 27:	überdurchschnittlich bewertete Primärstudien (Sortierung nach Bewertung)	170
Tabelle 28:	Synoptische Darstellung der Primärstudien.....	173
Tabelle 29:	Beispiel zur Einordnung einer optimalen Studie in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001).....	183

1 Einleitung

„Wir amüsieren uns zu Tode“ schreibt Neil Postman (1985) und bezieht sich damit auf die Urteilsbildung im Zeitalter der Unterhaltungsindustrie. „Problematisch am Fernsehen ist nicht, daß es uns unterhaltsame Themen präsentiert, problematisch ist, daß es jedes Thema als Unterhaltung präsentiert“ (a.a.O.:110).

„Wir evaluieren uns zu Tode“ wage ich zu behaupten und beziehe mich damit auf die Flut von Evaluationsstudien, die in vielfältiger Form, mit vielfältigen Ansprüchen und Definitionsgrundlagen und zu vielfältigen Themen durchgeführt werden.

Problematisch an der Evaluation ist nicht, dass damit der Erfolg bestimmter Maßnahmen gemessen werden soll, problematisch ist, dass die Ergebnisse als Erfolg oder Misserfolg der Maßnahme per se interpretiert werden. Da wir jedoch mit einer Flut von Studien überschüttet werden, sind deren Ergebnisse nicht mehr sicher zu deuten. Thomas L. Russel (1999)¹ listet in „no significant difference“ mehrere hundert Forschungsberichte, Zusammenfassungen, Tagungsbeiträge und andere wissenschaftlichen Publikationen auf, die sich mit der Frage des Lernerfolgs² von Studierenden in Verbindung mit verschiedenen Darbietungsmodi (Präsenz, Online, Blended) befassen. Das nicht überraschende Ergebnis der Sichtung dieser Sammlung ist: Ein Teil der dort zu findenden Arbeiten stellt keinen signifikanten Unterschied zwischen den Gruppen fest, ein anderer Teil sieht einen höheren Lernerfolg bei den Gruppen, die mit dem Computer arbeiten. Wiederum andere AutorInnen³ stellen fest, dass traditionell lernende Klassen besser abschneiden und manche Studien kommen zu gemischten Ergebnissen. Eine verlässliche Aussage kann also nicht getroffen werden – mehr noch – am Ende ist noch nicht einmal klar, was unter Lernerfolg verstanden wird, denn dieser wird in verschiedenen Studien verschieden definiert.

In der Forschungsphase ihrer Dissertation zum Thema „Evaluation als modernes Ritual“ bekam Christine Schwarz (2004) von einem Interviewpartner folgendes zu hören:

„Das darf doch nicht wahr sein, jetzt schreiben Sie über die Evaluation der Evaluation! Ja sind denn jetzt alle verrückt geworden? Haben Sie denn wirklich nichts Besseres vor, als über

¹ Als Ergänzung zu Thomas L. Russel (1999) entstand die Website <http://www.nosignificantdifference.org>, letzter Aufruf 21.06.2007.

² Der Begriff „Study outcomes“ wird mit Lernerfolg übersetzt (vgl. <http://dict.leo.org/forum/viewUnsolvedquery.php?idThread=214075&idForum=2&lp=ende&lang=de> letzter Aufruf 24.11.2006).

³ Anmerkungen zu geschlechtsneutraler Schreibweise: Sind Frauen und Männer gleichermaßen gemeint, werden entweder beide Geschlechter genannt, Passivformulierungen oder die „Binnen-I“-Schreibung verwendet. Bei der Angabe von Zitaten wurden, wenn möglich, jeweils die Vor- und Nachnamen der AutorInnen angegeben, um im Zuge von Gender Mainstreaming darauf hinzuweisen, ob der entsprechende Beitrag von einer Frau oder einem Mann verfasst worden ist (vgl. Annabell Preussler 2004).

die sowieso schon Zeit vergeudende Evaluation nun auch noch eine Studie zu schreiben?“ (a.a.O.:ii).

Auch ich möchte in dieser Arbeit an diesem Punkt anknüpfen und mich der Evaluation der Evaluation widmen und die Möglichkeiten und Grenzen im Bereich der Lernerfolgsmessung beim E-Learning aufzeigen.

1.1 Problemaufriss und Zielsetzung

Mediengestütztes Lernen bzw. E-Learning sind derzeit uneingeschränkt up to date. Alles, was „e“ ist – also E-Learning, E-Education, E-Inclusion, E-Government usw. scheint sich derzeit offenbar gut vermarkten zu lassen.

Doch wo liegen die Vorteile von E-Learning? Warum soll es besser sein als traditionelles Lernen? Dies sind Fragen, die immer wieder auftauchen und die man mit wissenschaftlichen Studien zu belegen versucht. E-Learning sei einfacher, verständlicher, knüpfe an den Erfahrungshorizont der Lernenden an, sei realitätsnäher, kostengünstiger, motivierender, schlichtweg effektiver als traditionelles Lernen. Ist es das tatsächlich und wenn ja, wie ist das messbar? Wird durch E-Learning höherer Lernerfolg erzielt als mit traditionellem Lernen?

Viele Studien zielen genau auf diesen Vergleich von Online- versus Präsenzlernen ab. Doch ist solch ein Vergleich sinnvoll? Die im Folgenden beschriebene Dissertation soll einen Beitrag leisten zur Evaluationsforschung im Bereich des mediengestützten Lernens. Sie soll Studien, die sich mit dem Zusammenhang von E-Learning und Lernerfolg bzw. dem Einfluss von E-Learning auf den Lernerfolg beschäftigen, analysieren und vergleichen. Die Arbeit ist somit als Methodenkritik anzusehen und gehört damit dem empirisch-methodischen Bereich an. Das Forschungsinteresse geht der Vermutung nach, dass Lernerfolg nicht eindeutig operationalisierbar ist und unspezifische Vergleiche von Online- versus Präsenzlernen nicht uneingeschränkt sinnvoll anwendbar sind.

Dies führt dazu, dass auch die Methode der Metaanalyse nur bedingt (da nur für bestimmte Aspekte) geeignet ist, um generalisierbare Rückschlüsse auf den Lernerfolg zuzulassen. Indifferenzierte Sichtweisen bezüglich einer Definition von (Lern-)Erfolg machen Vergleiche bzw. Zusammenfassungen einzelner Studien quasi unmöglich. Unumgänglich ist ein kritisches Hinterfragen des Begriffs „Lernerfolg“. Was wird überhaupt unter Lernerfolg verstanden und wie wird er gemessen? Subsummieren unterschiedliche Studien unterschiedliche Dinge in diesem Begriff?

Am Beispiel der Lernzieltaxonomie nach Lorin W. Anderson und David R. Krathwohl (2001) soll dieses Problem aufgegriffen werden. Voraussetzung dafür ist die theoretische Diskussion der Konstrukte Lernerfolg und Lernqualität, die auf die Bereiche E-Learning und traditionelles Lernen übertragen werden. Wie kann Lernerfolg überhaupt evaluiert werden? Mit der Durchführung einer Meta-Evaluation sollen Bereiche bzw. Dimensionen des Konstrukts Lernerfolg identifiziert werden,

die sich tatsächlich miteinander vergleichen lassen. Die Taxonomie scheint für die Untersuchung, ob die in einem Lernsetting intendierten Lernziele auch der Lernerfolgskontrolle entsprechen, geeignet zu sein.

In Kapitel 2 werde ich den aktuellen Forschungsstand zur Thematik des Online-Lernens darstellen. Dabei werde ich einerseits auf historische Aspekte eingehen, andererseits aber auch eine Beschreibung des Lernens in der Wissensgesellschaft skizzieren. Wesentlich scheint mir dabei die Debatte um die Frage, ob Medien das Lernen überhaupt beeinflussen, zwischen Richard E. Clark und Robert B. Kozma zu sein. Darüber hinaus soll eine kurze Übersicht zu metaanalytischen Untersuchungen in diesem Bereich gegeben werden.

In Kapitel 3 möchte ich auf das Verfahren der Evaluation zu sprechen kommen und dabei Begriffsklärungen und Methoden darstellen.

Eine Beschreibung und Erläuterung der Taxonomie der Lernziele nach Lorin W. Anderson und David R. Krathwohl wird in Kapitel 4 gegeben, bevor ich in Kapitel 5 auf die Lernqualität eingehen werde. Darin soll der Versuch unternommen werden, die Dimensionen von Lernerfolg sowie den Zusammenhang von Lernerfolg und Lernqualität darzustellen.

Kapitel 6 ist als Zwischenfazit zu sehen, das die bisherige Argumentation aufgreift und mit der Problematik der Messung von Lernqualität und Lernerfolg in Zusammenhang bringt.

In Kapitel 7 möchte ich die Verfahren von Meta-Evaluation und Meta-Analyse beschreiben, um meine eigene Untersuchung im praktischen Teil ab Kapitel 8 rechtfertigen zu können.

Kapitel 9 stellt sowohl Resumée als auch Ausblick dar, da die gefundenen Ergebnisse einerseits diskutiert werden, andererseits aber auch Empfehlungen für Evaluationen gegeben werden.

2 State of the Art – Online-Lernen

Obwohl ich mich dagegen sträube, die Forschungslage als „State“⁴ zu bezeichnen, denn der Forschungsprozess sollte – um Innovationen hervorzubringen – notwendigerweise nicht stillstehen, möchte ich nicht darauf verzichten, an dieser Stelle einen Überblick über die derzeitige Forschungslage zu geben. Beginnen möchte ich zunächst mit einem kurzen Abriss der Entwicklung mediengestützten Lernens. Anschließend werde ich versuchen, die bisherigen Evaluationsbestrebungen in diesem Bereich zusammenfassen.

2.1 Mediengestütztes Lernen: ein kurzer historischer Abriss

In der Erziehungswissenschaft ist das Lernen mit Medien schon seit langem ein wichtiges Thema, welches allerdings etlichen Wandeln und Paradigmenwechseln unterworfen war. Im Laufe der Auseinandersetzung mit medienpädagogischen Fragestellungen wurden dabei verschiedene erkenntnistheoretische Argumentationen zu Grunde gelegt (vgl. Peter Strittmatter & Helmut M. Niegemann 2000). Dazu gehören beispielsweise die „kognitive Wende“⁵ in den 1960er Jahren sowie die Konstruktivismus-Objektivismus-Debatte in den 1980ern (vgl. Patrick Blumschein 2004:2)

Bereits in den 1960er Jahren wurden Computer unter dem Sammelbegriff „computerunterstützte Lehre“ zu Lehrzwecken eingesetzt. Dabei handelte es sich um Großrechenanlagen, auf denen einfache Programme installiert werden konnten. Ihre Anschaffung war allerdings nur für Universitäten und nicht für Schulen lohnend. So wurde die von Konrad Zuse entwickelte Anlage „Zuse Z22“ beispielsweise 1958 an der Fachhochschule Karlsruhe in Betrieb genommen (vgl. Thomas Bartos 2004:1). In den 1960er Jahren zählten Rechenanlagen dieser Art bereits zur Standardausstattung an Hochschulen (vgl. ebd.).

Ebenfalls in den 1960er Jahren stellte der zu dieser Zeit eingeführte programmierte Unterricht eine der ersten Formen mediengestützten Lernens an Schulen dar, allerdings stand er zunächst unter dem Aspekt der Behebung des Mangels an Lehrkräften durch den Einsatz von Lernprogrammen (vgl. Annabell Preussler 2002:41 ff). Diese Lernform ist zurückzuführen auf die Idee der „programmierten Unterweisung“, die

⁴ oder auch „Stand“ der Forschung

⁵ Als „kognitive Wende“ wurde die Verdrängung des Behaviorismus durch den Ansatz der Informationsverarbeitung (kognitive/ kognitivistische Theorien) bezeichnet. Die „kognitive Wende“ wurde zeitlich etwa mit dem Erscheinen des Werks „Cognitive Psychology“ von Ulric Neisser eingeleitet (Ulric Neisser 1967). Allerdings gab es bereits zu früherer Zeit – bevor der Behaviorismus stark ausgebreitet war – Forschungsbemühungen in diese Richtung. Sigrun Richter (2002:2) weist darauf hin, dass beispielsweise das Stufenmodell Jean Piagets bzw. sein Einfluss auf didaktischen Konzepte – bereits durchaus dem entsprach, was in der modernen Kognitionsforschung unter „Kognitionen“ verstanden wird: „für den Organismus bedeutungsvolle und deshalb meist erfahrungsabhängige Wahrnehmungs- und Erkenntnisleistungen“ (Gerhard Roth 1994:29, zit. nach Sigrun Richter 2002:2).

sich wiederum aus der Theorie des operanten Konditionierens von Burrhus Skinner (1938) ableitete. Helmut M. Niegemann bezeichnet das Entstehen der „computergestützten Instruktion“ als einen „Versuch, aus einer Lerntheorie⁶ unmittelbar eine Lehrtechnologie abzuleiten“ (Helmut M. Niegemann 1995:26). Erste „Lehrmaschinen“ wurden Anfang der 1960er Jahre unter anderem von Burrhus Skinner entwickelt. Die Funktionalität derselben lag vor allem darin, den Lehrstoff in möglichst kleinen Einheiten anzubieten. Diese sogenannten „frames“ lagen in der Regel in Textform vor, denen eine Reihe von Fragen folgte. Die Lernenden hatten die Antworten anschließend in die Maschine einzugeben, welche die richtige Lösung anzeigte. Das Lernergebnis konnte auf diese Weise unmittelbar überprüft werden. Im Sinne des operanten Konditionierens sollten die positiven Lerneffekte verstärkt werden, was folgendes bedeutete: wenn es unumgänglich war, dass die Lernenden korrekte Antworten gaben, dann mussten die Fragen dementsprechend einfach gehalten sein, damit keine falschen Antworten zu erwarten waren, wie beispielsweise Lückentexte, die nur ein Wiedererkennen erforderten (vgl. a.a.O.:26).

Helmut M. Niegemann weist darauf hin, dass

„Skinner’s [sic!] Ansatz [...] in der heutigen Diskussion um computergestützte Instruktion nicht selten abqualifiziert [wird], ohne seinen Beitrag zur Entwicklung einer lernpsychologisch begründeten Instruktionstechnologie historisch zu würdigen. Obwohl sich die theoretische Deutung der Lernprozesse bei kognitiven Lehrinhalten als unzureichend erwiesen hat, sind die [...] praktischen Anweisungen Skinners zum Vorgehen bei der Entwicklung von Lernprogrammen [...] in wesentlichen Teilen noch immer aktuell, auch wenn die Begründungen heute andere sind“⁷ (a.a.O.:27).

Die in Deutschland entwickelten Lehrmaschinen⁸, die ab 1964 entwickelt wurden, waren in erster Linie für die Gruppenschulung gedacht. Auch hierbei erhielten die Lernenden die Anweisungen und Informationen über Kopfhörer und Diaprojektor.

„Beim Geromat wurden die optischen Informationen auf Bildschirme an jedem Arbeitsplatz übertragen, bei den beiden

⁶ „Ist das Auftreten einer operanten Verhaltensweise von der Darbietung eines verstärkenden Stimulus gefolgt, so wächst ihre Stärke an.“ (Skinner 1938:21, zit. nach Bower & Hilgard 1983:250).

⁷ Skinners praktische Anweisungen:
„Es muss zunächst eine klare, detaillierte, objektive Beschreibung erarbeitet werden, was es heißt, den Lehrstoff zu können. Es ist eine Folge von Frage-Antwort-Frames zu entwickeln, in welchen der Lernende in langsam ansteigendem Schwierigkeitsgrad mit dem Lehrinhalt konfrontiert wird und wobei dem Lernenden dieselben Fakten aus unterschiedlichen Perspektiven mehrmals dargestellt werden. Es muß dafür gesorgt werden, daß der Lernende aktiv ist, z.B. indem bei jedem Frame eine Antwort auf eine Frage verlangt wird. Auf jede Antwort muss eine sofortige Rückmeldung erfolgen. Die Fragen sollen so gestellt werden, daß die richtige Antwort eine sehr hohe Auftretenswahrscheinlichkeit hat. Jeder Lernende sollte das Lernprogramm in seinem persönlichen Tempo durchlaufen. Das Arbeiten mit dem Lehrprogramm soll durch zusätzliche Verstärkungen gefördert werden“ (vgl. Skinner 1958, nach Gordon H. Bower & Ernest R. Hilgard 1983:278 f., zit. nach Helmut M. Niegemann 1995:27 f.).

⁸ Robbimat O, Geromat III, Bakkalaureus

anderen Systemen wurde eine Projektionsfläche im Schulungsraum verwendet. Während Robbimat ein tonbandgesteuertes Gerät für bis zu 24 Lerner war, handelte es sich beim GEROMAT III bereits um ein rechnergesteuertes System für jeweils drei Lernende. Ein Fortschreiten im Lehrstoff war beim letzteren nur möglich, wenn alle drei Lerner eine richtige Antwort (bei Mehrfachwahl-Aufgaben) gefunden hatten; bei unterschiedlichen Lösungen erhielten alle die Aufforderung, sich zu einigen“ (a.a.O.:33).

Eine durchschlagende Bedeutung erlangte der programmierte Unterricht nicht, im Gegenteil: Es zeigte sich, dass er in erster Linie als Nachhilfe oder Trainingseinheit benutzt wurde. In der Schule erwies sich als Problem, dass durch diese Medien keine individuelle Differenzierung der Lernzeit möglich war. Als der Lehrermangel Ende der 1970er Jahre als behoben galt und auch die Sensibilität der Lehrkräfte gegenüber Entpersonalisierung und Fremdbestimmung wuchs, nahm auch der programmierte Unterricht deutlich ab (vgl. Annabell Preussler 2002:41 ff).

Anfang der 1970er Jahre hielten zunächst die Sprachlabore Einzug in den Unterricht. Beinahe jede Schule verfügte zu dieser Zeit über eine solche Einrichtung. Thomas Bartos (2004:4) weist darauf hin, dass hierbei wieder einmal die Lehrmethoden den technischen Möglichkeiten angepasst wurden und die anfängliche Euphorie schnell in Ernüchterung umschlug. Gründe dafür sind bei Dieter Schwittmann (1987) zu finden:

„Der Grund dafür lag [...] darin, dass die Forschungserwartungen nicht erfüllt wurden, die darin bestanden, eine begründete Entscheidungshilfe für die Medienwahl zu bekommen“ (a.a.O.:193).

Ende der 70er Jahre kam es mit dem Computer-Unterstützten Unterricht (CUU) auch in Deutschland⁹ zu einem erneuten Einsatz der Lehrprogramme in den Schulen, der Erfolg war allerdings ebenfalls nicht weit reichend. Erst mit der Entwicklung von Mikroprozessoren wurden die Geräte sowohl handlicher als auch leistungsfähiger. Es stellte sich rasch heraus, dass eigene Programmierfähigkeiten auf Grund der Anwendungsfreundlichkeit der meisten Programme nicht mehr von großem Nutzen waren. Deutlich wurde, dass neben der technischen Nutzungskompetenz erzieherische Fragen des verantwortlichen Umgangs mit dem Computer in den Vordergrund treten mussten (ebd.).

Wissenschaftliche Untersuchungen, die sich mit der Wirkung von mediengestütztem Lernen befassten, wurden Ende der 1970er Jahre unter anderem vom „Forschungs- und Entwicklungszentrum für objektivierte Lehr- und Lernverfahren in Paderborn“ durchgeführt. Erstmals wurde dabei in einem Forschungsprojekt ein Effizienzvergleich zwischen verschiedenen Medienarten durchgeführt – in diesem Fall Buch und Computer. Insbesondere an Hand der Vergleichskriterien Lernerfolg, Einstellung der Lernenden und Kosten hoffte man, die Überlegenheit eines Mediums nachweisen zu können (vgl. Helmut M. Niegemann 1995:34).

⁹ In den USA begann die Entwicklung bereits Anfang der 1960er Jahre unter dem Begriff Computer Assisted Instruction (CAI).

Die Ergebnisse sprachen jedoch keine klare Sprache – in einigen Untersuchungen wurden Unterschiede zwischen verschiedenen Medienarten festgestellt, in wiederum anderen nicht (a.a.O.:35, vgl. z.B. Simon 1977; Treitz 1984; Mandl & Hron 1985). Dass der herkömmliche, traditionelle Unterricht dem computergestützten Lernen überlegen war, konnte nicht nachgewiesen werden.

Als in den 80er Jahren die „personal computers“ (PCs) Einzug in die Haushalte und Büros hielten, machte diese Entwicklung vor Schulen und Universitäten nicht halt. Thomas Bartos (2004:2) spricht sogar davon, dass diese technische Errungenschaft „die Lehre entscheidend veränderte“. Dies hatte im Wesentlichen zwei weit reichende Konsequenzen für die Lehre: zunächst einmal wurden die PCs dafür verwendet, Wissen zu vermitteln und zu verbreiten. Zu diesem Zweck mussten aber zweitens diejenigen, die mit ihnen arbeiten sollten, zunächst technisch geschult werden, was ebenfalls mittels Lernprogrammen erfolgte. Aus dieser Entwicklung heraus gewann der Begriff des „lebenslangen Lernens“ an Bedeutung und Relevanz (vgl. Dieter Schwittmann 1987). Von der didaktischen Entwicklung unterschieden sich diese Programme allerdings kaum von den früher entstandenen Lehrmaschinen, da auch hier zumeist behavioristische Konzepte in so genannten drill-and-practice-Programmen Anwendung fanden (vgl. ebd.). Es verwundert daher wenig, dass zwar Novitätseffekte bei der Anwendung beobachtbar waren, der Erfolg jedoch schnell abnahm.

Mitte der 1980er Jahre wurde computergestütztes Lernen von Seiten der Wirtschaft – z.B. der Siemens AG, der Allianz AG oder der Deutschen Bundespost – finanziell gefördert, allerdings lag der Schwerpunkt bei diesen Investitionen in der Entwicklung von Lernprogrammen für die betriebliche Aus- und Weiterbildung (vgl. Helmut M. Niegemann 1995:34). Neben der technischen Nutzungskompetenz traten erzieherische Fragen des verantwortlichen Umgangs mit dem Computer in den Vordergrund. Aus diesem Grund verabschiedete die Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung (BLK) 1984 zunächst ein Rahmenkonzept (1984) und später ein Gesamtkonzept „für eine informationstechnische Grundbildung, die für alle Schülerinnen und Schüler in der Sekundarstufe I verbindlich sein sollte“ (BLK 1987, zitiert nach Hagemann 2001:30).

Seit den 1990er Jahren ist ein immer stärker werdender Trend in Richtung computergestütztes Lernen festzustellen. Bereits seit der Einführung der CD-ROM und dem damit ermöglichten Einsatz von Multimedia-Angeboten wird der Computer wieder in hohem Maße zu Lernzwecken eingesetzt. Die didaktischen Möglichkeiten waren jedoch der einseitigen Kommunikation und der mangelnden Interaktion wegen noch begrenzt (vgl. Thomas Bartos 2004:3).

Das WWW¹⁰ wird seit Beginn der 1990er Jahre zum Lernen genutzt, bot es in den Anfängen doch auch die Möglichkeit, schnell auf

¹⁰ „World Wide Web“: Hypertextsystem, das über das Internet zur Verfügung gestellt wird.

große Datenmengen und Informationen zuzugreifen. Zu dieser Zeit begannen erste Schulen, sich durch Mailbox-Netze untereinander zu vernetzen. Dabei waren das Deutsche SchulNetz und das Offene Deutsche Schulnetz die größten Netze dieser Art. „Im März 1995 waren in Deutschland 5 Schulen im WorldWideWeb präsent, 18 Monate später waren es bereits mehr als 500“ (Nicola Döring 1997:331).

Mittlerweile gab es etliche wissenschaftliche Projekte, die den Einsatz von neuen Medien zum Lernen zum Thema hatten, jedoch auch hier fielen die Ergebnisse dieser Studien unterschiedlich aus. Fanden beispielsweise Sheryl L. Boblin & Barbara A. Gibson (1986) einen Vorteil gegenüber computergestütztem Lernen¹¹, konnten Pauline Kot, Lynn D. Skillen & Marilyn Wales im selben Jahr keine Unterschiede im Lernen belegen¹². Chen-Lin C. Kulik, James A. Kulik und Barbara Shwalb, die ebenfalls 1986 eine groß angelegte Metaanalyse zum Thema „The Effectiveness of Computer-based Adult Education“ durchführten, hatten wiederum die Überlegenheit der Technologie zum Ergebnis¹³.

Einen weiteren Aufschwung erfuhr der Trend Mitte der 1990er Jahre. Die Einführung von ISDN¹⁴ wurde vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie und der Deutschen Telekom zum Anlass genommen, die Initiative „Schulen ans Netz“ zu gründen. Ziel war es, bis zum Jahre 2001 alle bundesdeutschen Schulen an das Internet anzuschließen und den Schülerinnen und Schülern ein gewisses Maß an Medienkompetenz zu vermitteln. Insgesamt wurde dies von den Schulen gut angenommen und fand große Akzeptanz, obwohl jeweils nur geringe Teile des Kollegiums damit arbeiteten (vgl. Frank Weinreich & Renate Schulz-Zander 2000:578). Andererseits war das Interesse seitens der Schülerinnen und Schüler im Bereich der Informations- und Kommunikationstechnologien enorm hoch – zumal die Mehrheit auf Grund von privater Nutzung bereits auf Erfahrungen in diesem Bereich zurückgreifen konnte (vgl. Annabell Preussler 2002:51).

Erstmals wurde ein „Wandel des Unterrichts“ festgestellt, da in der Arbeit mit den neuen Medien mehr Aufmerksamkeit und Kooperation wahrgenommen wurde und der Unterricht insgesamt lernendenzentrierter war (vgl. Heike Hunneshagen u.a. 2000:59). Trotz allem wurde das Internet nur in wenigen Fächern eingesetzt. Die Schwierigkeiten waren

¹¹ „... students in the class using CMI achieved higher marks for all of the unit tests, for the average of the unit tests and for the final course mark when compared to the previous three (non-CMI) classes.“ (Sheryl L. Boblin & Barbara A. Gibson 1986, zit. nach <http://www.nosignificantdifference.org/search.asp>, letzter Aufruf 04.09.2005)

¹² „Achievement scores and learning time for the two groups were equivalent, indicating equal amounts of learning.“ (Pauline Kot, Lynn D. Skillen & Marilyn Wales 1996, zit. nach <http://www.nosignificantdifference.org/search.asp>, letzter Aufruf 04.09.2005).

¹³ „In nineteen of the twenty-three studies, the CBE [computer-based education; Anm. AP] class had the higher examination average; in four studies, the conventionally taught class had the higher average. The difference in examination performance of CBE and control students was reported to be significant in eleven studies“ (Chen-Lin C. Kulik, James A. Kulik & Barbara Shwalb 1986:242)

¹⁴ Integrated Services Digital Network, ein internationaler Standard für Telekommunikationsnetze.

dabei selten pädagogischer Art, eher mangelte es an entsprechenden Kapazitäten wie Ausstattung, Kosten und Zeit (vgl. a.a.O.:178).

Bemerkenswert an der Initiative war, dass „der bildungspolitische Fortschritt [...] über einen technischen Vorgang ausgedrückt [wurde] [...], und nicht etwa über eine pädagogische Zielformel“ (Euler 2001, zit. nach Thomas Bartos 2004:3).

Zusammenfassend stellt Thomas Bartos fest:

„Betrachtet man die Entwicklung der multimedialen bzw. telemedialen Lehre in ihrem bisherigen Verlauf, werden ihre Charakteristiken schnell augenfällig. Der Anspruch der „Neuen Lehre“ wurde bisher hauptsächlich durch den Einsatz neuer Technik determiniert. [...] Innovationen wurden stets von einer übermäßigen Euphorie begleitet und endeten ein jedes mal in der Ernüchterung“ (Thomas Bartos 2004:4, vgl.auch Michael Kerres 2001; Bernd Weidenmann 1997; Maik Adomßent 2001:161f.).

Es wurde in diesem kurzen historischen Abriss deutlich, dass die Übergänge der genannten Entwicklungslinien beim Einsatz des Computers zum Lernen fließend waren. Nicht selten sorgte diese Entwicklung für Kritik innerhalb des Bildungswesens. Ebenso wurde deutlich, dass der Medieneinsatz in Schule und Hochschule auch stark von den technischen Möglichkeiten abhing.

2.2 Lernen in der Wissensgesellschaft

Durch die zunehmende Informatisierung und der damit einhergehenden Explosion von Information ist unsere Gesellschaft zunehmend zu einer Wissensgesellschaft geworden, in der Informationstechnologien eine wesentliche Rollen spielen und gleichsam als Faktor der wirtschaftlichen (und gesellschaftlichen) Entwicklung angesehen werden¹⁵. Doch was macht diese Gesellschaft aus?

Wie ich bereits an anderer Stelle gezeigt habe (vgl. Annabell Preussler 2002:29ff), gibt es in der Literatur dazu verschiedene Meinungen, jedoch wird ein Merkmal bei den meisten Definitionen genannt: die Informatisierung, womit eine „Durchdringung einer wachsenden Anzahl gesellschaftlicher Bereiche durch vielfältige Formen und Anwendungen von IKT¹⁶“ gemeint ist (Oliver Hauf 1996:83).

Bis vor einigen Jahren wurde dieser gesellschaftliche Wandel als Informationsgesellschaft definiert – dieser Begriff wurde 1963 eingeführt (Peter J. Weber 2005:47), da Information „zu einem zentralen Rohstoff, Produktionsfaktor und Wettbewerbsvorteil“ geworden war (Hans-Dieter Kübler 1999:30). Dazu kommt, dass nicht nur eine Veränderung des Alltags durch die Medien stattgefunden hat (vgl. Klaus Hurrelmann 1994:398), sondern Medien als Übermittler von Informationen diese die Wirklichkeit zunehmend mit strukturierten (vgl. Bildungskommission NRW

¹⁵ Vgl. <http://de.wikipedia.org/wiki/Informationsgesellschaft> [letzter Aufruf, 18.09.2005].

¹⁶ Informations- und Kommunikationstechnologien

1995:135). Dies ist damit zu begründen, dass die Menge, die Geschwindigkeit und die Effizienz bei der Informationsbeschaffung und deren Verarbeitung höchste Priorität haben (vgl. Nina Degele 2000:18). Dieser Priorität nehmen sich die Technologien an, da sie einen globalen Zugriff auf Informationen ermöglichen (vgl. Heinz Mandl, Gabi Reinmann-Rothmeier und Cornelia Gräsel 1998:6).

Inzwischen hat sich der Fokus mehr auf den Bereich des Wissens verlagert, denn Handlungsfähigkeit innerhalb dieser Gesellschaft kann nur durch Wissen erreicht werden (vgl. Hans-Dieter Kübler 1999:30). Dies kann jedoch nicht allein durch die Verarbeitung von Daten entstehen.

„Wissen ist eine notwendige, aber keine ausreichende Fähigkeit zum Handeln. Um etwas in Gang setzen zu können, müssen die Umstände, unter denen dies geschehen soll, der Kontrolle des Handelnden unterliegen“ (Nico Stehr 1994:242f).

Peter J. Weber (2005:47) fasst diese Aussage so zusammen, dass Informationen erst durch die Bewertung und Verknüpfung mit vorhandenen Informationen zu Wissen werden. Die Wissensgesellschaft wird also als eine Gesellschaft definiert, „in der Wissen in allen Bereichen zunehmend Grundlage und Richtschnur menschlichen Handelns wird“ (Nico Stehr 2001:10, zit. nach Peter J. Weber 2005:47).

Im Gegensatz zur Informationsgesellschaft, in welcher die Information an sich Fokus der Diskussion war, liegt der Schwerpunkt der Wissensgesellschaft auf dem einzelnen Menschen. Die „verantwortungsbewußte und sozial verträgliche Wissensnutzung“ wird mit dem Ziel „der Wahrung von Grundrechten und demokratischem Zusammenleben“ gefordert (Heinz Mandl & Gabi Reinmann-Rothmeier 1997:78). Die Wissensgesellschaft wird als das Leitbild des 21. Jahrhunderts angesehen.

Einen weiteren Grund dafür führt Stefan Aufenanger (2001) an, indem er vermutet, „dass die Herausforderungen in der globalen Gesellschaft nur durch die neuen Informations- und Kommunikationstechnologien bewältigt werden können“ (a.a.O.:256). Wichtig ist auch die Bedeutung dieses Wissens. Die Menschen, die in einer Wissensgesellschaft leben, müssen zumindest wissen, auf welche Art und Weise welche Sachverhalte erlernt werden können.

Die Definition von „Wissensgesellschaft“ zielt ab auf die Beschaffenheit moderner Gesellschaften, deren Organisation auf Wissen basiert. Wissen ist dabei nicht nur zentrale Voraussetzung gesellschaftlicher Entwicklung, sondern auch zur wichtigsten Produktivkraft geworden (vgl. Helmut F. Spinner 1998:72). Anders als der Begriff der Informationsgesellschaft umfasst die Wissensgesellschaft „zudem das kollektive Interesse am Ergründen von Gesellschaft und Natur durch systematische, nachvollziehbare, standardisierte Verfahren“ (Wikipedia 2005¹⁷).

¹⁷ <http://de.wikipedia.org/wiki/Wissensgesellschaft> [letzter Aufruf, 18.09.2005].

Es ist in der Wissensgesellschaft durch moderne Informations- und Kommunikationstechnologien möglich geworden, nahezu unbegrenzt Informationen einzusammeln. Dem trägt auch die stetig voranschreitende Vernetzung Rechnung, wodurch regelrechte „Wissenspools“ entstehen. Als Beispiel sei die Online-Enzyklopädie Wikipedia genannt, in der Informationen von tausenden von Nutzenden kollektiv zusammengetragen werden.

Natürlich führt ein solcher gesellschaftlicher Umbruch auch zu einem Wandel des Bildungssystems. Lebenslanges Lernen erscheint als ein Schlagwort (vgl. Stefan Aufenanger 2001:256), ebenso Kommunikation und Kooperation (vgl. Renate Schulz-Zander 2000:105). Es reicht allerdings nicht aus, nur Technologien zur Verfügung zu stellen. Solange sich nicht die didaktische Umsetzung verändert, „wird weder die Lernkultur noch das konkrete Lehren und Lernen verändert. Die spezifischen Potentiale der Neuen Medien müssen zielgerecht genutzt werden, hierfür sind durchdachte Konzepte zum Medieneinsatz notwendig“ (Heinz Mandl & Gabi Reinmann-Rothmeier 1997:84 f.).

Der Paradigmenwechsel zeigt sich darüber hinaus auch insofern, als der Zugang zu Informationen und Wissen nicht mehr nur privilegierten Mitgliedern der Gesellschaft vorbehalten ist (z.B. Expertinnen und Experten in den jeweiligen Gebieten), sondern allgemein verfügbar wird. Dies beinhaltet auch einen veränderten Status der Lehrenden. Waren sie im traditionellen Unterricht noch VermittlerInnen von Informationen, zeichnet sich deutlich eine Funktionsänderung in Richtung ModeratorIn oder Coach von Bildungsprozessen ab. So wandelt sich das Lernen nach der konstruktivistischen Sichtweise zu einem aktiven Prozess, den die Lernenden eigenständig konstruieren und in welchem sie selbst mitverantwortlich für die Bildungsinhalte sind (vgl. Annabell Preussler 2002:39).

Diese neuen Möglichkeiten erfordern neue Kompetenzen: Medienkompetenz ist zu einer Schlüsselqualifikation in der Wissensgesellschaft geworden (vgl. Renate Schulz-Zander 1997:10). Sie soll „zu einer konstruktiven und zugleich kritischen Nutzung der Medien führen“ (Bildungskommission NRW 1995:137).

„Medienkompetenz bedeutet ‚Handlungsfähigkeit‘ in Bezug auf Medien und – umfassender gesehen – Handlungsfähigkeit in einer durch Medien bestimmten Welt. [...] Medienkompetenz hat immer eine Zieldimension: Mündigkeit, Selbstbestimmung, Selbstverwirklichung, Befriedigung von Bedürfnissen, Bildung, Ich-Identität werden – in sozialer Verantwortung – angestrebt“ (Jochen Hettinger & Wolfgang Wunden 2000:5 f.).

Im Zuge dieser Entwicklung haben sich in den letzten Jahren viele Bildungsinstitutionen dieser Herausforderung angenommen. Neben reinem Online-Lernen hat dort auch der Ansatz des Blended Learning¹⁸ Einzug gehalten. Dabei werden die neuen Medien nicht alternativ, sondern begleitend zu bestehenden Lehr- und Lernformen eingesetzt.

¹⁸ oder auch: hybrides Lernen

Dieses Lernarrangement setzt auf die Vorteile des E-Learnings (Flexibilität, Orts- und Zeitunabhängigkeit etc.) und kombiniert dies mit Präsenzangeboten. Dabei ist festzuhalten, dass diese nicht zusätzlich angeboten werden, sondern von vornherein ein didaktisch begründeter Bestandteil des Blended-Learning-Konzepts sind (vgl. Michael Kerres, Claudia de Witt & Jörg Stratmann 2002:6f.).

Dieser Trend zu kooperativem und kollaborativem Lernen setzt sich mit den neuesten Entwicklungen fort. Web 2.0¹⁹ ist ein Schlagwort, das heutzutage in aller Munde ist und die Veränderung des World Wide Web im Hinblick auf neue interaktive Techniken und Dienste meint. Der wichtigste Aspekt ist m.E. daran, dass Web 2.0 bzw. „Social Software“²⁰ damit auch gleichzeitig ein verändertes Verständnis des Netzes impliziert. Gab es noch vor einigen Jahren überwiegend statische Webseiten, so ist in den letzten Jahren eine Entwicklung zu dynamischen Modellen – beispielsweise Weblogs – zu beobachten. Dabei liegt besonderes Augenmerk auf der Eigenverantwortlichkeit der Nutzenden. Während die Technologien der 90er Jahre vor allem auf Veröffentlichung fokussiert waren, bietet Web 2.0 die Möglichkeit der aktiven Teilnahme. So ist die „Encyclopædia Britannica“ als reines Nachschlagewerk in ihrem Umfang und ihrer Aktualität von der wikibasierten Online-Enzyklopädie Wikipedia abgelöst worden (vgl. Stefan Kreml 2005), die wesentlich davon profitiert, dass Nutzende ihr Wissen dort zur Verfügung stellen und Seiteninhalte ständig neu und eigenverantwortlich aktualisieren. „User generated content“ ist einer der Leitsätze unserer Zeit und liefert eine Erklärung dafür, warum Web 2.0-Anwendungen auch als „social software“ verstanden werden. Die „Verknüpfung von Personen“ – und nicht Dateien, wie Peter Baumgartner (2006a:1) das zentrale Merkmal von Web 2.0 charakterisiert, baut sich dabei in einem „bottom-up“-Prozess auf. UserInnen setzen sich selbst mit anderen UserInnen in Beziehung.

„Das Auffinden und Herstellen von sozialen Kontakten mit gleicher Interessensbasis ist dabei das wesentliche – die Definition von Social Software bestimmende – Kriterium“ (a.a.O.:2).

Diese neue Art der Webkultur birgt bemerkenswerte Potentiale für das Lernen, nehmen doch Interaktion und Aktivität einen hohen Stellenwert ein und wird formales Lernen nun endlich informalisiert (vgl. ebd., Sebastian Pütz 2006). Es verwundert nicht, dass sich diese Ansätze wie ein Lauffeuer auch in der Bildungslandschaft verbreiten. So fanden sich zunächst in den USA, später auch in Deutschland, Österreich und in

¹⁹ Der Begriff Web 2.0 hat seinen Ursprung durch die gleichnamige Konferenzreihe des O'Reilly-Verlags und des Konferenzveranstalters MediaLive International, die erstmalig im Oktober 2004 stattfand (vgl. http://de.wikipedia.org/wiki/Web_2.0 , letzter Aufruf 25.01.2007).

²⁰ „Web 2.0“ und „Social Software“ haben im Prinzip die selbe Bedeutung. „Web 2.0“ IST „Social Software“

der Schweiz Hochschulen, die mit Weblogs und Wikis²¹ arbeiten. Aber gerade, weil es einen Widerspruch darstellt, informales Lernen in formale Lernarrangements zu integrieren, sind die didaktischen Potentiale und der praktische Nutzen von social software für viele Bildungseinrichtungen noch unklar (vgl. a.a.O.:1ff,5).

Bemerkenswert ist – am Ende bemerkt – noch die Tatsache, dass sich InternetnutzerInnen heute weitgehend auf dem neuesten Stand bewegen (vgl. Martin Fisch & Christoph Gscheidle 2006:431). Bei fast jedem zweiten deutschen Internetzugang handelt es sich um eine DSL-Verbindung, davon in 77 Prozent der Fälle um Flatrates – also Angebote zu einem monatlichen Festpreis (a.a.O.:432). Ein Viertel der Nutzenden ist am aktiven Umgang mit dem Internet interessiert. So hat sich ein Drittel aller Anwendenden bereits in einem Forum eingebracht, bei der Nutzung von DSL-Verbindungen sind es sogar 44 Prozent (a.a.O.:434f.). Der Paradigmenwechsel hin zu Web 2.0 zeigt sich also deutlich.

Es dürfte klar geworden sein, dass auch der Bildungsbereich nicht mehr an der Nutzung von WWW und Internet vorbeikommt. Doch inwieweit wird das Lernen überhaupt durch den Einsatz von Software beeinflusst?

2.3 Die Kozma-Clark-Debatte²²

Wie bereits erwähnt, können vergleichende empirische Studien nicht eindeutig belegen, ob Veränderungen beim Lernen tatsächlich auf den Medieneinsatz zurückzuführen sind (vgl. Martin Freudenreich & Carsten Schulte 2002:1). Dieses Problem der Evaluationsforschung hat bereits Mitte der 1990er Jahre vor allem durch die Diskussion zwischen Robert B. Kozma und Richard E. Clark große Prominenz erzielt, die seitdem als Kozma-Clark-Debatte bekannt ist (ebd.)²³.

Durch Reviews und Metaanalysen zeigte Clark (1983) auf, dass die Effektivität des Lernens durch die Wahl des Mediums nicht beeinflusst wird. Zur Erläuterung dieser Aussage bedient er sich der Metapher eines Lieferwagens: Medien seien Fahrzeuge, die die Instruktion an die Lernenden lieferten, diese aber nicht selbst beeinflussten. Ebenso wenig beeinflusse der Lieferwagen für Lebensmittel die Ernährung. Die Wahl des

²¹ „Ein Wiki, auch WikiWiki und WikiWeb genannt, ist eine im World Wide Web verfügbare Seitensammlung, die von den Benutzern nicht nur gelesen, sondern auch online geändert werden kann. Dazu gibt es in der Regel eine Bearbeitungsfunktion, die ein Eingabefenster öffnet, in dem der Text des Artikels bearbeitet werden kann. Wie bei Hypertexten üblich, sind die einzelnen Seiten und Artikel eines Wikis durch Querverweise (Links) miteinander verbunden. Wikis ähneln damit Content Management Systemen. Der Name stammt von wiki, dem hawaiischen Wort für ‚schnell‘“ (<http://de.wikipedia.org/wiki/Wiki> , letzter Aufruf 16.06.2007).

²² Da diese Debatte unter „Kozma-Clark-Debatte“ bekannt wurde, wird dies auch an dieser Stelle fortgeführt, selbst wenn Clark in einer alphabetischen Sortierung an erster Stelle genannt werden müsste (vgl. Martin Freudenreich und Carsten Schulte 2002:1).

²³ Bereits Anfang der 70er Jahre war die Untersuchung von Methode und Medium in Deutschland und den USA ein Thema innerhalb der Unterrichtswissenschaft (vgl. Patrick Blumschein 2003:4).

Lieferwagens mag bestenfalls Einfluss auf die Kosten oder das Ausmaß der Instruktion haben, aber nur dessen Inhalt kann tatsächlich die Leistung beeinflussen (vgl. Richard E. Clark 1983:445; Patrick Blumschein 2003:5).

Kozma (1994a) greift diese Position auf und stellt dar, dass das Medium sehr wohl einen Einfluss auf die Auseinandersetzung mit einer Problemstellung seitens der Lernenden ausübe. Dies sei der Fall, wenn die Darstellung der Inhalte mittels Film oder Multimedia erfolge. Darüber hinaus könnten die Lernenden die dargebotenen Informationen auf diese Weise besser an bereits bestehendes Wissen anknüpfen (vgl. Patrick Blumschein 2003:5). Er stellt sich also weniger die Frage, ob Medien das Lernen beeinflussen, sondern eher wie sie das tun (vgl. Robert B. Kozma 1994b:7). Kozma weist darüber hinaus darauf hin, „dass eine Beziehung von eingesetztem Medium und Lernerfolg sich nicht wie in naturwissenschaftlichen Kontexten gleichsam *auffinden* lasse, sondern *gestaltet* werden müsse“ (Martin Freudenreich & Carsten Schulte 2002:2; vgl. Robert B. Kozma 1994b:7). Freudenreich & Schulte stellen fest:

„Die Kritik Kozmas an Vergleichsstudien zum Medieneinsatz bezieht sich auf die Art und Weise, in der diese durchgeführt werden: Medieneffekte würden meist auf Grund eines standardisierten Vor-/Nachtest Designs nachgewiesen, dagegen fehlten in diesen Studien kognitive, affektive und soziale Aspekte, unter denen aktives Lernen stattfindet“ (ebd.).

Zur Veranschaulichung verwendet Kozma das Bild eines Tornados, der eine Stadt verwüstet. Es reiche in diesem Fall nicht aus, nur Fotografien dieses Ereignisses zu betrachten, um es wirklich zu verstehen. „Gerade der prozesshafte Charakter des Geschehens ist von Interesse für ein vertieftes Verständnis sowohl von Tornados als auch von Lernprozessen, so dass der Umgang der Lernenden mit der Software, der Lernprozess selbst, untersucht werden müsse“ (ebd.).

Richard E. Clark (1994a) ist dagegen der Auffassung, dass lediglich die Methode bzw. die Lehrstrategie das Lernen beeinflusse. Er betont, dass ein Medium dann wertlos sei, wenn dessen Ersetzen nicht zu einem Verlust an instruktionaler Qualität führe (vgl. Patrick Blumschein 2003:6; vgl. Clark 1994a:22). Er verweist darüber hinaus auf Forschungsergebnisse, in denen keine Effekte des Instruktionsmediums feststellbar waren²⁴ (vgl. ebd.).

Verschiedene AutorInnen greifen die dargestellte Debatte zwischen Kozma & Clark auf²⁵. So begibt sich beispielsweise Gavriel Salomon (2000) auf den Standpunkt, dass Lernqualität sich nicht allein aus der Medienwahl²⁶ bestimmen lasse. Dennoch blieben Effekte der neuen Medien für das Lernen weitestgehend aus. Salomon ist der Ansicht, dass eine Reihe von Umständen für diese „Enttäuschungen“ ausschlaggebend seien: Neue Technologien werden zwar für das Lehren und Lernen eingesetzt, die Implementation ihrer Möglichkeiten wird allerdings nicht

²⁴ Z.B. die Metaanalysen von Kulik & Kulik (1989)

²⁵ Vgl. auch Patrick Blumschein 2003:6ff.

²⁶ Er bezieht sich dabei auf die neuen Medien.

vollzogen. Salomon spricht an dieser Stelle sogar von einer Domestizierung der Technologien:

„A most powerful and innovative technology is taken and is domesticated such that it does more or less what its predecessors have done, only it does it a bit faster and a bit nicer“ (ebd.).

Technologie mache laut Salomon keinen Unterschied im Lernen, dies liege jedoch an der bestehenden Praxis, wie die Technologie verwendet werde. Er kritisiert den in diesem Zusammenhang vorherrschenden Fokus zahlreicher Veröffentlichungen auf technische Aspekte, was zu einer Vernachlässigung der Lernendenperspektive führe. Nicht das Wissen, sondern der Computer werde damit zum Herzstück des Lernens und Lehrens:

„The computer, like the Biblical apple, commands much attention, as it is a far more tempting object than a new approach to teaching“ (ebd.)

Salomon steht zusammenfassend auf dem Standpunkt, dass sich die Technologien zwar dazu eignen, Probleme zu lösen – beispielsweise als Nachschlagewerke oder Anschauungsobjekte für Simulationen o.ä., es sei ihnen aber nicht möglich, Informationen in Wissen umzuwandeln.

Auch in neueren Publikationen wird der Zusammenhang zwischen Medium und Methode thematisiert. Michael Kerres, Claudia de Witt und Jörg Stratmann (2002:2) stellen beispielsweise fest, dass „die Wahl der didaktischen Methode [...] in weitem Masse [sic] von der Wahl des Mediums unabhängig [ist]“. Dabei sei die Wahl der Methode von einer Reihe von Parametern – Zielgruppe, Lehrinhalte, Lehrziel usw. – abhängig (vgl. Ewald Terhart 1997, zit. nach Michael Kerres, Claudia de Witt & Jörg Stratmann 2002:2).

2.4 Metaanalyse im Bereich des mediengestützten Lernens

Die Frage, welche Methode bzw. welches Lernsetting für das Lernen besser geeignet ist, stellt schon seit vielen Jahren ein großes Thema in der Forschung dar. Da es – wie ich bereits gezeigt habe – eine Vielzahl an Untersuchungen gibt, die unterschiedliche Ergebnisse aufweisen, wurde mittels Metaanalysen versucht, diese Erkenntnisse zu vereinheitlichen. Eine bedeutende Studie dieser Art stellt die Metaanalyse zum computerunterstützten Lernen von Chen-Lin C. & James A. Kulik (1991) dar, die eine Nachfolgestudie einer bereits 1986 publizierten Untersuchung darstellt (vgl. Chen-Lin C. Kulik, James A. Kulik & Barbara Schwalb 1986). Darin wurden 248 vergleichende Studien ausgewertet, von denen 46 (entspricht 19%) ein besseres Ergebnis für herkömmliches Lernen und 202 (entspricht 81%) einen höheren Lernerfolg für das computerbasierte Lernen aufweisen²⁷ (vgl. ebd.).

²⁷ Nur in 100 Fällen waren diese Ergebnisse signifikant (vgl. Gerhard Tulodziecki 2001:3).

Gerhard Tulodziecki (2001:3) weist in diesem Zusammenhang auf ein bekanntes Problem bei metaanalytischen Verfahren hin, dass nämlich Studien, die signifikante Ergebnisse aufweisen, öfter publiziert werden als Studien, bei denen dies nicht der Fall ist (vgl. dazu die Ausführungen in Kapitel 7.1 Metaanalyse). Insofern sei das Ergebnis der Kulik'schen Metaanalyse unter Umständen zu Gunsten des computergestützten Lernsettings verzerrt. Andererseits müsse jedoch berücksichtigt werden, dass zu dieser Zeit noch nicht die technischen Möglichkeiten der Multimedialität vorhanden waren (vgl. ebd., vgl. Michael S. Jenks & John M. Springer (2002:52f.).

Patrick Blumschein (2004:10) stellt die wichtigsten Arbeiten der 1980er Jahre in diesem Bereich folgendermaßen zusammen:

„Die bekanntesten sind ohne Zweifel die Metaanalysen von Kulik/Kulik/Cohen (1980) und Kulik/Bangert/Williams (1982) zum computergestützten Lernen, von Kulik/Kulik (1989), Kulik/Kulik/Shwalb (1986), Fricke/ Treinies (1985) zu CBT. Andere Forscher haben sich mit der Effektivität von Videodisk-Lernprogrammen befasst (Bosco/Wagner 1988). In diesem Zusammenhang ist eine von Fletcher (1990) durchgeführte Metaanalyse zu „computer controlled interactive videodisc instruction“ zu nennen, die in drei Settings realisiert wurde: higher education, industrial training und military training. [...] Ebenfalls zu erwähnen ist eine von Kulik/Kulik (1991) durchgeführte Metaanalyse von 254 Studien, die konventionelle Lernumgebungen mit CBI (Computer-Based- Instruction) verglichen. Mehr oder weniger wurden mit dieser Studie die Ergebnisse früherer Studien bestätigt (etwa Kulik/Kulik 1987)“ (Patrick Blumschein 2004:10).

Die ersten Metaanalysen zu interaktiven Multimedialprogrammen wurden ab Mitte der 1990er Jahre durchgeführt (vgl. Edwin P. Christmann, John L. Badgett & Robert Lucking 1997; Claire M. Fletcher-Flinn & Breon Gravatt 1995; Mohammad N. Yaakub 1998, nach Patrick Blumschein 2004:12). John Schacter & Cheryl Fagnano (1999) fassen den Stand der Forschung der 90er Jahre zusammen, indem sie weitere wichtige Metaanalysen auflisten (vgl. auch Patrick Blumschein 2004:13). Ich möchte an dieser Stelle darauf verzichten, alle Studien namentlich zu erwähnen und möchte statt dessen die Ergebnisse diskutieren, in denen die Befunde von über 700 Einzelstudien eingegangen sind. Demnach zeichnen sich bessere Leistungen der SchülerInnen im computergestützten Unterricht ab (vgl. John Schacter & Cheryl Fagnano 1999:9). Auch Jay Sivin-Kachala und Ellen R. Bialo (2000:10) stellen in einem Forschungsbericht²⁸, der die Effektivität von Technologien in der Schule zusammenfasst, fest, dass in allen Wissensbereichen über alle Schularten hinweg positive Lerneffekte beim Einsatz neuer Medien auftraten (vgl. Patrick Blumschein 2004:14f.).

Auch aus jüngerer Zeit sei an dieser Stelle eine metaanalytische Arbeit vorgestellt: 2003 veröffentlichte ein Forschungsteam der University of Houston (Hersh C. Waxmann, Meng-Fen Lin & Georgette M. Michko 2003) Ergebnisse aus 42 Primärstudien, die sich mit der Effektivität

²⁸ Es handelt sich hierbei um eine Metaanalyse, die 311 Studien umfasst (vgl. Patrick Blumschein 2004:14).

technologiegestützten Lernens beschäftigt hatten. Ziel der Metaanalyse war die Untersuchung der Effekte in Bezug auf den Lernerfolg: „to estimate the effects of teaching with technology on students' cognitive, affective, and behavioral outcomes of learning“ (a.a.O.:1). Die WissenschaftlerInnen stellten einen (kleinen) positiven Effekt – jedoch mit hoher Effektstärke – zu Gunsten des computergestützten Unterrichts im Gegensatz zu traditionellen Lehrformen fest (vgl. a.a.O.:13).

Zusammengefasst lässt sich die Forschungslage also im Sinne von Gerhard Tulodziecki charakterisieren: Die empirischen Befunde sprechen insgesamt dafür, „dass mit geeigneter Software eine Verbesserung des Lernens im Sinne besserer Lerngewinne und verkürzter Lernzeiten zu erwarten ist“ (Gerhard Tulodziecki 2001:3). Diese Verbesserung stellt jedoch keinen Automatismus dar, sondern findet nur unter bestimmten Bedingungen statt, die weiterer Forschung bedürfen (vgl. ebd.).

Darüber hinaus ist zu bedenken, dass sich sowohl die Technologie selbst sowie auch deren Einsatz immer weiter entwickelt und fortsetzt. War der Computereinsatz im Unterricht in den 1980er Jahren noch durch isoliertes Rezipieren und bestenfalls kleine Simulationen geprägt, werden heutzutage bereits interaktive webbasierte Lernszenarien in Schulen und Hochschulen gefördert, die besonders auf die Aspekte der Kommunikation und der Verfügbarkeit von Wissen fokussieren. Das bedeutet, dass auch metaanalytische Verfahren den aktuellen Stand nur bedingt abbilden können, da sie überdies methodisch schwierig sind (vgl. Kapitel 7.1 Metaanalyse ab Seite 65).

3 Evaluation – eine Annäherung

Evaluation – was ist das? Ich möchte zu Beginn dieses Kapitels drei Beispiele von Martin Bonsen & Andreas Büchter (2005:7ff) aufgreifen:

War beispielsweise Beethoven ein Evaluator, weil er seine Sinfonien dadurch verbesserte, dass er sie von einem Orchester immer wieder probespielen ließ?²⁹ Kann man es als Evaluation bezeichnen, wenn eine Reihe von Zahnarztpraxen, die Hypnose als neue Schmerztherapie einsetzen, eine Befragung ihrer PatientInnen hinsichtlich ihres Schmerzempfindens vor und nach der Behandlung durchführen und diese Werte mit denen von Praxen vergleichen, die nach herkömmlichen Methoden arbeiten? Ist es Evaluation, wenn in einem Kindergarten die Entscheidung ansteht, ob zukünftig Halloween oder St. Martin gefeiert werden soll und unter den Kindern nach der Durchführung eines Halloweenfestes eine Befragung durchgeführt wird, wie ihnen dies gefallen habe?

Es wird deutlich, dass Evaluation viele Bedeutungen haben und unterschiedlich ausgelegt werden kann. Christine Schwarz bezeichnet Evaluation gar als modernes Ritual:

„Was Evaluation ist, wie man sie richtig macht und wozu sie da ist, darüber gibt es keinen Konsens, das bleibt diffus. Vielleicht liegt genau in dieser Flüchtigkeit und Flexibilität das Potential von Evaluation, das sie zur heutigen Legitimation bürokratischer Herrschaft so unerlässlich macht: Mit Evaluation legt man sich nicht wirklich fest. Vielmehr hilft Evaluation, sich möglichst viele Handlungsoptionen zu eröffnen. So bleiben die Einzelnen gegenüber der Flexibilität anderer allumfassend flexibel“ (Christine Schwarz 2004:257).

3.1 Begriffsklärung

Betrachten wir zur Klärung des Begriffes zunächst ein Standardwerk zur Evaluation. Jürgen Bortz & Nicola Döring (1995) legen ihrer Arbeit nach Peter H. Rossi & Howard E. Freeman (1989) folgende Definition zu Grunde:

„Evaluationsforschung beinhaltet die systematische Anwendung empirischer Forschungsmethoden zur Bewertung des Konzepts, des Untersuchungsplanes, der Implementierung und der Wirksamkeit sozialer Interventionsprogramme“ (a.a.O.:96).

Systematisch bedeutet hier, dass Ziele, Zwecke, Verfahren und Methoden im Vorfeld der Evaluation geklärt werden, dass die Datenerhebung mittels geeigneter Messverfahren und unter Berücksichtigung relevanter Indikatoren und Kriterien erfolgt, dass die Bewertung auf der Basis einer nachvollziehbaren Analyse geschieht und dass das gewonnene Ergebnis als Planungs- bzw- Entscheidungshilfe zur

²⁹ vgl. Beethoven-Haus Bonn (o.J.)

Optimierung des Evaluationsgegenstandes verwendet wird (vgl. Martin Bensen & Andreas Büchter 2005:3).

Die Deutsche Gesellschaft für Evaluation³⁰ – ein Zusammenschluss von Personen und Institutionen, die im Bereich der Evaluation tätig sind – beschreiben Evaluation als

„die systematische Untersuchung des Nutzens oder Wertes eines Gegenstandes. Solche Evaluationsgegenstände können z. B. Programme, Projekte, Produkte, Maßnahmen, Leistungen, Organisationen, Politik, Technologien oder Forschung sein. Die erzielten Ergebnisse, Schlussfolgerungen oder Empfehlungen müssen nachvollziehbar auf empirisch gewonnenen qualitativen und/oder quantitativen Daten beruhen“ (DeGEval 2002:13).

Heinrich Wottawa und Heike Thierau (1990) legen daher Kennzeichen einer *wissenschaftlichen* Evaluation fest:

1. „Evaluation dient als Planungs- und Entscheidungshilfe und hat somit etwas mit der Bewertung von Handlungsalternativen zu tun.“
2. „Evaluation ist ziel- und zweckorientiert. Sie hat primär das Ziel, praktische Maßnahmen zu überprüfen, zu verbessern und über sie zu entscheiden.“
3. Evaluationsmaßnahmen sollten „dem aktuellen Stand wissenschaftlicher Techniken und Forschungsmethoden angepasst sein“ (a.a.O.:13).

Damit ist nicht nur klar geworden, was als Evaluation bezeichnet bzw. nicht bezeichnet werden kann, sondern auch, warum diese überhaupt durchgeführt wird.

Kehren wir noch einmal zu den anfangs genannten Beispielen zurück, so wird deutlich, dass keines von ihnen den Anforderungen einer wissenschaftlichen Evaluation gerecht werden kann. So ist Beethovens „Probehören“ für ihn selbst sicherlich als eine eigene Bewertung seiner Arbeiten zu verstehen, diese ist jedoch keinesfalls für andere nachvollziehbar und entspricht keinem vorab festgelegten Schema. Im zweiten Beispiel – welches durchaus als Evaluation bezeichnet werden kann – wird die Problematik deutlich, mit denen sich Evaluationen auseinandersetzen müssen. Es werden zwar Gegenstandsbereiche, Rahmenbedingungen und Standards festgelegt, dennoch ist die Messung der Daten ungenau, da unterschiedliche PatientInnen nun einmal unterschiedliches Schmerzempfinden haben. Beispiel drei ist wiederum kaum als Evaluation zu bezeichnen, da den ProbandInnen (den Kindergartenkindern) nur eine Alternative vorgelegt wird und die Entscheidung über die Fortsetzung der Feiern im Kindergarten somit nicht getroffen werden kann. Die Evaluation ist also nicht auf den Gegenstand bezogen.

³⁰ Inzwischen umbenannt in „Gesellschaft für Evaluation e.V.“

Es ist also weder ohne weiteres möglich, Evaluation zu definieren, noch diese praktisch umzusetzen:

„Evaluation – more than any science – is what people say it is; and people currently are saying it is many different things“ (Glass & Ellet 1980:211, zit. nach William R. Shadish, Laura C. Leviton, Thomas Cook 1991:30).

Dieses Zitat stellt eine sehr relativistische Position des Evaluationsverständnisses dar. Welche Sichtweisen von Evaluation und welche Definitionsversuche existieren nun eigentlich? Peter Baumgartner (1999a) benennt einige dieser Positionen aus dem wissenschaftlichen Diskurs und strukturiert das eher uneinheitliche Bild:

Evaluation wird zum Einen oftmals mit einer relativistischen Sichtweise aufgefasst, wie ich es zuvor bereits angedeutet habe. Diese Position wird mit dem Fehlen einer Definition bzw. sogar eines Definitionsversuchs begründet und besagt letztlich: „Evaluation ist alles und nichts“ (a.a.O.:65). Die Praxis der Evaluation und die verschiedenen Ansichten dazu seien so vielfältig, dass eine Definition unmöglich gemacht werde (ebd.).

Zum Anderen wird Evaluation als ein quantitatives Analyseverfahren angesehen, also beispielsweise im Sinne der Entwicklung und Durchführung von Tests. Dabei kommt jedoch die ursprüngliche Bedeutung der Evaluation als Bewertung (lat. „valere“ = Bewertung) zu kurz (ebd.).

Drittens wird Evaluation als Methodenlehre begriffen. Peter Baumgartner gibt zu bedenken: „Evaluation bloß als ein Bündel von Methoden zu betrachten, würde es mit untergeordneten Spezialgebieten wie Feldforschung, experimentelles Design oder Psychometrie gleichsetzen“ (a.a.O.:66ff). Auch in diesem Fall wird der Aspekt der Bewertung vernachlässigt.

Evaluation wird häufig auch als eine Verbesserung praktischer Maßnahmen wahrgenommen. Dagegen wendet Peter Baumgartner ein, dass die Zuweisung eines Werturteils zwar sinnvoll ist, konkrete Verbesserungsvorschläge daraus in der Regel jedoch nicht abgeleitet werden können, da das fachliche Themengebiet nicht notwendigerweise dem der EvaluatorInnen entspricht (a.a.O.:68).

Die in der Praxis am häufigsten anzutreffende Position ist die, Evaluation als angewandte Sozialforschung anzusehen. Daran kritisiert Peter Baumgartner (a.a.O.:69), dass damit einerseits Evaluationsfelder, die sich nicht auf (soziale) Interventionsmaßnahmen beziehen, ausgeblendet werden (z.B. Produktevaluation). Ebenso wird mit der Bezeichnung „angewandte“ suggeriert, dass „grundsätzliche philosophische oder ethische Fragestellungen (Erkenntnistheorie, Werttheorie) nicht besonders ausgewiesen, sondern nur mehr unter der Rubrik „Sozialwissenschaft“ behandelt [werden]“ (a.a.O.:69f.).

„Während die Sozialwissenschaften dem Postulat der Wertfreiheit verpflichtet sind, stellt die Zuweisung von Werten in Evaluationen gerade das zentrale Element dar. Während in den Sozialwissenschaften die Problemdefinition und die Verwertung aus dem eigentlichen Forschungszusammenhang ausgeklammert wird, werden Evaluationen (Inhalt, Methodologie und Design) gerade durch die Interessen und Ziele von Auftraggebern und Beteiligten charakterisiert“ (a.a.O.:76).

Er plädiert dafür, Evaluation – ähnlich wie beispielsweise Statistik – als eigene (Hilfs-)wissenschaft anzusehen und nicht als einen Anwendungsfall angewandter Sozialforschung.

Letztlich wird nach Peter Baumgartner Evaluation auch als Bewertung angesehen. Der Evaluation wird dabei sowohl Produkt- als auch Prozesscharakter zugewiesen. Der Autor stimmt dieser Position selbst zu und definiert Evaluation daher so:

„Unter Evaluation sind alle Aktivitäten und/oder Ergebnisse zu verstehen, die die Bedeutung, Verwendbarkeit, (Geld-)Wert, Wichtigkeit, Zweckmäßigkeit,... einer Sache beurteilen bzw. bewerten. Nur dieses weit gefaßte Verständnis von Evaluation kann sowohl die Charakteristika besonderer Evaluationsfelder berücksichtigen als auch einen adäquaten Beitrag zur Theoriebildung leisten (a.a.O.:70)“.

Fassen wir alle genannten Sichtweisen zusammen, handelt es sich bei einer Evaluation also immer um eine Bewertung. Diese geschieht jedoch nicht auf einer alltäglichen Grundlage, sondern muss nachvollziehbar und an bestimmten Rahmenbedingungen orientiert sein. Sagt eine Politikerin beispielsweise: „Die Welt ist schlecht!“, so ist das zwar eine Bewertung, aber längst keine Evaluation, denn die Behauptung fußt nicht auf einer wissenschaftlichen und durch objektive Daten belegbaren Basis und folgt keiner Systematik.

3.2 Typisierung und Ablauf von Evaluationen

An dieser Stelle möchte ich noch auf eine Typisierung von Evaluationen eingehen, die hilfreich ist, das breite Feld zu schematisieren. Diese Typisierung bezieht sich auf den Gegenstandsbereich und das Ziel von Evaluationen. Peter H. Rossi, Howard E. Freeman et al. (1988) unterscheiden drei wesentliche Typen von Evaluationsstudien.

1. „Analysen zur Programmentwicklung incl. Konzeptualisierung und Ausarbeitung einer geplanten Intervention – Planungsevaluierung,
2. Begleitforschung oder Monitoring als laufende Überwachung der Umsetzung und Ausführung eines Programms – Prozessevaluierung und
3. Abschätzung von Programmwirkungen und Programmnutzen – Ergebnisevaluierung oder Nutzen-evaluierung“ (a.a.O, zit. nach Bernhard Bührlen-Armstrong 1998:105).

Die Punkte 1 und 2 stellen also eine formative Evaluation, Punkt 3 eine summative Evaluation dar. Unter einem formativen Vorgehen versteht man die Begleitung eines Prozesses. Die Evaluation ist danach ausgerichtet, diesen Prozess noch während der Durchführung zu optimieren. Eine summative Evaluation wird dagegen erst nach Abschluss der Maßnahme durchgeführt bzw. soll die Maßnahme selbst nicht beeinflussen. Das Ergebnis hat keine Wirkung mehr auf die Maßnahme, dient aber dazu, diese zu resümieren und ggf. bei einer Wiederholung zu verbessern.

Zu unterscheiden sind Evaluationen auch dahingehend, ob sie intern oder extern durchgeführt werden. Intern bedeutet, dass diejenigen, die eine Maßnahme oder ein Produkt initiiert haben, auch gleichzeitig deren Evaluation ausführen. So wird beispielsweise bei der Evaluation von Lehrveranstaltungen von interner Evaluation (auch: Selbstevaluation) gesprochen. Externe Evaluation meint dagegen, dass die EvaluatorenInnen nicht an der Maßnahme beteiligt sind oder waren. Die wissenschaftliche Evaluation einer Unterrichtsmaßnahme durch ein unabhängiges Forschungsinstitut kann also als externe Evaluation bezeichnet werden.

Die unterschiedlichen Arten von Evaluationen sind in der folgenden Abbildung dargestellt:

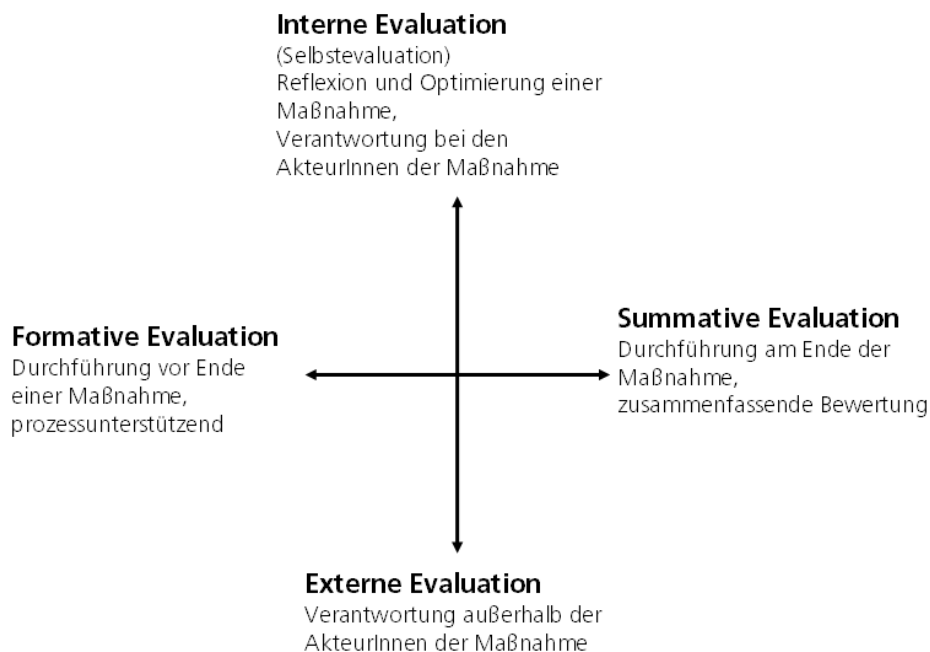


Abbildung 1: Formen von Evaluationen
(angepasst nach Martin Bonsen & Andreas Büchter 2005:5)

So wie es in Bezug auf die Definition von Evaluation keine einheitliche Meinung gibt, wird auch der Ablauf von Evaluationen unterschiedlich gesehen. Verbreitet zu sein scheint die Ablauflogik nach Michael Scriven (1980 und 1991):

1. „Formulierung von Wertkriterien: In der ersten Phase werden jene Kriterien ausgewählt und definiert, die der Evaluand erfüllen muß, um als gut, wertvoll etc. gelten zu können.
2. Formulierung von Leistungsstandards: Für jedes einzelne Kriterium wird eine Norm definiert, die der Evaluand erreichen muß, damit das Kriterium als erfüllt angesehen werden kann (Operationalisierung).
3. Messung und Vergleich (Analyse): Nun wird jedes Kriterium beim Evaluanden untersucht, gemessen und mit den jeweils vorgegebenen Leistungsstandards verglichen.
4. Werturteil (Synthese): In dieser letzten und wohl schwierigsten Phase von Evaluationen müssen die verschiedenen Ergebnisse zu einem einheitlichen Werturteil integriert werden“ (a.a.O., zit. nach Peter Baumgartner 1999a:76).

Andere AutorInnen wie beispielsweise Hans-Günter Rolff (2001:93) sehen den Evaluationsprozess eher als Kreislauf. Dies ist dann sinnvoll, wenn eine Verbesserung von Maßnahmen oder Produkten angestrebt wird.

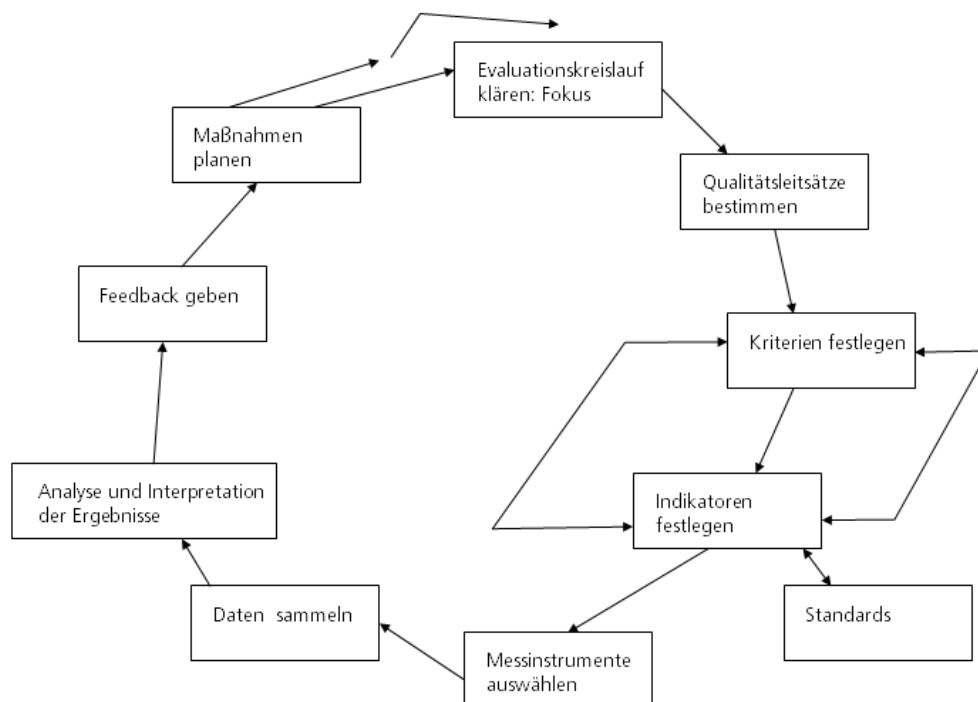


Abbildung 2: Evaluationskreislauf (angepasst nach Hans-Günther Rolff 2001:93)

Zusammenfassend lege ich dieser Arbeit nunmehr folgendes Verständnis von Evaluation zu Grunde und schließe mich damit Martin Bonsen & Andreas Büchter (vgl. 2005:4ff) an:

Evaluation kann sich demnach auf unterschiedliche Gegenstände beziehen, z.B. Projekte, Produkte, Maßnahmen, Leistungen, Organisationen, Technologien usw.. Evaluation kann dabei verschiedene Zwecksetzungen verfolgen (z.B. Vorbereitung von Entscheidungen, Verbesserung von Organisationen oder Programmen, Unterstützung von Lern- und Reflexionsprozessen). Evaluation kann in unterschiedlichen Phasen der Entwicklung eines Evaluationsgegenstandes ansetzen und beispielsweise entwickelnden, gestaltenden oder bilanzierenden Charakter haben.

3.3 Vergleichendes Evaluationsdesign

In Evaluationen werden in der Regel auch die Erfolge und Misserfolge bestimmter Maßnahmen bewertet. Um dies zu messen, werden oft vergleichende Studien durchgeführt – ein Lernsetting (beispielsweise ein Online-Kurs) wird mit einem anderen Lernsetting (beispielsweise einem Präsenzkurs) verglichen. Je nach Abschneiden der Lernenden wird entweder das eine oder das andere Setting als erfolgreich angesehen. Im Rahmen dieser Arbeit sind besonders solche Studien von Interesse, die sich mit dem Zusammenhang von Lernerfolg und Lernmethode beschäftigen.

Oftmals sind solche Untersuchungen nach (quasi-)experimentellem Schema aufgebaut: Die zu evaluierende Maßnahme wird an einer Gruppe getestet, während eine andere Gruppe dieser Maßnahme nicht ausgesetzt wird³¹. Dieses Evaluationsdesign zählt zu den wichtigsten Versuchsanordnungen: Es zielt darauf ab,

„kausale Zusammenhänge zwischen einer oder mehreren unabhängigen und einer abhängigen Variablen unter Konstanzhaltung aller übrigen Einflussfaktoren festzustellen. Ein kausaler Zusammenhang zwischen einer unabhängigen und einer abhängigen Variablen ist dann gegeben, wenn sie in einem empirisch nicht umkehrbaren, asymmetrischen Zusammenhang stehen: wenn X, dann Y, aber nicht umgekehrt.“ (Reinhard Wittenberg 1999:77).

In diesem Fall würde also beispielsweise eine Gruppe von Lernenden in einem Online-Kurs mit einer Gruppe Lernender eines Präsenzkurses verglichen. Die Maßnahme der Untersuchungsgruppe stellt die Nutzung des Internets dar, während die andere Gruppe – die nach üblichem Schema arbeitet – als Kontrollgruppe dient.

Die Wirkung der Maßnahme kann nur nachgewiesen werden, wenn in der Vergleichsgruppe tatsächlich keine Seiteneffekte auftreten (wie zum Beispiel erhöhter Lernerfolg o.ä.). Sollte dies der Fall sein, können die auftretenden Effekte folglich nicht auf die Maßnahme zurückzuführen sein. Um also ein Ursache-Wirkungs-Prinzip belegen zu können, müssen Störeinflüsse möglichst gering gehalten werden, was in

³¹ Experimentelle Designs machen eine randomisierte Zuteilung zu Test- und Kontrollgruppe erforderlich.

der Praxis oft nicht realistisch leistbar ist, da die interne Validität in der Feldforschung meist gering ist (vgl. Annabell Preußler i.Vorb.).

Einen Zusammenhang zwischen zwei Variablen bezeichnet man als Korrelation, also beispielsweise: „Je häufiger Schülerinnen mit dieser Software arbeiten, desto höher ist der Lernerfolg.“ Korrelation ist jedoch nicht mit Kausalität gleichzusetzen, obwohl dies immer wieder angetroffen wird:

„Hat man zwischen zwei Variablen x und y eine statistisch abgesicherte, d.h. signifikante Korrelation gefunden, kann diese Korrelation im kausalen Sinne folgendermaßen interpretiert werden:

1. x beeinflusst y kausal,
2. y beeinflusst x kausal,
3. x und y werden von einer dritten oder weiteren Variablen kausal beeinflusst,
4. x und y beeinflussen sich wechselseitig kausal.

Der Korrelationskoeffizient liefert keine Informationen darüber, welche der 4 Interpretationen richtig ist. [...] Eine Korrelation zwischen zwei Variablen ist eine notwendige, aber keine hinreichende Voraussetzung für kausale Abhängigkeiten. [...] Kausalität läßt sich korrelationsstatistisch nur widerlegen, aber nicht eindeutig bestätigen“ (Jürgen Bortz 1993:217).

Auf die Kritik an Verfahren der Leistungsmessung werde ich im Kapitel „Lernerfolg und Leistungsmessung “ genauer eingehen.

4 Lernqualität und Lernerfolg

Wann erreicht das Lernen eine hohe Qualität? Ist diese beim E-Learning höher als bei traditionellem Lernen?

DIE Lernqualität gibt es nicht und kann es auch nicht geben. Qualität ist ein Konstrukt, das sich aus verschiedenen Dimensionen zusammensetzt. So kann auch Lernqualität vielfältige Bedeutungen haben. Einerseits kann sie sich auf die Qualität des Lernergebnisses, aber auch die Qualität des Lernprozesses oder die der Behaltensleistung beziehen.

„Konstrukte [...] unterscheiden sich von *Indikatoren* (beobachtbare Merkmale) vor allem dadurch, dass sie nicht ohne weiteres messbar sind. Konstrukte sind theoretische Größen, die innerhalb von Theoriekontexten sinnvoll sind, deren empirische Existenz jedoch nur behauptet werden kann. Für die empirische Forschung ergibt sich daraus die Notwendigkeit, Konstrukte operationalisieren zu müssen, wobei die Operationalisierungen wiederum nur einen hypothetischen Charakter haben, sich also als falsch erweisen können“ (Ute Clement & Bernd Martens 2000:106; Hervorhebungen im Original).

Dennoch sollen im Folgenden definitorische Gedanken aufgegriffen werden, bevor der Zusammenhang von Lernqualität und Lernerfolg deutlich gemacht wird.

4.1 Das Konstrukt Qualität

Um Lernqualität bestimmen zu können, muss vorab geklärt werden, was generell unter Qualität zu verstehen ist. Zunächst einmal leitet sich der Begriff von dem Lateinischen „qualitas“ ab, was soviel wie Beschaffenheit oder Eigenschaft bedeutet. Darüber hinaus ist „Qualität“ in der Umgangssprache mit Wertungen verbunden und bringt „die Zweckangemessenheit eines Ergebnisses menschlicher Arbeitstätigkeit bzw. der Beschaffenheit eines Produktes oder einer Dienstleistung zum Ausdruck“ (Wikipedia³²). Am Ende des Eintrags in der Wikipedia findet sich eine einfache, aber dennoch sehr einleuchtende Faustformel für die Definition von Qualität: „Qualität ist die Übereinstimmung von Ist und Soll“ (ebd.).

Was zunächst plausibel klingt, ist m.E. jedoch zu einfach, denn was Qualität tatsächlich ausmacht, lässt sich daraus nicht ableiten. Die Schwierigkeit ist darin zu sehen, dass Qualität sich nicht nur aus einer Perspektive zusammensetzt, sondern – wie bereits am erwähnt – ein Konstrukt darstellt, das aus dem Zusammenspiel vieler Faktoren entsteht (vgl. Ulf-Daniel Ehlers 2004:25, Annabell Preussler & Peter Baumgartner 2006:74). Selbst im wissenschaftlichen Diskurs dürfte es schwierig sein, zu einem Konsens über die Definition von Qualität zu gelangen. Ulf-Daniel Ehlers (a.a.O.:51) merkt an, dass noch keine übereinstimmende Meinung darüber herrscht, „wie etwa Bildungsqualität definiert ist“.

³² <http://de.wikipedia.org/w/index.php?title=Qualit%C3%A4t&oldid=37442789>, letzter Aufruf: 16.10.2007)

Was bleibt uns also? Wir bewegen uns nicht im definitionsleeren Raum, sondern können zumindest Eigenschaften beschreiben, die Qualität trotz dieser relativistischen Darstellung charakterisieren.

Fest steht: Qualität ist schwer zu standardisieren (vgl. Lee Harvey & Diana Green 2000). Sie sagt zwar etwas über die Beschaffenheit eines Gegenstandes aus, setzt sich aber nicht mit dem Gegenstand selbst auseinander (vgl. Ulf-Daniel Ehlers 2004:56). Mit anderen Worten: Qualität lässt sich nicht durch den Inhalt eines Produkts beschreiben, sondern entsteht erst durch das Zusammenwirken verschiedener Dimensionen. Beispielsweise ist die Qualität eines Restaurants nicht nur durch dessen Inneneinrichtung gekennzeichnet, sondern vor allem auch durch die Güte der Speisen, deren Preis, den Service, die Freundlichkeit des Personals und so weiter.

Diese generellen Eigenschaften von Qualität lassen sich auch auf die Lernqualität im Bildungsbereich übertragen. So ist ein Faktor von Lernqualität auch immer der/die Lernende selbst. Qualität kann nur und erst dann entstehen, wenn die Lernenden mit der Lernsituation in Interaktion treten. Lernqualität entsteht folglich erst, wenn auch tatsächlich gelernt wird (vgl. Gerhard M. Zimmer & Elena Psaralidis 2000:265; Ulf-Daniel Ehlers 2002:7ff). Im Zuge dieser Entwicklung muss auch der Qualitätsbegriff erweitert werden, so dass er die Rahmenbedingungen des Lernens und auch die Lernenden selbst einbezieht (Sigmar-Olaf Tergan 2003). Lee Harvey & Diana Green (2000:31) beschreiben „Qualität als Transformation“ damit ist – dieser Argumentation folgend – gemeint, dass sich die Qualität aus dem Verhältnis zwischen Lernangebot und Lernendem als „Weiterentwicklung des Lernenden durch einen Lernprozess“ konstruiert (Ulf-Daniel Ehlers 2002:16, vgl. auch Annabell Preussler & Peter Baumgartner 2006:75).

4.2 Lernerfolg als Indikator für (Lern-)Qualität

Seit dem Aufkommen des E-Learnings wurden eine Reihe von Qualitätskonzepten entwickelt, die allerdings ohne große Erfolge blieben, da

„ein auf alle Lerner gleichermaßen übertragbarer Zusammenhang zwischen didaktischen oder medialen Attributen eines Lernsystems [...] nach derzeitigem Stand nicht nachgewiesen werden [konnte]“
(Ulf-Daniel Ehlers 2004:24, vgl. auch Bernd Weidenmann 1997).

Für die Messung und Sicherung von (Lern-)Qualität ist Evaluation also ein entscheidender Faktor, da dadurch der Versuch unternommen wird, sowohl Lernangebote als auch Lernergebnisse an Hand bestimmter Kriterien und im Hinblick auf die Lernziele zu messen und zu bewerten. Dabei stellt der Lernerfolg oftmals eines dieser Kriterien dar und dient in sozialwissenschaftlichen Studien sogar als ein Indikator für Lernqualität.

Die Messung des Lernerfolgs ist demnach ein beliebter Weg, um Lernqualität zu belegen. So geht man davon aus, dass mit einer Software oder einem Bildungsangebot umso bessere Lernerfolge erzielt werden

können, je höher die Qualität dieser Software ist. Plant man beispielsweise die Anschaffung eines Lernprogramms, überprüft man es in der Regel auf ein Qualitätssiegel oder eine Testnote (siehe Abbildung 3).



Abbildung 3: Beispiel für ein Testsiegel auf einer Software (ibis acam GmbH 2004).

Ein größerer Lernerfolg wird von dem Produkt erhofft, das eine höhere Qualität verspricht. Umgekehrt wird aber auch die Qualität einer Software darin gemessen, wie gut überhaupt mit ihr gelernt werden kann, also wie hoch der erzielte Lernerfolg ist. Wer also mit der Evaluation einer Software beauftragt ist, wird – um Aussagen zur Qualität treffen zu können – den Lernerfolg messen (vgl. Annabell Preussler & Peter Baumgartner 2006:76). Die Qualität wird somit als Voraussetzung für den Lernerfolg gesehen (vgl. Reiner Fricke, 2000).

Wie bereits an anderer Stelle dargestellt wurde (vgl. Annabell Preussler & Peter Baumgartner 2006:82), kann Qualität von E-Learning aus zwei unterschiedlichen Richtungen betrachtet werden: Einerseits kann sie sich auf die Qualität des Lehrmaterials, andererseits auf die Qualität des Lernergebnisses beziehen. Im ersten Fall könnte die Evaluation z.B. darin bestehen, dass das Material von ExpertInnen analysiert und bewertet wird. Der zweite Fall macht eine Evaluierung des (abgeschlossenen) Lehr- und Lernvorganges notwendig. Auffällig ist, dass diesen beiden Evaluationsformen unterschiedliche Methodiken zu Grunde liegen, die zu unterschiedlichen Ergebnissen führen können. So kann ein Kurs nach Sichtung seines Materials sehr gut abschneiden, in der Praxis jedoch sehr schlechte Lernergebnisse hervorbringen und ein rudimentärer Kurs kann – vielleicht gerade weil hier besonders viel Eigenleistung erforderlich ist – besonders gute Ergebnisse liefern (vgl. Peter Baumgartner & Annabell Preussler 2005:6).

Sigmar-Olaf Tergan & Peter Schenkel (2004:5) haben m.E. jedoch Recht, wenn sie feststellen, dass „der Zusammenhang zwischen einem qualitativ hochwertigen Lernangebot und großem Lernerfolg [...] nicht

linear [ist].“ Die Übertragung von Wissen über ein Medium auf einen Lernenden ist seiner Auffassung nach ausgeschlossen, da es sich beim Lernen – egal, was darunter im Einzelnen verstanden wird – immer um eine Leistung einer Person handelt. Die Auseinandersetzung der Lernenden mit dem Wissen in einer bestimmten Lernumgebung ermöglicht letztlich den Lernerfolg, aber nie das Medium allein. Daher sind gute Lernangebote keine Garantie für den Lernerfolg (vgl. ebd.). Obwohl es also zwischen der Qualität des Angebotes und der Qualität des Lernergebnisses einen Zusammenhang gibt, so ist meiner Auffassung nach zu betonen, dass das, was wirklich zählt, die Qualität des Lernergebnisses – also der Lernerfolg – ist.

4.3 Lernerfolg – Definitionsversuche

Was macht Lernen erfolgreich und wie kann man das feststellen?
Was genau wird nun also unter Lernerfolg verstanden?

Michael Kerres (2001:111) bezeichnet Lernerfolg als das „Ergebnis didaktischer Aktivitäten“, aber dabei nicht als das bloße „Behalten von Fakten, Ereignissen oder Vorgängen“. Bei einer solchen Definition würden seiner Ansicht nach beispielsweise die Persönlichkeit der Lernenden oder ihre Denkschemata außer Acht gelassen (ebd.). Ebenfalls fällt auf, dass der Lerntransfer nicht ausreichend beachtet wird – also ob das Gelernte auch praktisch umgesetzt werden kann und der Erfolg in einer Lernsituation auf den Alltag übertragbar ist (vgl. Annabell Preussler & Peter Baumgartner 2006:76). In der Praxis wird es sich m.E. mit der Beschreibung von Lernerfolg oft etwas zu einfach gemacht, wenn die Notengebung nur auf die Beherrschung des Lehrstoffes abzielt. Reicht diese Beschreibung aus? Was ist das geforderte Ausmaß und vor allem, wie wird der Lernerfolg gemessen? Ein Beispiel soll diesen Aspekt näher beleuchten:

„Ein Schüler, der sich auf den Unterricht konzentriert, Mitschriften anfertigt, bei Unklarheiten stutzig wird und entsprechend nachfragt, der am Nachmittag liest oder mit Freunden den Unterrichtsstoff diskutiert, der sich Unklares von Eltern oder älteren Geschwistern erklären lässt, der Textstellen unterstreicht, sich durch stilles Wiederholen wichtige Informationen einprägt und der sich vergewissert, dass er auch wirklich den Stoff für die nächste Klassenarbeit beherrscht, dessen Leistung sollte einer guten Note entsprechen“ (Nadine Spörer 2003:69).

Es wird an diesem Zitat deutlich, dass eine große Schwierigkeit darin besteht, Lernerfolg an Hand einzelner Kriterien festzumachen. Könnte nicht eine Schülerin, die den Lernstoff bloß auswendig gelernt wiedergibt, sogar eine bessere Note erzielen als der im Zitat erwähnte Schüler? Der festgestellte Lernerfolg wäre bei ihr u.U. also sogar höher als bei dem Schüler, der sich tatsächlich mit seinem Lernstoff auseinandersetzt. Lernerfolg ist m.E. also nicht unmittelbar messbar, sondern nur die „Performance“, also das Zeigen einer erwarteten Leistung. Das bedeutet aber nicht, dass diese Leistung auch beherrscht wird. Obwohl also bloße Reproduktion von vorhandenem Wissen offenbar nicht allein dazu dienen kann, den Lernerfolg festzustellen, orientiert sich dessen Messung in diesem Beispiel, wie auch in vielen Evaluationsstudien,

oftmals an standardisierten Wissensabfragen (vgl. Annabell Preussler & Peter Baumgartner 2006:76). Peter Baumgartner (1999:206) bezeichnet diese Verwechslung von Denkprozessen und Denkprodukten als einen Kategorienfehler. Ein solcher werde dann begangen, „wenn Begriffe auf unterschiedlichen Ebenen miteinander verglichen bzw. in Beziehung gesetzt werden“ (ebd.). Er führt ein anschauliches Beispiel von Gilbert Ryle (1969) dafür an:

„Einem Besucher werden alle Gebäude der Universität (Hörsäle, Rektorat, Dekanat, Mensa, Studienberatung,...) gezeigt. Nachdem der Besucher alles eingehend betrachtet, besucht und studiert hat, fragt er uns: ‚Schön ich habe jetzt viele Gebäude und Räumlichkeiten gesehen, aber wo ist die Universität?‘“ (a.a.O., zit. nach Peter Baumgartner 1999:206).

Diese Frage macht deutlich, wie fälschlicherweise zwei verschiedene Ebenen miteinander in Beziehung gesetzt werden. Ähnlich verhält es sich bei der Beschreibung von prozeduralem Wissen und Kompetenzen bzw. Fertigkeiten. „Das ‚Wissen, wie‘ ist immer noch (ähnlich, wie das ‚Wissen, daß‘ etwas der Fall ist) grundsätzlich ein theoretisches Wissen und keine praktische Fertigkeit“ (ebd.). Wenn ich weiß, wie etwas gemacht wird, heißt das längst nicht, dass ich es auch kann. Daher ist es problematisch, die Messung von Lernerfolg rein von Wissensabfragen abhängig zu machen.

Eine weitere Schwierigkeit bei der Feststellung von Lernerfolg stellen die Anforderungen an modernes Lernen dar, wie Ute Clement & Bernd Martens (vgl. 2000:98) feststellen. Auf der einen Seite werden zwar Interaktivität und ein hohes Maß an Selbststeuerung beim Lernen gefordert, andererseits soll der Lernerfolg aber dennoch kontrollierbar und möglichst vorhersagbar sein. Überdies wird Wissen der konstruktivistischen Sichtweise und dem Ansatz des situierten Lernens zufolge ständig in sozialen Kontexten konstruiert (vgl. Michael Kerres 2001:111) und also nicht im Kopf „abgespeichert.“ So werden demzufolge auch Evaluationsansätze unbrauchbar, die auf das Individuum ausgerichtet sind. Es sei lediglich möglich zu prüfen, inwieweit die Lernangebote soziale Settings beeinflussen und die Wissenskonstruktion innerhalb dieser Settings unterstützen (ebd.).

Eine Umsetzung dessen erscheint in der Forschungspraxis allerdings schwierig. So ist es nicht verwunderlich, dass Lernerfolg oft mit Behaltensleistung gleichgesetzt wird (vgl. a.a.O.:112). Michael Kerres zufolge sollte es hierbei jedoch um andere Leistungsdimensionen gehen – dabei müssten kompetenzorientierte Ansätze und die theoretischen Diskussion des didaktischen Designs berücksichtigt werden:

„statt *Wissen*, wird über das *Verstehen*, *Anwenden* und *Handeln* diskutiert. Es geht um (z.B. metakognitives oder metamotivationales) strategisches Wissen, um Selbstregulation und Planungskompetenz“ (ebd., Hervorhebungen im Original).

Abhängig sei der Lernerfolg davon, „inwieweit es geling[e], die angestrebten kognitiven und emotionalen Lernprozesse anzuregen“. Lernangebote sind demzufolge „so aufzubereiten, dass ihre Bearbeitung

bestimmte Lernprozesse möglich machen.“ Dabei sei – will man Aussagen zum Funktionieren von mediengestütztem Lernen treffen – überdies zu beachten, dass Medien Angebote sind, die Lernende auf verschiedene Arten aus der sozialen Umwelt heraus nutzen. Diese regen unterschiedliche Aktivitäten an. Darüber hinaus sei es denkbar, dass Lernende das Lernangebot nicht in der gewünschten Weise nutzen. Daher müsse geprüft werden, wie sie mit den Lernangeboten umgingen (a.a.O.:146 f.)



Abbildung 4: Bildungsmedien als Angebote zur Anregung von Lernprozessen (nach Michael Kerres 2001)

Abbildung 4 stellt diesen Prozess dar. An allen seinen Übergängen erfolgt eine Re-Konstruktion des Angebots bzw. des angebotenen Materials. Hier sei zu prüfen, inwieweit von dem erwünschten Prozess abgewichen werde (ebd.). Michael Kerres schließt mit diesem Modell – abweichend von früheren Ansätzen – bereits die konstruktivistische Vorstellung ein. Dennoch bleiben Kerres’ Fragen für mögliche Evaluationen: „Was soll mit diesem Lernprozess letztlich für ein Lernziel verbunden sein?“ und „Wie kann sichergestellt werden, dass die Lernenden das Angebot tatsächlich annehmen, und der Lernprozess tatsächlich stattfindet?“ (ebd.).

Auch Sigmar-Olaf Tergan (2004) kritisiert, dass in vielen Evaluationsansätzen davon ausgegangen werde, dass ein Lernangebot, welches entsprechend den Vorgaben eines Instruktionsdesign-Modells gestaltet wurde und den Kriterien und Standards von Evaluation entspricht, zu entsprechendem Lernerfolg führe. Diese Vorstellung sei deshalb falsch – und damit bezieht sich Tergan auf Gerhard M. Zimmer & Elena Psaralidis (2000:264), weil es „keinen kausalen Zusammenhang zwischen objektiven Merkmalen eines Lernangebotes und den Lernergebnissen [gäbe]“ (Gerhard M. Zimmer & Elena Psaralidis 2000:264). Was Lernen erfolgreich macht (oder eben nicht), sind die kognitiven Aktivitäten der Lernenden, gepaart mit der Motivation. Für sie stellen Medien nur Informations- bzw. Lernangebotsträger dar, aber die Rezeption erfolgt selbstgesteuert und aktiv (vgl. Sigmar-Olaf Tergan 2004:17).

Eine Schwierigkeit bei der Messung der Lernqualität besteht also darin, dass auch der Lernerfolg als ein Konstrukt bezeichnet werden muss. Dies macht empirische Forschung ungleich schwieriger, denn damit werden zwei Konstrukte miteinander in Verbindung gesetzt, deren Verallgemeinbarkeit aber nicht gegeben ist. Es bleibt also zu prüfen, wie der Lernerfolg in der Praxis verstanden wird.

4.4 Dimensionen von Lernerfolg

Um das Konstrukt Lernerfolg zu kennzeichnen, stellt Michael Kerres (2001:112) einige Kriterien aus dem Kontext des mediengestützten Lernens heraus. Dazu gehören beispielsweise die erlebte Qualität des Lernangebotes, die emotionale Reaktion und Lernmotivation, das Lernverhalten, die subjektive Zufriedenheit mit dem Lernverhalten/-ergebnis, die faktische Nutzung und Akzeptanz, das erzielte Kosten-Nutzen-Verhältnis oder auch strukturelle Implikation und Veränderungen institutioneller und gesellschaftlicher Organisation von Bildung. Darüber hinaus nennt Kerres auch den „objektiven“ Lernerfolg. Er bezieht sich damit auf die Messung von Lernerfolg „in verschiedenen zeitlichen Abständen und in unterschiedlicher Anwendungsnähe (Erinnern, Anwenden, Transfer)“ (vgl. ebd.). Wie bereits dargestellt wurde, ist „objektiver“ Lernerfolg allerdings nur schwer messbar. Es bleibt die Prüfung dessen, was Lernen zu einem erfolgreichen Prozess macht. Im Rahmen dieser Arbeit wird Lernerfolg als ein übergreifendes Konstrukt verstanden, welches im Folgenden an Hand von Dimensionen konkretisiert werden soll.

4.4.1 Modell der lernrelevanten Kontexte und Komponenten nach Sigmar-Olaf Tergan (2004)

Um zu zeigen, wie das Konstrukt Lernerfolg in einzelne Dimensionen gegliedert werden kann, möchte ich beispielhaft auf das Modell der lernrelevanten Kontexte und Komponenten nach Sigmar-Olaf Tergan (2004) eingehen.

Dieser schlüsselt das Konstrukt zunächst auf, indem er *Lernende* und *Lernszenarien* unterscheidet. Lernerfolg kann sich demnach auf die *Lernaktivitäten* und *Lernprozesse* der Lernenden beziehen. Dabei sind kognitive und psychomotorische Lernaktivitäten zu unterscheiden. Ebenso wird zwischen kognitiven und prozeduralen Lernprozessen differenziert. Darüber hinaus tragen auch die Kontexte und Komponenten der Lernszenarien zum Lernerfolg bei.

„Die Frage danach, was Lernen erfolgreich macht, bezieht sich auf die Lernaktivitäten (kognitive, psychomotorische) und Lernprozesse (kognitive, prozedurale) der Lernenden sowie die Komponenten und Kontexte von Lernszenarien, die dazu beitragen, entsprechende Aktivitäten und Prozesse auszulösen, zu unterstützen und zu fördern“ (a.a.O.:17).

Abbildung 5 stellt diese Überlegungen in einer Übersicht dar.

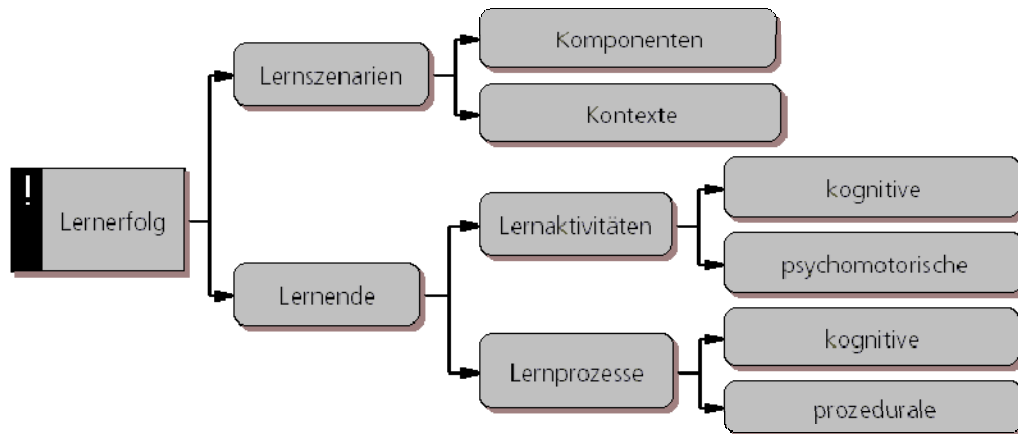


Abbildung 5: Dimensionen von Lernerfolg
(erstellt nach Sigmar-Olaf Tergan 2004:17)

Tergan geht dabei von der Annahme aus, dass sich unterschiedliche Lernbedingungen für unterschiedliche Lernziele als förderlich bzw. hinderlich erweisen, d.h. effektive Lernprozesse nur unter bestimmten Bedingungen stattfinden. Solche Lernziele könnten z.B. der Erwerb von psychomotorischen Fertigkeiten, Sachwissen, kognitive Kompetenz oder praktisch anwendbares Wissen sein. Er betont, dass eine konkrete Lernsituation immer die spezielle Situation der Lernenden einbeziehen muss. Es sei jedoch schwierig, einzelne Bedingungen erfolgreichen Lernens zu ermitteln, da sie nur in Wechselwirkung mit den Lernenden zum Tragen kommen (vgl.a.a.O.:18f.).

Ein Modell soll daher die lernrelevanten Kontexte und Komponenten technologiebasierter Lernszenarien verdeutlichen (vgl. Abbildung 6). Dies trägt dazu bei, die einzelnen Konstruktdimensionen deutlicher herauszustellen.

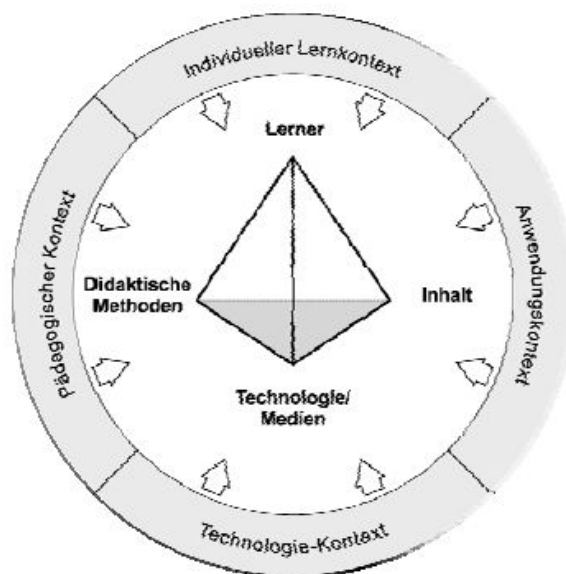


Abbildung 6: Lernrelevante Kontexte und Komponenten technologiebasierter Lernszenarien (Sigmar-Olaf Tergan 2004:18)

In diesem Modell werden vier Kontexte bzw. Dimensionen unterschieden:

1. Individueller Lernkontext,
2. Anwendungskontext,
3. Pädagogischer Kontext,
4. Technologie-Kontext.

Diese Dimensionen bilden die Rahmenbedingungen für die entsprechenden Komponenten auf der Ebene der Lernszenarien: die Lernenden, den Inhalt, die Technologie und die didaktischen Methoden (vgl. ebd.). Ich möchte im Folgenden skizzenhaft auf die einzelnen Dimensionen eingehen und dabei auch die relevanten Komponenten kurz darstellen.

Der individuelle Lernkontext ist auf Rahmenbedingungen und Hintergründe bezogen, die Einfluss auf das Lernen ausüben (können). Tergan nennt hier beispielsweise den Bildungsabschluss und die Berufserfahrung, Zugangsmöglichkeit zu Technologien, Lernort, finanzielle Ressourcen usw. Dies bezieht er nun auf die Komponente der einzelnen Lernenden und stellt lernrelevante Merkmale heraus. Dazu gehören beispielsweise Persönlichkeitsmerkmale – Tergan nennt hier Kennzeichen wie Visualisierer - Verbalisierer, Lernstrategie-Typ usw., kognitive Merkmale (z.B. Vorwissen oder Medienkompetenz), Merkmale individueller Emotion (z.B. Motivation am Lerninhalt; Akzeptanz des Lernangebotes), Merkmale individueller Interessen (z.B. Allgemeines Interesse an Fort- und Weiterbildung; gezielte Prüfungsvorbereitung) und letztlich auch soziale Merkmale wie Teamfähigkeit, Kommunikationsfähigkeit usw.

Tergan stellt explizit dar, dass Lernerfolg nur erwartet werden kann, wenn die Voraussetzungen auf Seiten der Lernenden auch den Anforderungen des Lernangebotes entsprechen (vgl. ebd.).

Mit dem Anwendungskontext ist vor allem die Anwendung des Wissens gemeint. Tergan nennt hier vor allem zwei Kennzeichen: Ort und Situation der Wissensanwendung sowie kognitive, emotionale und soziale Anforderungen bei praktischen Aufgabenstellungen. Es sei daher von hoher Wichtigkeit, die allgemeinen Anforderungen gründlich zu analysieren, bevor ein Lernangebot erschaffen werde. Komponenten des Lerngegenstandes bzw. Lerninhalts stellen in dieser Dimension vor allem die Authentizität eines Lernangebotes, sachliche Korrektheit, Art der Inhalte (z.B. didaktisch aufbereiteter oder abstrakter Text usw.), Kodierungsformen und angesprochene Sinnesmodalitäten dar. Die kognitiven Anforderungen sind vor allem für die Ziele des Lernens relevant. Zusammengefasst stellt Tergan für diese Dimension fest, dass ein Lerngegenstand dann zur Lernförderung beitragen könne,

„wenn er den Zielen des Lernangebotes und den Voraussetzungen auf Seiten der Lernenden entspricht, durch die Technik und die verwendeten Medien angemessen dargestellt und interaktiv zugänglich ist und seine aktive kognitive Verarbeitung durch die

realisierten didaktischen Methoden unterstützt und gefördert wird“ (a.a.O.:19).

Die dritte Dimension bezieht sich auf den pädagogischen Kontext und meint vor allem die pädagogischen und psychologischen theoretischen Annahmen zur Lernförderung sowie das Modell des Instruktionsdesigns, auf dem ein Lernangebot basiert. Die Komponente der didaktischen Methoden stellt insbesondere die konkreten Maßnahmen dar, die erforderlich sind, didaktische Überlegungen in die Gestaltung eines Lernangebotes zu implementieren. Dazu gehören Methoden zur Aufbereitung, Mediendidaktik und Maßnahmen zur Lernunterstützung. Diese tragen dann zu erfolgreichem Lernen bei, wenn sie die Voraussetzungen der Lernenden berücksichtigen und sie zu einer aktiven kognitiven und konstruktiven Verarbeitung von Informationen anregen. Darüber hinaus sollten sie den Zielen des Lernens und den entsprechenden Anforderungen an Gegenstandsbereiche und Technologien entsprechen (vgl. a.a.O.:20).

Die letzte Dimension bezieht sich auf den Technologie-Kontext. Damit sind zum einen der aktuelle Stand der Informations- und Kommunikationstechnologien sowie zum anderen die technischen Rahmenbedingungen gemeint, die die Medienverwendung in einem Lernsetting fördern oder behindern. Die in dieser Dimension relevante Komponente Technologie und Medien betrifft Merkmale der Medien und Technologien, die in einem Lernangebot verwendet werden. Dazu gehören Merkmale der medialen Darstellung von Inhalten, Merkmale der Lerner-System-Interaktion sowie Merkmale der medialen Umsetzung didaktischer Methoden. Laut Tergan tragen Technologie und Medien in gleicher Weise zum Lernerfolg bei, wie bei der Dimension des pädagogischen Kontexts, nämlich genau dann,

„wenn sie den Voraussetzungen auf Seiten der Lernenden entsprechen, die Lernenden bei der aktiven und konstruktiven Verarbeitung von Informationen bei der Bewältigung von Aufgabenstellungen unterstützen und den Anforderungen des Gegenstandsbereichs gerecht werden“ (a.a.O.:21).

Betrachtet man Tergans Modell als Ganzes, stellt man fest, dass nicht eine einzelne Dimension oder Komponente, sondern nur das „Gefüge von Rahmenbedingungen“ (a.a.O.:21) zur Bestimmung des Lernerfolgs beiträgt. Dabei ist m.E. eine vollständige explizite Isolierung und Beschreibung aller Faktoren erfolgreichen Lernens nicht möglich. Dies liegt in mancher Hinsicht an der Komplexität dieses Konstrukts. Der entscheidende Punkt ist hier jedoch, dass das Gefüge durch mehr Aspekte bestimmt wird als sich nur durch die Auflistung seiner Dimensionen ausdrücken lässt. Vielmehr sind die einzelnen Faktoren innerhalb des Gefüges stark miteinander verwoben.

Da diese Betrachtungsweise eine Menge von empirischen Herausforderungen birgt, wird in der Forschung versucht, einzelne Aspekte dieses Gefüges – beispielsweise die Motivation – zu isolieren und einzeln zu betrachten. Vor dem Hintergrund, dass alle Dimensionen zusätzlich miteinander in Beziehung stehen, ist dieses Verfahren jedoch nicht möglich. Würde man einen Teil des Gefüges herausnehmen, ist eine ganzheitliche Betrachtung nicht mehr möglich. Hierbei sind nicht die

einzelnen Dimensionen relevant, sondern deren Relationen zueinander. Diese Dynamik lässt sich jedoch wiederum nur im Gesamtzusammenhang begreifen. Um es mit den Worten Peter Baumgartners auszudrücken:

„Statt allgemein von Lernerfolg zu sprechen - und dabei meistens sprachlich formulierbares Wissen zu meinen, müssen bei der Evaluation interaktiver Medien vor allem die durch die Interaktivität angestoßenen Lernprozesse evaluiert werden. Diese lassen sich jedoch meistens gerade nicht sprachlich erfassen“ (Peter Baumgartner 1999a:93).

Dass es also DEN Lernerfolg nicht geben kann, wird auch durch Tergans Modell deutlich. Diese Überlegungen werden im an späterer Stelle dieser Arbeit erneut aufgegriffen, wenn es darum geht, Lernziele und Lernleistungen miteinander in Beziehung zu setzen.

4.4.2 (ökonomische) Effizienz mediengestützten Lernens

Der Vollständigkeit wegen möchte ich an dieser Stelle noch den Aspekt der Effizienz mediengestützten Lernens erwähnen, auf den auch Michael Kerres (2001:13) in seiner Darstellung verweist.

„Die Effektivität drückt das Ausmaß der Zielerreichung unabhängig vom Einsatz von Ressourcen aus. In der Bildungspraxis steht jedoch die Frage der Effizienz von Bildungsmedien im Vordergrund, bei der Effekte auf den Mitteleinsatz bezogen werden“ (ebd.).

So geht es bei der Frage der Effizienz nicht darum, ob Maßnahme X oder Maßnahme Y den besseren Effekt erzielt, „sondern in welchem Verhältnis dieser Effekt mit dem Aufwand zu ihrer Konzeption, Entwicklung und Durchführung steht“ (ebd.).

Als Beispiel wird eine Studie der Autovermietung AVIS angeführt, die berichtet, dass sie durch den Einsatz von CBT über eine Million Dollar einsparen konnte – der Lernerfolg zeigt sich für AVIS demnach in anderen Faktoren: einer verkürzten Bearbeitungsdauer für Reservierungen, die Trainingsdauer der Maßnahme usw.

In dieser Arbeit werde ich auf diese Sichtweise von Effizienz nicht näher eingehen, da sie einen anderen Aspekt des Konstrukts beleuchtet. Es sollen hier weniger die wirtschaftlichen als die pädagogischen Sichtweisen von „Lernerfolg“ in den Fokus genommen werden.

4.5 Lernziele als Faktor

Eine klare Definition von Lernerfolg ist resümierend ebenso wenig problemlos möglich, wie dessen empirische Erfassung. Jedoch scheint eine Lösung dieser Problematik gerade darin zu liegen, das Konstrukt näher aufzuschlüsseln. So spricht nicht nur Michael Kerres (2001:112) von der unterschiedliche Anwendungsnähe (Erinnern, Anwenden, Transfer) des „objektiven“ Lernerfolgs, sondern auch Sigmar-Olaf Tergan (2004:18) weist darauf hin, dass der Lernerfolg den Anforderungen des Lernangebotes entsprechen muss (vgl. ebd.).

Das bedeutet m.E., dass bei der Bewertung der Lernleistung auch immer eine Vorstellung über die Lernziele erforderlich ist: Für die Abfrage von bloßem abstraktem Faktenwissen ist beispielsweise das Reproduzieren des Stoffes zur Feststellung des Lernerfolgs ausreichend. Bei der Bewertung von komplexeren Problemen sind dagegen auch Anwendungswissen bzw. Methodenwissen gefordert. Die unterschiedlichen Dimensionen des Konstrukts Lernerfolg dürfen – auch in Bezug auf die Lernziele – also nicht durcheinander gebracht werden. Dies bedeutet aber auch, dass eine Bewertung des Lernerfolgs nur, aber eben auch genau dann möglich zu sein scheint, wenn eine Berücksichtigung der Dimensionen erfolgt. Bezieht sich ein Lernziel beispielsweise auf das Erinnern von Faktenwissen, so kann auch der Lernerfolg unter Berücksichtigung dieses Lernziels erfasst werden. Wesentlich für eine realistische Beurteilung des Lernerfolgs ist daher die differenzierte Festlegung von Lernzielen, auf deren Basis die Operationalisierung der Variablen und die Entscheidung über das Forschungsdesign erfolgen müssen. Dafür bietet sich die Taxonomie nach Lorin W. Anderson und David R. Krathwohl (2001) an, auf die in Kapitel 6 eingegangen werden soll.

Zuvor möchte ich skizzieren, wie sich die eben erfolgte Argumentation auf den Vergleich von Lernsettings anwenden lässt.

5 Lernerfolg und Leistungsmessung in Evaluationen und Vergleichsstudien

Um jedoch den Erfolg von E-Learning im Vergleich zu traditionellen Lernformen zu erfassen, werden oftmals Vergleichsstudien herangezogen. Diese eröffnen – unter Berücksichtigung der bisherigen Argumentation – jedoch nur bedingt die Möglichkeit, Aussagen über den Erfolg des einen oder anderen Lernarrangements zu treffen (vgl. Annabell Preussler & Peter Baumgartner 2006:81). Betrachtet man die aufgezeigten Schwierigkeiten, die bei der Messung von Lernerfolg auftreten, wird schnell klar, dass auch Evaluationen, die eine Maßnahme in Bezug auf ihren Lernerfolg untersuchen, nicht problemlos durchzuführen sind. So weisen beispielsweise Vergleichsstudien das Problem auf, „herkömmliche traditionelle Maße zu erfassen und so das neue Medium nicht in seinen emergenten Funktionen zu erfassen“ (Patrick Blumschein 2004:7).

Werden verschiedene Lernsettings miteinander verglichen, werden oft nur Funktionalitäten untersucht. Die EvaluatorenInnen gehen der Frage nach, welches Setting besser geeignet ist, also den höheren Erfolg bringt. Nicht ausreichend berücksichtigt wird hingegen oftmals die Frage „besser wofür?“

„Leider haben wir es auch bei der Evaluation von multimedialen Lehr-Lernumgebungen häufig mit der Tatsache zu tun, daß Forscher etwas evaluieren sollen, bei dem die Autoren, die diese Multimediaapplikation erstellt haben, gar nicht wußten, von welchen Lernzielen sie ausgingen. [...] Die Einstellung zum Thema Evaluation wird oft folgendermaßen umschrieben: ‚Wir machen unsere Vorlesungen wie bisher, haben einige Multimedia-Applets dazu entwickelt, und nun evaluiert mal schön‘. Eine gemeinsame Diskussion über Lehrziele kommt sehr, sehr schwer in Gang“ (Reiner Fricke 1999:196).

Dass wir uns „zu Tode evaluieren“, wie der Titel dieser Arbeit andeutet, wird durch diese Aussage von Reiner Fricke besonders anschaulich zur Geltung gebracht. Welche Aspekte müssen bei Evaluationen also beachtet werden und welche Maße gilt es zu erfassen?

„Die Evaluation der ‚Neuen Medien‘ als Lehr- und Lernmedien im tertiären Bildungssektor hat die wichtige Aufgabe, die medialen Materialien, Produkte und Projekte im Hinblick auf ihre Eignung, Qualität und Angemessenheit zu untersuchen und so zu ihrer Weiterentwicklung und Optimierung beizutragen. Neben die konkrete Evaluation einzelner Materialien und Produkte, Software-Anwendungen und interaktiver Lehr- und Lernmedien tritt die nicht minder wichtige – aber häufig nicht wahrgenommene, d.h. die nicht gesehene ebenso wie nicht ausgeführte – Aufgabe der Medienevaluation als Basis für die kritische Auseinandersetzung mit den Zwecken, den Mitteln und den Zielgruppen der (multi-)medialen Lehr- und Lernumgebung“ (Christine von Prümmer 1999:221).

Einen methodischen Forschungsstandard stellten viele Jahre lang Vergleichsuntersuchungen im herkömmlichen Sinne dar – ein Lernsetting wurde mit einem anderen verglichen. Auf diese Weise wurde erhofft,

Rückschlüsse auf den Erfolg eines Settings zulassen zu können. Um beispielsweise den bereits erwähnten programmierten Unterricht zu erforschen, wurde genau solch ein Verfahren angewendet (vgl. Patrick Blumschein 2004:8, Peter Strittmatter & Helmut Niegemann 2000, Almut Köbberling 1971). Nach wie vor erfreuen sich derartige Vergleichsstudien großer Beliebtheit. In vielen Bereichen mag dies auch ein sinnvolles Verfahren in der pädagogischen Forschung sein, im Bereich der Medienwirkungsforschung ist es m.E. durchaus diskussionswürdig.

Ein Trugschluss bei der Evaluation mediengestützten Lernens ist die Auffassung, man könne verschiedene Lernsettings ohne Schwierigkeiten miteinander vergleichen. Ein solches Verfahren führt in der Regel zu nicht-signifikanten Ergebnissen, wie an Hand zahlreicher Studien in diesem Bereich festgestellt werden kann (vgl. Peter Baumgartner 1999a:91f., Sigmar-Olaf Tergan 2003, Jane M. Carey 2001:7).

Peter Baumgartner (1997, 1999a) kritisiert, dass der Lernerfolg in den meisten Fällen durch das Testverfahren definiert wird, das aus – nachvollziehbaren – Gründen der Vergleichbarkeit das gleiche Prüfverfahren für alle Testgruppen festlegt. Dies sei jedoch in zweierlei Hinsicht problematisch: Handelt es sich um einen Vergleich rezeptiven Lernens, in dem die Vermittlung von Wissen hauptsächlich durch Text oder Grafik erfolge, würden gerade die spezifischen Eigenschaften (wie z.B. Interaktivität) der neuen Medien nicht genutzt.

„Trotz aufwändiger und kostenintensiver Methodik unterliegen sie einem Zirkelschluss: Zuerst wird Lernen kognitivistisch eingeschränkt und anschließend behauptet, dass ‚Lernen‘ durch interaktive Software generell nicht besser unterstützt wird, als es traditionelle Mittel (z.B. ein gutes Buch) ohnehin schon tun bzw. getan haben“ (Peter Baumgartner 1997:135).

Diesen gedanklichen Trugschluss macht Baumgartner an einem Beispiel deutlich: Möchte man verreisen, würde man – in Analogie der eben geschilderten Annahme – danach fragen, ob ein PKW oder ein LKW für diese Urlaubsreise besser geeignet sei. Damit schränkt man den „Fahrerfolg“ auf diese spezielle Reiseart ein. Man hebt also gerade den Vorteil hervor, den eine der zu untersuchenden Maßnahmen aufweist (vgl. ebd.).

Der zweite Fall wäre der, dass in Evaluationen nachgewiesen werden soll, dass interaktive Medien für bestimmte Zwecke besser geeignet seien.

„In diesen Untersuchungen wird zwar ‚Lernerfolg‘ nicht rezeptiv definiert, doch kann dies meist in einem vergleichenden Evaluationsdesign nicht adäquat eingelöst werden, weil – qua Evaluationsdesign – der Lernerfolg mit gleichen Inhalten und gleicher Prüfmethode festgestellt werden soll. Welchen Sinn machen vergleichende Evaluationen der interaktiven Performance von Fertigkeiten, wenn sie nur durch eines der zu untersuchenden Medien (zB. Softwaresimulation) vermittelt bzw. geübt wurden und im anderen (zB. Buch) nur in ihrem Ablauf *beschrieben* wurden?“ (ebd., Hervorhebungen im Original).

Um das eben genannte Beispiel aufzugreifen: Peter Baumgartner führt an dieser Stelle die Frage an, ob für den Transport von großen Lasten ein PKW oder ein LKW besser geeignet sei (ebd.). Auch hierbei wird deutlich, dass die Frage bereits die bestgeeignete Lösung impliziert.

Ein anderes Beispiel stellt Abbildung 7 dar, welche m.E. sehr gut verdeutlicht, welche Schwierigkeiten das Prüfverfahren mit sich bringt. Die Karikatur von Hans Traxler (1999:36)³³, die mir ursprünglich aus dem sonderpädagogischen Bereich³⁴ bekannt wurde, bringt die Problematik auf den Punkt. Es wird versucht, gerecht zu prüfen, welcher Prüfling die Aufgabe, auf den Baum zu klettern, am besten erfüllt. Mit diesem Testverfahren werden in diesem Fall spezielle Eigenschaften „gemessen“, die nur ein Teil der Prüflinge aufweist. Gerade die Eigenschaften, die andere Prüflinge auszeichnen und von den anderen abheben, werden mit dieser Prüfung jedoch nicht erfasst. Dabei wird vernachlässigt, dass diese Eigenschaften in anderen Situationen durchaus förderlich sein können.

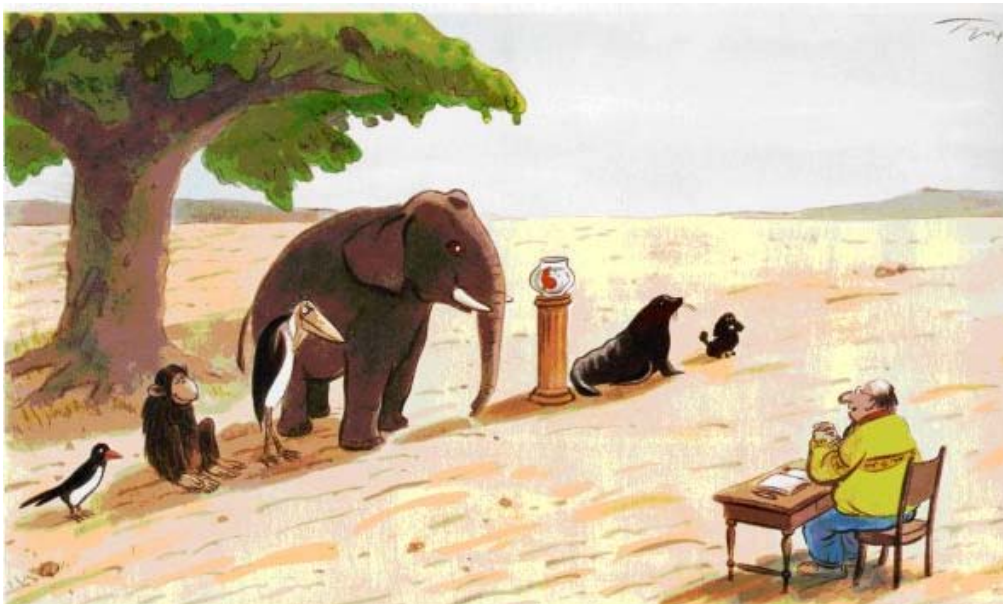


Abbildung 7: Karikatur „Zum Ziele einer gerechten Auslese lautet die Prüfungsaufgabe für Sie alle gleich: Klettern Sie auf den Baum!“ (Cartoon von Hans Traxler³⁵ 1999:36/37)

So ist es zwar legitim zu behaupten, dass der Affe diese Aufgabe am besten lösen wird, man betrachtet dabei jedoch nur einen kleinen Aspekt – nämlich das Klettern. Aber sind damit alle Dimensionen erfasst?

³³ Ursprünglich 1975 veröffentlicht, 1999 neu aufgelegt.

³⁴ Bei der ersten Veröffentlichung der Karikatur im Jahr 1975 ging es dem Zeichner in erster Linie darum, das Normenabitur zu kritisieren. Die Zeichnung erfreute sich großer Beliebtheit und wurde im Jahr 2001 noch einmal in neuer Form in der Zeitschrift „Erziehung und Wissenschaft“ (2/2001:1-3) veröffentlicht. Statt der Tiere waren nun aber verschiedene Menschen zu sehen – im Rollstuhl, mit einer Leiter, mit großem Körperumfang usw. Ziel war es, zu zeigen, wie sehr das Herkunftsmilieu die Berufs- und Lebenschancen der Kinder bestimmt. (vgl. Schroedel-Verlag 2001:4). Im Rahmen sonderpädagogischer Seminare wird die Zeichnung nach wie vor eingesetzt (vgl. beispielsweise Universität Dortmund 2004).

³⁵ Verwendung mit freundlicher Genehmigung des Zeichners.

Sind nicht eine Reihe anderer Settings denkbar, in denen andere Dimensionen an Bedeutung gewinnen?

Überträgt man diese Überlegungen zurück auf den Bereich von Evaluationen mediengestützter Lernsettings, wird deutlich, dass ein Testverfahren, das alle Testgruppen mit der gleichen Methode prüft, nicht alle Funktionen und Eigenschaften der einzelnen EvaluandInnen berücksichtigen kann. Das bedeutet jedoch: Wenn ein Untersuchungsdesign in der Komplexität reduziert werden soll und deshalb die unabhängigen Variablen konstant gehalten werden, dann darf sich auch die Interpretation der Ergebnisse nicht auf diese Variablen beziehen (vgl. Peter Baumgartner 1997:135). Er weist darauf hin, dass es bei der Evaluation interaktiver Software eben nicht um die Analyse des Materials geht, sondern darum, inwiefern Lernprozesse durch diese Software angeregt bzw. initiiert werden können (1999a:95). Die Frage nach diesen Lernzielen wird oftmals vernachlässigt, weil die Aufmerksamkeit vor allem auf die Methodik ausgerichtet ist. Dies führt dazu, dass sich das Ergebnis häufig auf die Bewertung einfacher Formen der Wissensvermittlung (präsentative Lehrformen) und weniger auf die Evaluierung von Fähigkeiten und Kompetenzen bezieht (vgl. Peter Baumgartner & Annabell Preussler 2005:5).

Zusammenfassend möchte ich mit Rolf Schulmeister et al. (2005:42) feststellen, dass

„aus dem Darbietungsmodus der Lerninhalte und der Unterscheidung zwischen verschiedenen Lernmedien [...] keine Unterschiede im objektiv messbaren individuellen Lernerfolg oder der von den Lernern subjektiv empfundenen Lernqualität ableitbar [sind]“ (a.a.O.).

Oder, wie Joachim Hasebrook bereits vor 10 Jahren feststellte:

„Nach dem heutigen Erkenntnisstand ist es nicht möglich, prinzipielle Aussagen über die Lernwirkungen von Multimedia zu machen“ (Joachim Hasebrook 1995:101).

Wie ich jedoch bereits an anderer Stelle gezeigt habe (vgl. Annabell Preussler & Peter Baumgartner 2006:81), halte ich Evaluationen für die Qualitätssicherung nach wie vor für notwendig. Zwar möchte ich auf der einen Seite Gerhard M. Zimmer & Elena Psaralidis (2000:263f) zustimmen, „dass die Bewertung eines Lernangebotes auf der Grundlage seiner Produktmerkmale noch nichts über die erzielbaren Wirkungen aussagt“, auf der anderen Seite kommt man in der Praxis um realistische Bewertungen nicht umhin. Wie soll also die Qualität eines Lernsettings bewertet werden? Gerhard M. Zimmer & Elena Psaralidis (2000:265, vgl. auch Ulf Ehlers 2002:7ff) stellen fest, dass Qualität erst durch das Lernen selbst hergestellt werden kann. Diese sollte nach Sigmar-Olaf Tergan (2003a) nicht an den Lernaktivitäten festgemacht werden,

„sondern nur im Hinblick auf die Bewältigung bestimmter Anforderungen von Problem- und Aufgabenstellungen“, die in unterschiedlicher Weise von den Lernenden bzw. vom Lernangebot ausgehen können“ (a.a.O.:425).

Daher lautet die Schlussfolgerung an dieser Stelle nicht: „Evaluationen sind nach derzeitigem Erkenntnisstand ohnehin nicht möglich“, sondern: „Das (vergleichende) Untersuchungsdesign muss der Komplexität der untersuchten Wechselwirkungen entsprechen!“ (Annabell Preussler & Peter Baumgartner 2006:81) Mit dieser Vorgehensweise macht man es sich jedoch nicht gerade leicht. So soll einerseits das Gesamtgefüge des Konstrukts Lernerfolg beachtet werden, andererseits sollen mindestens zwei Evaluanden ganzheitlich miteinander verglichen werden, um für die Praxis verlässliche Aussagen treffen zu können.

Eine Lösung dieses Problems stellt die Betrachtung der Lernziele dar: Aus der bisherigen Diskussion wurde deutlich, dass der Lernerfolg adäquat zum Lernangebot sein muss – er muss dessen Anforderungen entsprechen. Es ist also erforderlich, im Vorfeld der Bewertung der Lernleistung eine Vorstellung über die Lernziele zu haben. Was soll in diesem Lernsetting eigentlich gelernt werden? Geht es um das Auswendiglernen von Faktenwissen oder um die Analyse komplexer Texte? Lernziele helfen an dieser Stelle zu überprüfen, ob das, was in einem Lernsetting intendiert war, auch dem entspricht, was letztlich tatsächlich abgefragt wurde. Hierbei handelt es sich um zwei Dimensionen des Konstrukts Lernerfolg, die nicht miteinander verwechselt und nicht durcheinander gebracht werden dürfen. Wie ich bereits dargestellt habe, scheint die Taxonomie der Lernziele nach Lorin W. Anderson und David R. Krathwohl (2001) dafür besonders geeignet zu sein, da man sowohl die Lernziele als auch die Verfahren der Lernzielüberprüfung kognitiven Prozessen zuordnen kann. Auf diese Taxonomie soll im Folgenden eingegangen werden.

6 Taxonomie der Lernziele nach Lorin W. Anderson und David R. Krathwohl (2001)

Die Grundfrage, die im Laufe dieser Arbeit immer wieder auftaucht, ist die nach dem Lernerfolg in dem einen oder anderen Lernsetting. Ist der Lernerfolg höher, wenn eine virtuelle Lehrveranstaltung besucht wird, oder eine traditionelle? Konsens scheint darin zu bestehen, dass als Lernerfolg das (beabsichtigte) Ergebnis eines Lernprozesses bezeichnet werden kann³⁶. Damit aber ein Ergebnis überhaupt beabsichtigt sein kann, müssen im Vorfeld Lernziele festgelegt worden sein. Wie ich bereits in den letzten Kapiteln dargestellt habe, können sie auch zur Überprüfung des Lernerfolgs in Bezug auf seine verschiedenen Dimensionen nützlich sein. In der Regel dienen Lernziele dazu, den Lernprozess zu steuern und auszurichten (vgl. Marc Jelitto 2006:1).

„Sie [die Lernziele; Anm. AP] können von Lehrenden zur Planung des Unterrichts, Auswahl von Materialien und dem Entwurf von Prüfungsaufgaben verwendet werden. Von Lernenden können sie zur Orientierung und Selbstkontrolle genutzt werden. Weiterhin können auch Bedingungen für die Ausführung der Tätigkeit durch die Lernenden und Kriterien für die Messung von der Zielerreichung angegeben werden“ (ebd.).

Möchte man Lernenden etwas beibringen, wird in der Regel die Frage „Was soll vermittelt werden?“ gestellt. Kurt Schobel & Ulrike von Holdt (2004:17) weisen darauf hin, dass bei der Formulierung von Lernzielen eigentlich eher die Folge hinterfragt werden müsste: „Was soll dabei herauskommen?“³⁷ Anstelle einer Input-Orientierung sollte sich ein zielgerichtetes Studienmodul nach der Output-Orientierung richten und dies zur Zielfrage für die Lehrenden machen.

„Eine einheitliche, globale Zielsetzung für einen Studiengang wäre z. B.: „Die Studierenden sind befähigt, den Ingenieurberuf auszuüben“. Die Gestaltung der einzelnen Studienmodule und Veranstaltungen muss dementsprechend gewährleisten, dass dieses Ziel erreicht werden kann“ (a.a.O.:17f.).

Grundsätzlich werden in der Debatte um Lernziele drei Arten unterschieden³⁸: Während sich kognitive Lernziele auf Wissen und intellektuelle Fertigkeiten der Lernenden beziehen, fokussieren die affektiven Lernziele auf deren Einstellungen und Werthaltungen. Als letzter Punkt seien noch die psychomotorischen Lernziele genannt, die sich auf manuelle Fertigkeiten beziehen und eher formalen Charakter haben. Obwohl natürlich alle drei Lernzielarten beim Lernprozess eine Rolle spielen, liegt der Fokus dieser Arbeit auf den kognitiven Lernzielen, deren Klassifizierung Lorin W. Anderson und David R. Krathwohl (2001) als Hilfestellung zum Verständnis und zur Implementierung standardisierter Curricula für Lehrpersonen in einer Taxonomie vorgenommen haben.

³⁶ Näheres in Kapitel 4.3 „Lernerfolg – Definitionsversuche“

³⁷ Siehe auch die Ausführungen zu Input-, Output- und Prozessqualität im Kapitel 4 „Lernqualität“.

³⁸ Vgl. <http://de.wikipedia.org/wiki/Lernziel>, letzter Aufruf 12.03.2007

„Unter einer Taxonomie³⁹ wird ein systematisches Klassifikationsschemata zur Ordnung von Dingen, Erscheinungen, Prozessen etc. nach einheitlichen sachlogischen Prinzipien, Verfahren und Regeln verstanden“ (Peter Baumgartner 2006:1).

Diese von Lorin W. Anderson und David R. Krathwohl (2001) erstellte Klassifizierung baut auf der von Benjamin S. Bloom (1956) entwickelten Taxonomie der Lernziele auf, die aus den Kategorien „Knowledge“, „Comprehension“, „Application“, „Analysis“, „Synthesis“ und „Evaluation“ bestand. Diese erfüllte bereits zu dieser Zeit mehr Funktionen als ein reines Messinstrument.

„He believed it could serve as a

- common language about learning goals to facilitate communication across persons, subject matter, and grade levels;
- basis for determining for a particular course or curriculum the specific meaning of broad educational goals, such as those found in the currently prevalent national, state, and local standards;
- means for determining the congruence of educational objectives, activities, and assessments in a unit, course, or curriculum; and
- panorama of the range of educational possibilities against which the limited breadth and depth of any particular educational course or curriculum could be contrasted“ (David R. Krathwohl 2002:212).

2001 entwickelten Lorin W. Anderson und David R. Krathwohl eine revidierte Version dieser Taxonomie. Sie bezeichnen mit dem Begriff „learning outcomes“ Aussagen darüber, über welche Handlungsfähigkeit Studierende nach Abschluss eines Lernvorgangs verfügen sollen. Als Beispiel sei der Besuch einer Vorlesung genannt: Was wurde damit initiiert und was können die Studierenden anschließend besser? Der Fokus bei Lorin W. Anderson und David R. Krathwohl liegt also auf den intendierten Lernzielen, deren Klassifizierung und Einordnung durch die entwickelte Taxonomie erleichtert werden soll, und nicht den Inhalten (vgl. auch Kurt Schobel & Ulrike von Holdt 2004:23).

6.1 Struktur der Taxonomie

Die Taxonomie der Lernziele verbindet Wissensarten mit kognitiven Prozessen (vgl. Susanne Heyer 2006:12). Sie beschreibt auf der senkrechten Achse zunächst die Wissensarten und auf der waagerechten Achse die kognitiven Prozess-Dimensionen.

³⁹ von griechisch. taxis „Ordnung“, -nomia „Verwaltung“ (vgl. <http://de.wikipedia.org/wiki/Taxonomie> letzter Aufruf 12.03.2007)

DIE WISSENS-DIMENSIONEN	DIE KOGNITIVE PROZESS-DIMENSION					
	1. Erinnern	2. Verstehen	3. Anwenden	4. Analysieren	5. Bewerten	6. (Er)schaffen
A. Faktenwissen						
B. Begriffliches Wissen						
C. Verfahrensorientiertes Wissen						
D. Metakognitives Wissen						

Tabelle 1: Taxonomiematrix nach Lorin W. Anderson & David R. Krathwohl (übersetzt nach Kurt Schobel & Ulrike von Holdt, 2004)

Dabei werden die Lernziele zunächst sprachlich erfasst:

„Objectives that describe intended learning outcomes as the result of instruction are usually framed in terms of (a) some subject matter content and (b) a description of what is to be done with or to that content. Thus, statements of objectives typically consist of a noun or noun phrase—the subject matter content—and a verb or verb phrase—the cognitive process(es). Consider, for example, the following objective: The student shall be able to remember the law of supply and demand in economics. ‚The student shall be able to‘ (or ‚The learner will,‘ or some other similar phrase) is common to all objectives since an objective defines what students are expected to learn“ (David R. Krathwohl 2002:213).

Es werden vier Wissensdimensionen unterschieden: Faktenwissen, begriffliches Wissen, verfahrensorientiertes Wissen und metakognitives Wissen.

„Faktenwissen“ beschreibt ein Basiswissen, das erforderlich ist, um mit Probleme in einer bestimmten Disziplin lösen zu können. Es beinhaltet einerseits das Wissen über Terminologien und andererseits das Wissen über spezifische Details und Elemente. „Begriffliches Wissen“ zielt ab auf die Interrelationen zwischen den Basiselementen innerhalb einer größeren Struktur, um ein gemeinsames Funktionieren zu ermöglichen. Dazu gehören das Wissen über Klassifikationen und Kategorien, das Wissen von Prinzipen und Generalisierungen sowie das Wissen über Theorien, Modelle und Strukturen. „Verfahrensorientiertes Wissen“ beschreibt das Wissen darüber, wie etwas getan wird. Hierbei geht es um die Methoden des Nachforschens und Kriterien zur Anwendung von Fähigkeiten, Algorithmen, Techniken und Methoden. Diese Wissensart kann wiederum unterteilt werden: Wissen über fachspezifische Fähigkeiten und Algorithmen, Wissen über fachspezifische Techniken und Methoden sowie das Wissen über die Kriterien zur Anwendung bestimmter Prozeduren. „Metakognitives Wissen“ beschreibt letztendlich ein generelles kognitives

Wissen sowie das Bewusstsein und die Kenntnis über die eigene Kognition. Es beinhaltet strategisches Wissen, Wissen über kognitive Aufgaben unter Einbeziehung des zuvor erworbenen Wissens sowie das Wissen über sich selbst (vgl. a.a.O.:214).

Darüber hinaus unterscheiden Lorin W. Anderson und David R. Krathwohl sechs kognitive Lernzielkategorien: „Erinnern“, „verstehen“, „anwenden“, „analysieren“, „bewerten“ und „(er)schaffen“. Diese Kategorien beschreiben die kognitiven Abläufe des Lernprozesses (vgl. ebd.; Kurt Schobel & Ulrike von Holdt 2004:42).

„Die dargestellten Lernzielkategorien dienen dazu, die jeweils geforderte kognitive Leistung (auf Erkenntnis- und Wahrnehmungsfähigkeit bezogene Leistung) der Studierenden auf verschiedenen Stufen einzuordnen und die Anforderungen an die Studierenden auf diese Weise transparent zu machen“ (Kurt Schobel & Ulrike von Holdt, 2004:20).

So beschreibt „erinnern“ das Abrufen von relevantem Wissen aus dem Langzeitgedächtnis. Mit „verstehen“ ist gemeint, die Bedeutung oder Intention von instruktionalen Botschaften zu erkennen (unter Einbezug on mündlicher, schriftlicher und bildhafter Kommunikation). „anwenden“ meint das Ausführen von bestimmten Verfahren in vorgegebenen Situationen. Mit „analysieren“ wird die vierte Prozessebene erreicht. Hier soll das Material in seine Bestandteile gegliedert werden und bestimmt werden, inwiefern es eine Verbindung zu übergeordneten Strukturen aufweist. „Bewerten“ zielt auf Urteile ab, die auf Grund von Standards und Kriterien gefällt werden sollen. Auf der höchsten Stufe beschreibt „(er)schaffen“ das Zusammenführen von Elementen zu einem neuen, kohärenten Produkt (vgl. ebd.).

In jeder Kategorie gibt es Subkategorien, die zur näheren Bestimmung des jeweiligen Prozesses dienen. Vor dem Hintergrund der Lernzielüberprüfung ist dies besonders relevant. So gehört zur Prüfung von „verstehen“ die Subkategorie „erklären“. Das bedeutet: Wenn ich in der Lage bin, einen Sachverhalt zu erklären, erfülle ich damit das Lernziel „verstehen“. In einer Prüfungssituation kann dies beispielsweise durch die selbstständige Konstruktion eines eigenen Beispiels abgefragt werden. „Verstehen“ kann aber nicht durch „erkennen“ geprüft werden. Die Subkategorien sind also bereits als ein Teil der Operationalisierung der Lernziele zu begreifen. Wichtig dabei ist, dass auch die Subkategorien nur Teile des Gesamtgefüges ausmachen. Selbstverständlich sind zusätzliche Operationalisierungen möglich.

Die Subkategorien seien an dieser Stelle kurz angeführt⁴⁰:

- In der Kategorie „Erinnern“ wird zwischen „erkennen“ und „ins Gedächtnis rufen“ unterschieden.
- „Verstehen“ kann in sieben Subtypen unterschieden werden: „Interpretieren“, „Veranschaulichen“, „Klassifizieren“, „Zusammenfassen“, „Folgern“, „Vergleichen“ und „Erklären“.

⁴⁰ Vgl. Abschnitt III im Anhang ab Seite LXXXVIII.

- Zu „Anwenden“ gehören die Subkategorien „Ausführen“ und „Implementieren“.
- „Analysieren“ lässt sich einteilen in „Differenzieren“, „Organisieren“ und „Kennzeichnen“.
- Die zwei wesentlichen Komponenten der kognitiven Ebene „Bewerten“ sind „Überprüfen“ und „Kritisieren“.
- Zu „(Er)Schaffen“ gehören Generieren, Planen und Produzieren.

Die Lernziele werden dementsprechend in der Taxonomie angeordnet: Die sogenannten Wissensdimensionen sind senkrecht, die kognitiven Prozess-Dimensionen waagrecht platziert und bilden auf diese Weise den Lernprozess ab. Die höheren Wissensarten, wie metakognitives Wissen, schließen dabei die niedrigeren Wissensarten (wie Faktenwissen) ein. Das bedeutet – nur wer bereits über Faktenwissen verfügt, kann sich auch begriffliches Wissen aneignen usw. Selbige hierarchische Anordnung gilt für die Prozess-Dimensionen: Nur wer in der Lage ist, sich zu erinnern kann auch verstehen; nur wer verstehen kann, kann auch anwenden usw. In diesem Fall steigt die Komplexität der Prozesse von links nach rechts.

6.2 Praktische Relevanz der Taxonomie zur Erfassung von Lernzielen und Lernerfolg

Die vorgestellte Taxonomie der Lernziele dient dazu, Lernziele aus der Sicht der Lernenden zu untersuchen – also: Welche Fähigkeiten sollen Lernende erwerben? Genügt es, Wissen abzufragen, oder müssen sie es auch anwenden können? Darüber hinaus dient die Taxonomie als Analyseinstrument, „bei dem der Komplexitätsgrad eines Lernziels in Verhältnis zu den Lernzielklassen⁴¹ gebracht wird“ (Franziska Keller 2004:13). Werden Aufgaben eines Assessments nach demselben Vorgehen eingeordnet, wie die Lernziele, kann die Taxonomie dazu beitragen, diese zu entwickeln, zu strukturieren und hilft, sie zu operationalisieren. Durch die Klassifizierung kann auf diese Weise ein Zusammenhang zwischen Lehrmethoden und dem Assessment hergestellt werden (vgl. ebd.).

Abbildung 8: Ablauf der Taxierung einer Prüfungsaufgabe (Franziska Keller 2004:17) stellt eine Möglichkeit des Ablaufs der Entwicklung einer Prüfungsaufgabe unter Bezugnahme eines Lernzieles dar. Werden Assessments auf diese Weise durchgeführt, können Bewertungen auf einer relativ objektiven Basis vergeben werden. Eine Schwierigkeit besteht darin, dass die Curricula von Schulen und

⁴¹ Unter Lernzielklassen sind Intensitätsstufen oder "Cognitive process dimension" zu verstehen. Der Komplexitätsgrad bezieht sich dabei auf den Abstraktionsgrad des Inhalts einer Aufgabe. Lernzielklassen sprechen hingegen das kognitive Verhalten an (vgl. Franziska Keller 2004:13).

Universitäten bereits bestimmte Ziele beinhalten, die schwer in konkrete Lernziele umgewandelt werden können, da sie auf einer höheren Abstraktionsebene liegen (vgl. a.a.O.:11). Eine weitere Kritik besagt,

„dass Wissen nicht in die genannten Gruppen zerbrochen werden kann, denn im Rückschluss stellen die Gruppen bei der Zusammenführung nicht mehr das ursprüngliche Wissen dar“ (Karin M. Wiburg 2003, zit. nach Susanne Heyer 2006:10).

Darüber hinaus weist diese Einordnung weitere Schwächen auf: Lorin W. Anderson und David R. Krathwohl (2001) haben sehr genaue Angaben darüber gemacht, welche Tätigkeiten unter die sechs kognitiven Prozesse fallen. Franziska Keller stellt hier beispielsweise jedoch „entscheiden“ sowohl unter „Analysieren“ als auch unter „Beurteilen“ dar. Eine trennscharfe Einordnung ist insbesondere in Bezug auf Prüfungsaufgaben jedoch notwendig.

Die Taxonomie scheint geeignet zu sein, um Lernziele und damit auch eine den intendierten Lernerfolge zu vergleichen und stellt eine Vorarbeit zur Operationalisierung der Lernziele dar. Allerdings beschreiben Lorin W. Anderson und David R. Krathwohl nicht, wie Lernziele in ihrem Erfolg erfasst werden können. Dies soll im nächsten Kapitel beleuchtet werden.

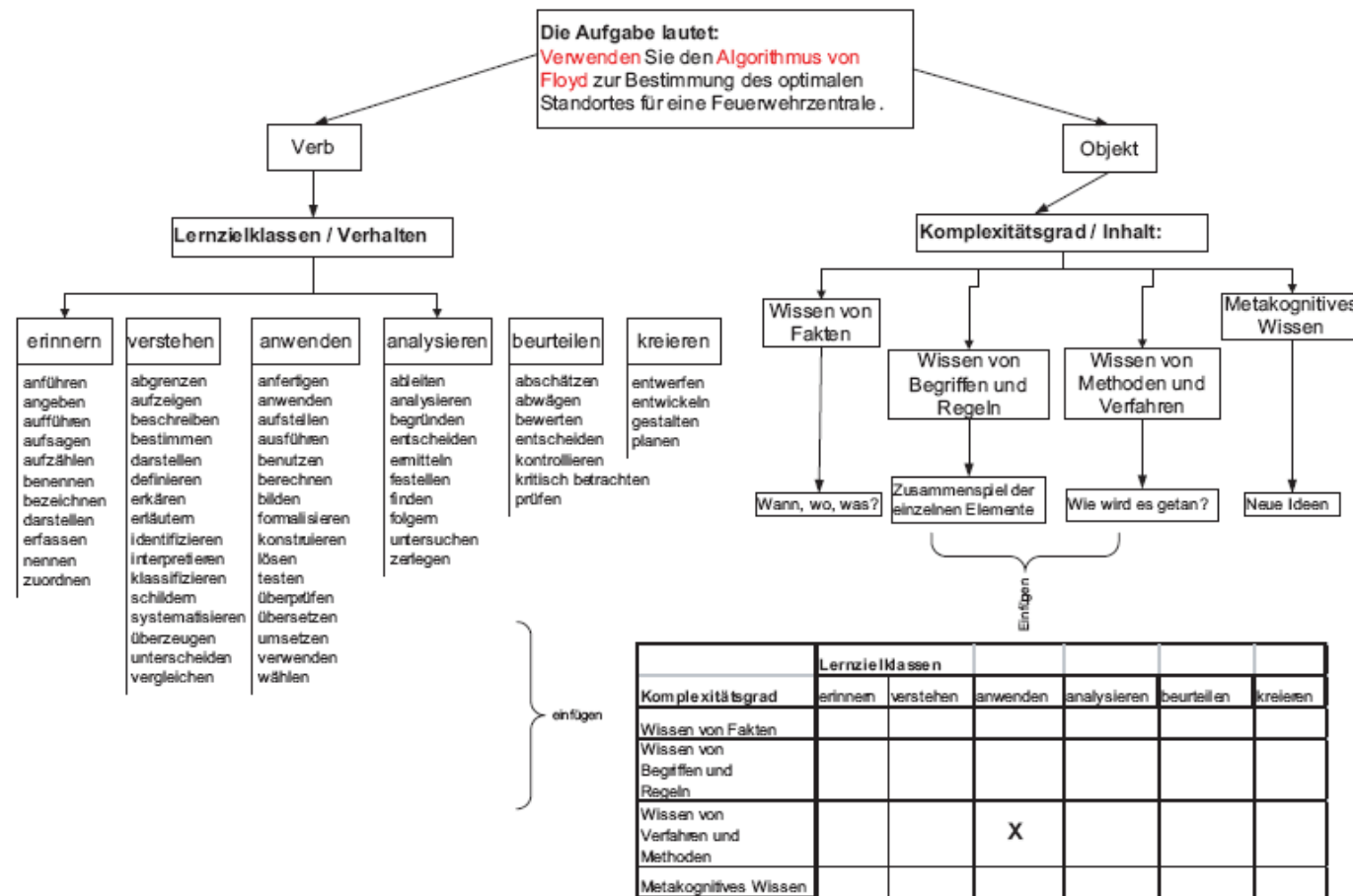


Abbildung 8: Ablauf der Taxierung einer Prüfungsaufgabe
(Franziska Keller 2004:17)

6.3 Aufgabentypen und Lernziele

Um das Verhältnis von Lernziel und Lernerfolgskontrolle zu beschreiben, soll zunächst die Ontologie von Aufgabentypen betrachtet werden, die Norbert Meder & Andrea Frick (2006:75) vornehmen. Sie verstehen Aufgaben als

„interaktive Wissensarten, die zwischen der reinen Vorgabe von Informationen und den noch komplexeren Szenarien kooperativen Lernens stehen. Aufgaben geben nämlich einerseits Informationen, lassen aber anderes aus, was der Lernende zu ergänzen hat. Aufgaben produzieren gezielt Leerstellen, die der Lernende auszufüllen hat“ (a.a.O.:71).

Dabei haben Aufgaben einerseits die Funktion, Wissen zu generieren, andererseits dienen sie aber auch der Überprüfung des Wissens. Ob das Lernziel erreicht wurde, kann durch eine dem Lernziel angepasste Aufgabe überprüft werden (vgl. .a.a.O.:72f.).

Aufgaben können – den kognitiven Operationen entsprechend - unterteilt werden in

1. Entdeckende Aufgaben,
2. Ordnungsaufgaben,
3. Antwortaufgaben,
4. Ankreuzaufgaben und
5. Distinktionsaufgaben (vgl. a.a.O.:75f.)

In „entdeckenden Aufgaben“ sollen vor allem „Regeln, Gesetzmäßigkeiten oder Zusammenhänge“ (ebd.) identifiziert werden. So gehören „Fehler entdecken“, „Differenz erkennen“ und „Problem lösen“ zu den entdeckenden Aufgaben (vgl. ebd.).

„Ordnungsaufgaben“ dienen dem Verständnis von Regeln und Strukturen im Umfeld des Lerngegenstandes. Zu den Ordnungsaufgaben gehören Sequenzierungsaufgaben (etwas muss in eine Reihung gebracht werden), Konstruktionsaufgaben (aus Teilen soll ein Ganzes zusammengesetzt werden, z.B. Satzkonstruktionen) und Zuordnungsaufgaben (einem Begriff soll ein anderer zugeordnet werden) (vgl. ebd.).

„Antwortaufgaben“ zielen darauf ab, Antworten zu geben. Diese können in freier Form (Kurzantwort Aufgabe), in Aufsatzform (Exposé Aufgabe), mündlich (Aussprache Aufgabe), in Form eines Lückentextes (Lückentext-Aufgabe) sowie als Buchstabier-Aufgabe gegeben werden (vgl. ebd.).

„Ankreuzaufgaben“ werden in Einfachauswahl-, Mehrfachauswahl und ja/nein-Aufgabe unterschieden und beinhalten klassische Multiple-Choice-Tests (vgl. ebd.).

„Distinktionsaufgaben“ beziehen sich vor allem auf das Lernen von Begriffen. Es werden unbestimmte Limitationen (der Gegenstand wird durch Negation eingegrenzt), bestimmte Negationen (der Gegenstand wird von ähnlichen Merkmalen abgegrenzt) sowie die Bandbreite des Begriffes in Begriffsumfang (Gegenstände, die unter den Begriff fallen, müssen benannt werden) und Begriffsinhalt (Merkmale des Begriffs müssen aufgelistet werden) unterschieden (vgl. ebd.).

Die folgende Abbildung stellt die Aufgabentypen in einer Mindmap dar:

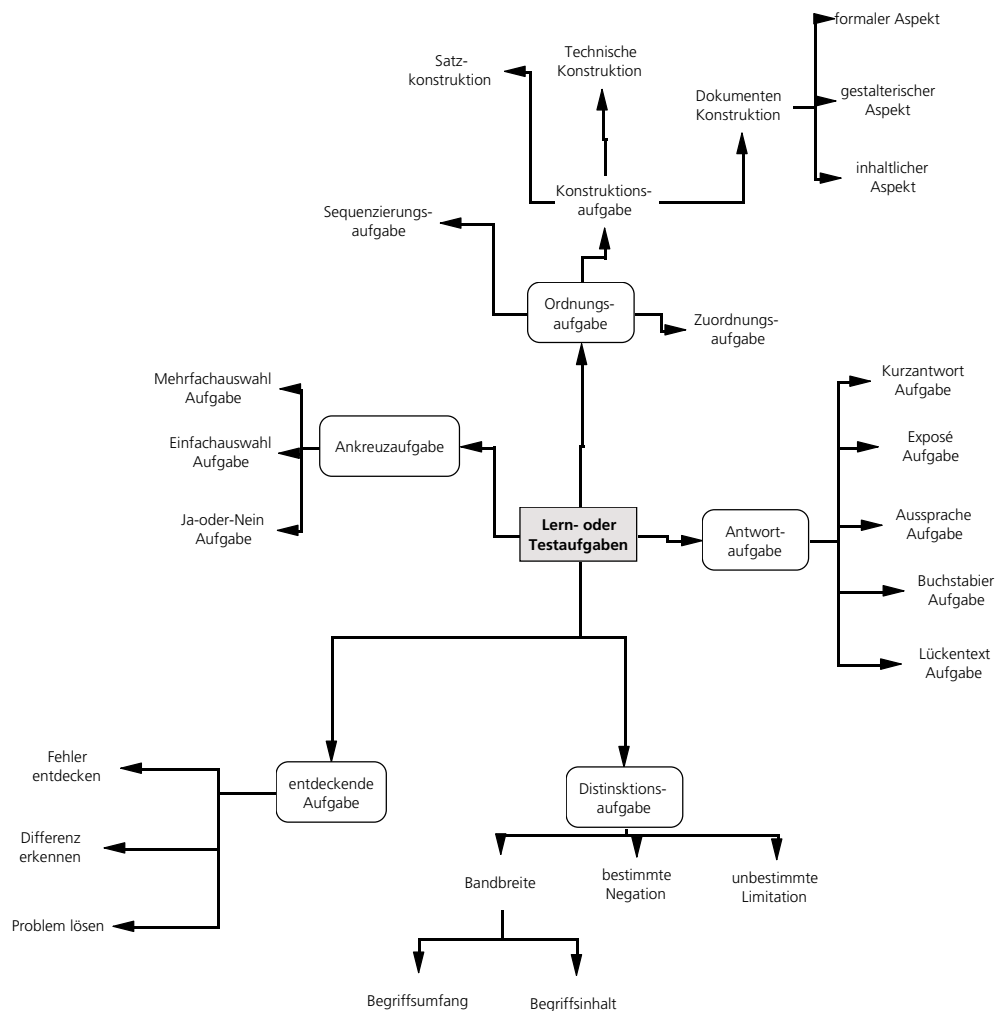


Abbildung 9: Mindmap Aufgaben (nach Norbert Meder & Andrea Frick 2006:75)

6.3.1 Zuordnung von Aufgabentypen zu Lernzielen

Es soll im Folgenden versucht werden, diese Ontologie auf die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001) zu übertragen.

Zur besseren Übersicht sind diese in der folgenden Abbildung strukturiert dargestellt. Die Hierarchie der Prozesse verläuft dabei von unten nach oben.

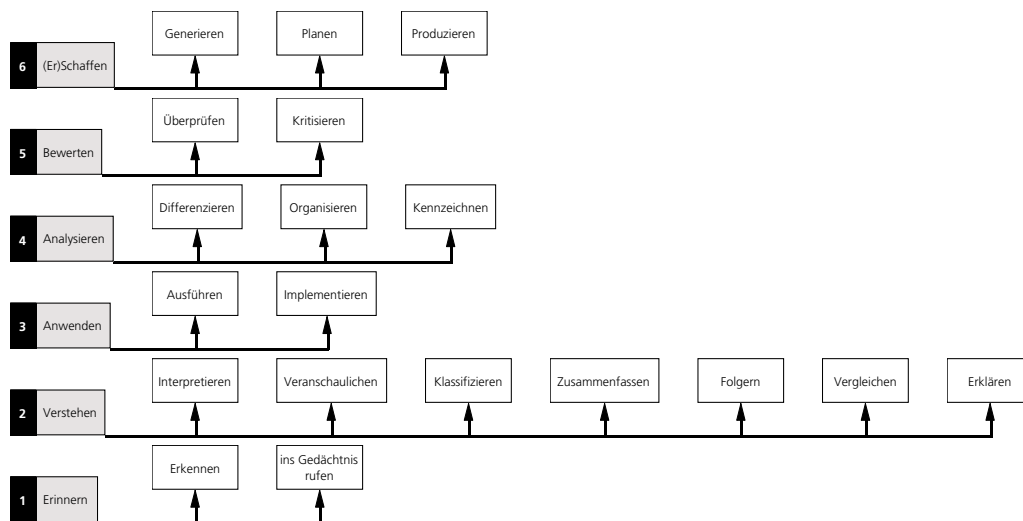


Abbildung 10: Kognitive Prozesse (nach David R. Krathwohl 2002:212)

Es wird deutlich, dass sich jedes Lernziel in weitere Subkategorien unterteilen lässt, die wiederum eine Operationalisierung dieses Lernziels darstellen. Wenn jemand in der Lage ist zu kritisieren, erfüllt diese Person damit das Lernziel „bewerten“. Dies lässt sich jedoch nicht durch eine reine Wissensabfrage überprüfen – damit würde ja nur die Behaltensleistung gemessen, die sich wiederum auf das Lernziel „erinnern“ bezieht. Das bedeutet, dass jedes Lernziel einer spezifischen Prüfungsform bedarf.

Aus diesem Grund werden den einzelnen Lernzielen im nächsten Schritt die Aufgabentypen zugeordnet, die zur Überprüfung der Lernziele erforderlich sind. Diese Zuordnung ist in Abbildung 11 dargestellt.

Hierzu dient die bereits beschriebene Ontologie nach Norbert Meder & Andrea Frick (2006:75f.). Die Beschreibung der Aufgabentypen und die Operationalisierung der Lernziele werden somit zusammengebracht. So geben Norbert Meder & Andrea Frick (ebd.) beispielsweise an, dass eine entdeckende Aufgabe durch das Erkennen von Differenz erfüllt wird. Dieser Überlegung entspricht die Operationalisierung des Lernziels „bewerten“, da es die Subkategorie „kritisieren“ enthält. Das bedeutet: Das Lernziel „bewerten“ kann durch eine entdeckende Aufgabe überprüft werden.

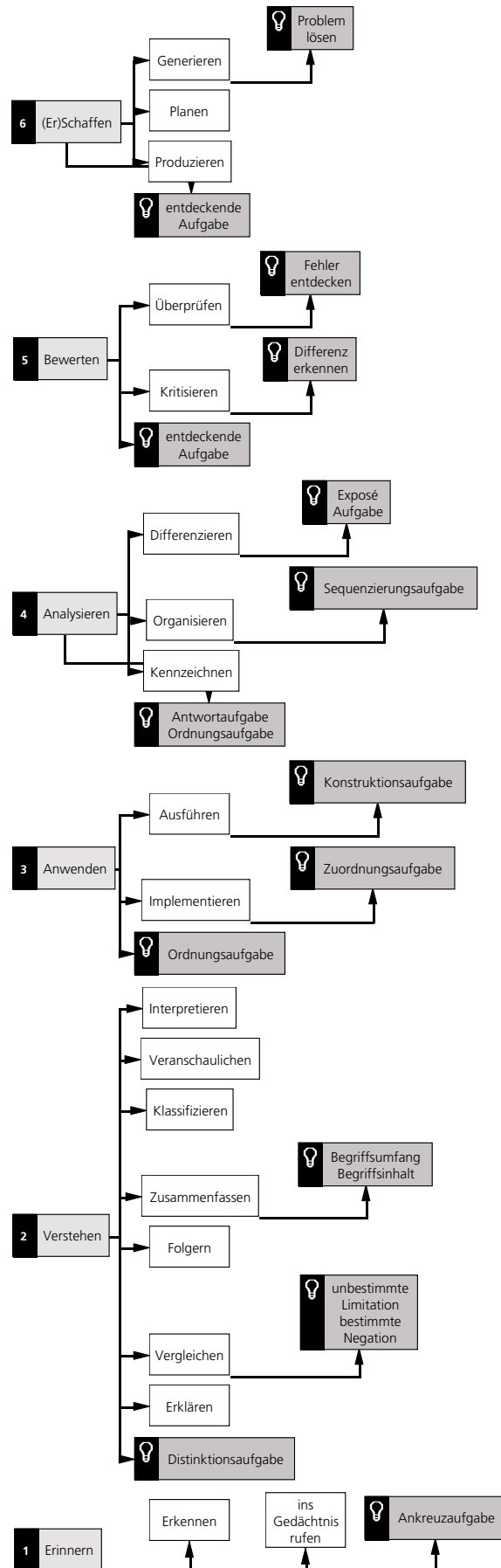


Abbildung 11: Kognitive Prozesse und Aufgabentypen

An dieser Stelle seien zur Verdeutlichung weitere Beispiele angeführt:

Ankreuzaufgaben werden von Norbert Meder & Andrea Frick (ebd.) beispielsweise als klassische Multiple-Choice-Tests beschrieben. Diese werden dementsprechend dem „erinnern“-Prozess zugeordnet, da an dieser Stelle „erkennen“ und „ins Gedächtnis rufen“ abgeprüft werden. Dies leistet ein Multiple-Choice-Test eher als ein schriftlicher Aufsatz, in den andere Prozesse hineinspielen würden. Die Ankreuzaufgabe misst also den Prozess „erinnern“ am genauesten.

Anders verhält es sich mit „Begriffsumfang“ und „Begriffsinhalt“. Diese wurden dem Prozess „verstehen“ angegliedert, da dieser „zusammenfassen“ enthält. Das bedeutet, dass Lernziele des Prozesses „verstehen“ durch Distinktionsaufgaben überprüft werden können.

Um ein letztes Beispiel zu geben: Der Aufgabentyp „Fehler entdecken“ wurde dem kognitiven Prozess „bewerten“ zugeordnet, da dieser „überprüfen“ enthält. Es ist also eine entdeckende Aufgabe erforderlich. Eine Überprüfung des Lernziels mit einem Lückentext wäre an dieser Stelle beispielsweise unsinnig.

Obwohl in Abbildung 11 dargestellte Zuordnung weder vollständig, noch eindeutig trennscharf ist, wird – auch unter Bezugnahme auf die vorherigen Kapitel – eines deutlich: für unterschiedliche Lernziele bedarf es unterschiedlicher Prüfungsformen. Nicht jeder Test ist für jeden Prozess geeignet. Es wird offensichtlich, dass es für „erinnern“ anderer Prüfungsmethoden bedarf als für „analysieren“ und für „(er)schaffen“ anderer als für „anwenden“. Hier ist noch weitere Forschungsarbeit nötig: Evaluationen sollten – um aussagekräftige Ergebnisse zu liefern – also berücksichtigen, dass Lernziele mit adäquaten Tests überprüft werden müssen, deren Dimensionen nicht vermischt werden dürfen, indem beispielsweise der gleiche Test für alle Gruppen eingesetzt wird – deren Lernziele sich jedoch qua Lernsetting bereits unterscheiden.

6.4 Zusammenfassung

Um einen Lernprozess zu prüfen (Lernerfolgskontrolle), muss vorab festgelegt worden sein, welche Ziele mit diesem Prozess genau erreicht werden sollen. Diese Überlegung mag überflüssig erscheinen, da sie zu naheliegend zu sein scheint – andernfalls ergäben schließlich auch didaktische Bemühungen keinen Sinn. M.E. wird es sich bei der Formulierung von Lernzielen in der Praxis jedoch oft zu einfach gemacht, die Lernziele werden zu universal formuliert und gehen wenig ins Detail.

Ein Hilfsmittel, um spezifische Lernziele entsprechend ihrer kognitiven Prozessdimensionen⁴² einzuordnen, stellen Lorin W. Anderson

⁴² Die Wissensdimensionen sind für die Arbeit nicht weiter relevant, da die Argumentation nur auf die kognitiven Prozesse ausgerichtet ist. Es dürfte darüber hinaus schwer sein, die entsprechenden Wissensdimensionen aus den Primärstudien abzuleiten.

und David R. Krathwohl (2001) mit der Taxonomie der Lernziele bereit. Dabei werden die Lernziele einerseits Wissensarten, andererseits kognitiven Prozessen zugeordnet. Auf diese Weise werden Lernprozesse geleitet und strukturiert.

Es bleibt jedoch die Frage, wie einzelne Lernziele genau überprüft werden können. Aus diesem Grund werden die Aufgabentypen nach Norbert Meder und Andrea Frick (2006:75f.) zu Rate gezogen und den entsprechenden kognitiven Prozessen zugeordnet: Je nach Lernziel müssen unterschiedliche Leistungen von den Lernenden erbracht werden. Darüber hinaus muss sich die Prüfung genau auf dieses Lernziel beziehen. Andernfalls würden möglicherweise Aspekte fokussiert, die nicht erfasst werden können oder die andere Dimensionen des Lernprozesses beleuchten, als die getesteten.

Beziehen wir diese Erkenntnis wiederum auf die Forschungsfrage, wie der Lernerfolg in unterschiedlichen didaktischen Szenarien bewertet werden kann, wird deutlich, dass es nötig sein wird, vorhandene Studien dahingehend zu untersuchen, ob das, was in einem Lernsetting intendiert war (also die Lernziele) auch tatsächlich dem entspricht, was letztlich überprüft wurde (also die Lernerfolgskontrolle).

Diese Überlegungen spielen eine wichtige Rolle bei der meta-evaluativen Untersuchung der Primärstudien. Bisher wurde auf das Problem aufmerksam gemacht, dass der Lernbegriff in Vergleichsstudien oftmals zu allgemein gehalten wird – es werden unterschiedliche Ausprägungen von Lernerfolg miteinander verglichen. Diese Problematik lässt sich gut mit Hilfe einer Analogie deutlich machen: Das beschriebene übliche Vorgehen entspricht unserem umgangssprachlichen Ausdruck vom Vergleich von Äpfeln mit Birnen. Ein Lernsetting (ein Apfel) wird mit einem anderen verglichen (einer Birne), jedoch berücksichtigen diese jeweils unterschiedliche Dimensionen. Der Lernbegriff wird dabei so allgemein gehalten, als würde man nicht von „Äpfeln“ und „Birnen“, sondern ganz einfach von „Obst“ sprechen.

Genau dieser Aspekt bleibt in den meisten Vergleichsstudien jedoch unberücksichtigt. Wenn beispielsweise die Untersuchungsgruppe Drill-and-Practice-Aufgaben in einem CBT erhält, die Kontrollgruppe jedoch das gleiche Material im Klassenraum diskutiert, unterscheiden sich die Lernziele deutlich voneinander und können sinnvollerweise nicht in einem einheitlichen Multiple-Choice-Test für beide Gruppen abgefragt werden.

Die Einordnung der Lernziele in die Dimension der kognitiven Prozesse ermöglicht jedoch eine spezifischere Operationalisierung: Wenn bedacht wird, dass es 6 kognitive Prozesse gibt, dann existieren somit bereits mindestens 6 Ausprägungen von Lernerfolg. Demnach kann man argumentieren, dass ein sinnvoller Vergleich von Untersuchungs- und Kontrollgruppen überhaupt nur dann möglich ist, wenn die untersuchten Faktoren dieselbe Ausprägung zu Grunde legen. Auf diese Weise können – um beim o.g. Beispiel zu bleiben – Äpfel mit Äpfeln und Birnen mit Birnen verglichen werden.

Durch eine Meta-Evaluation sollen im praktischen Teil dieser Arbeit bestehende Vergleichstudien im Hinblick auf diese Problematik analysiert werden. Wie bereits angekündigt, ist dabei von besonderem Interesse, ob die intendierten Lernziele der Lernerfolgskontrolle entsprechen und wie der Lernerfolg in den einzelnen Studien überhaupt operationalisiert wird. Darüber hinaus sollen Bereiche bzw. Dimensionen des Konstrukts Lernerfolg identifiziert werden, die sich tatsächlich miteinander vergleichen lassen.

7 Meta-Evaluation und Metaanalyse

Zur Reanalyse von wissenschaftlichen Untersuchungen, zur Integration von Forschungsergebnissen oder zur Beschreibung eines Forschungsfeldes sind Metaanalysen und Meta-Evaluationen übliche Verfahren. In der Regel werden dabei Originaldaten bestehender Studien miteinander verglichen, um zu allgemeingültigeren Ergebnissen zu gelangen. Dabei sind die Begriffe der Meta-Evaluation und Metaanalyse nicht besonders trennscharf.

„Metaevaluation kann als Oberbegriff verstanden werden, von dem eine Spielart die quantitative Metaanalyse ist, oder als die Beurteilung einer oder mehrerer Evaluationsstudien („Evaluation der Evaluation“)" (Bernhard Bühlren 1999:1).

Es fällt auf, dass in der Literatur beide Begriffe oftmals synonym verwendet werden (z.B. bei Petrei Uusikylä & Petri Virtanen 2000:50), dabei aber auch jeweils das eine oder andere meinen. Ich möchte der vorliegenden Arbeit die vorgestellte Definition zu Grunde legen und im Folgenden auf beide Begriffe eingehen.

7.1 Metaanalyse

Als Metaanalysen werden Verfahren bezeichnet, mit denen die Ergebnisse verschiedener Untersuchungen – so genannter Primärstudien – zur gleichen Thematik zusammengefasst werden. Dabei werden i.d.R. quantitative Methoden angewendet. Es handelt sich dabei um ein quantitativ-statistisches Verfahren.

„Mit einer Metaanalyse werden die zentrale Tendenz und die Variabilität der Ergebnisse über alle integrierten Einzeluntersuchungen hinweg ermittelt, um so zu einem aussagekräftigeren und sichereren Ergebnis zu kommen, als es durch eine einzelne Untersuchung oder ein Review möglich wäre“ (Patrick Blumschein 2004:XI).

Der Begriff wurde maßgeblich von Gene V. Glass (1976) geprägt, der darunter eine quantitative Integration der empirischen Forschungsergebnisse unter Berücksichtigung von Studienmerkmalen verstand.

„Primary analysis is the original analysis of data in a research study. [...] Secondary analysis is the re-analysis of data for the purpose of answering the original research question with better statistical techniques, or answering new questions with old data. [...] Meta-analysis refers to the analysis of analyses [...] for the purpose of integrating the findings“ (Gene V. Glass 1976:3).

Metaanalytische Verfahren sind wissenschaftliche Methoden und orientieren sich an messbaren Gütekriterien. Sie sind darauf ausgerichtet, sich einen Überblick über den aktuellen Stand der Forschung zu verschaffen und werden in der Regel als quantitative Methoden begriffen,

„die eine Forschungsfragestellung über die statistische Integration von Stichprobenkennwerten aus den Ausgangsstudien

beantworten“ (Erik Farin 1995, zit. nach Bernhard Bührlen-Armstrong 1998:112).

Es handelt sich also um eine verallgemeinernde Methode, die verschiedene homogene Primärstudien unter einem gemeinsamen Nenner subsumiert. Besondere Bedeutung kommt dabei der Effektstärke zu. Wenn beispielsweise in unterschiedlichen Primärstudien unterschiedliche Effekte einer Maßnahme gefunden werden, soll mithilfe der Metaanalyse der „wahre“ Effekt herausgefunden werden.

Auf welche Weise Metaanalysen durchgeführt werden, hängt davon ab, wie vollständig die Befunde aus Primärstudien zur Verfügung stehen (vgl. Reiner Fricke & Gerhard Treinies 1985:19). Sie können mit inferenzstatistischen und deskriptiven Methoden durchgeführt werden: Entweder wird die Größe des mittleren Korrelationskoeffizienten bzw. des mittleren Effekts bestimmt. Ebenso kann „die Wahrscheinlichkeit für das Zustandekommen der zu integrierenden Ergebnisse unter Annahme der Gültigkeit der Null- oder einer anderen Hypothese“ (a.a.O.:18f.) berechnet werden. Diese wird im Sinne von Hypothesentests zur Grundlage dafür gemacht, ob eine Hypothese beibehalten oder verworfen werden soll (vgl. ebd.).

„Sie [die Metaanalyse, Anm. AP] bedient sich quantitativ-statistischer Methoden, um die zentrale Tendenz und die Variabilität der Ergebnisse von empirischen Untersuchungen zu analysieren sowie die Variabilität der Resultate zu erklären. Wo der Rückgriff auf quantitative Techniken nicht möglich oder sinnvoll ist, werden qualitative Verfahren eingesetzt“ (Erik Farin 1994:7).

Eine Metaanalyse wird grob in 6 Phasen durchgeführt:

1. Explizite Formulierung der Fragestellungen;
2. Systematische Erfassung und Erhebung der empirischen Primärbefunde;
3. Kodierung und Bewertung inhaltlicher und methodischer Studienmerkmale;
4. Aggregation der Primärbefunde;
5. Interpretation der Ergebnisse;
6. Darstellung und Präsentation der Ergebnisse.
(Stephan Jeff Rustenbach 2003:XVII).

Anwendungsfelder metaanalytischer Verfahren sind vor allem Medizin, Psychologie und empirische Sozialforschung (ausführlich dazu Reiner Fricke & Gerhard Treinies 1985:21ff). Metaanalysen werden meist dann durchgeführt, wenn die Stichproben der Primärstudien zu klein sind, um verlässliche Ergebnisse zu liefern. In diesem Fall können mehrere Untersuchungen zusammengefasst werden und das Ergebnis konkretisieren (vgl. Wikipedia⁴³).

⁴³ <http://de.wikipedia.org/wiki/Metaanalyse>

7.1.1 Probleme der Metaanalyse⁴⁴

Probleme bei der Metaanalyse liegen u.a. in der Auswahl der Primärstudien. Es müssen geeignete Ein- bzw. Ausschlusskriterien formuliert werden, um einer sogenannten Publikationsverzerrung vorzubeugen. Nach Stephan Jeff Rustenbach (2003) ist dies die „bedeutendste potenzielle Validitätseinschränkung“ (a.a.O.:38). Man unterscheidet dabei zwischen objektiver und subjektiver Publikationsverzerrung. Ersteres ist dann gegeben, wenn die Publikationsentscheidung von der Signifikanz der Ergebnisse abhängig gemacht wird (vgl. ebd.). So bezeichnet z. B. das so genannte File-Drawer-Problem die Tendenz vieler WissenschaftlerInnen, nur erwartungskonforme Ergebnisse zu veröffentlichen „und andere in der Schublade [file-drawer] zu lassen“ (Bernhard Bührlen 1999:2). Eine subjektive Publikationsverzerrung entsteht dann, wenn AutorInnen insignifikante Ergebnisse gar nicht erst darstellen (vgl. Stephan Jeff Rustenbach 2003:38). Das „file-drawer-Problem“ kann reduziert werden, wenn sogenannte graue Literatur in die Untersuchung einbezogen wird. Darüber hinaus empfiehlt sich die Berechnung der „Fail-Safe-N“-Zahl nach Robert Rosenthal (1979). Diese Zahl gibt an, wie viele nicht-signifikante Studien erforderlich wären, damit das Gesamtergebnis der Metaanalyse falsch bzw. statistisch unbedeutend wäre (vgl. Bernhard Bührlen 1999:3).

Ein weiteres Problem bei der Metaanalyse stellt das „Apples and Oranges⁴⁵“- Problem (Uniformitätsproblem) dar. Es benennt die oftmals nicht in ausreichendem Maße gegebene Vergleichbarkeit der Primärstudien. Dies kann verschiedene Gründe haben – unterschiedliche Auswertungsmethoden, unterschiedliche Eigenschaften der Stichprobe, unterschiedliche Operationalisierung usw. (vgl. Martin Eisend 2004:20). Eine Lösung dieses Problems sieht beispielsweise Bernhard Bührlen (1999:3) darin, sich auf vergleichbare Merkmale der Studien zu beschränken, die so auch auf alle Studien anwendbar sind. In der Analogie der Benennung des Problems würde man also nicht mehr von „Äpfeln und Orangen“, sondern allgemein von „Obst“ sprechen. Dabei müssten aber gleichzeitig strengere Selektionskriterien angewendet werden.

Metaanalysen wird überdies der Vorwurf gemacht, dass sie Studien unterschiedlicher Qualität miteinander in Beziehung setzen, dass also jede Einzelstudie unabhängig von ihrer methodischen Qualität in der Untersuchung berücksichtigt wird. Dieses Problem wurde unter der Parole „garbage-in, garbage-out“ (Hans J. Eysenck 1978) bekannt – also, wo nur „Schrott“ hineinkommt, kann auch nur „Schrott“ herauskommen (vgl. Bernhard Bührlen 1999:2, Martin Eisend 2004:22f.). Andreas Beelmann und Thomas Bliesener (1994:218) empfehlen, diesem Problem durch getrennte Analysen (methodisch gute Studien versus methodisch schlechte Studien) zu begegnen und im Anschluss zu kontrollieren, inwiefern sich die Treatmenteffekte unterscheiden.

⁴⁴ Eine detaillierte Übersicht über die Probleme und Strategien bei metaanalytischen Untersuchungen ist bei Andreas Beelmann & Thomas Bliesener (1994) zu finden.

⁴⁵ In deutscher Sprache müsste man wahrscheinlich vom „Apfel und Birnen“-Problem sprechen (vgl. Reiner Fricke & Gerhard Treinies 1985:170).

Einen wichtigen Kritikpunkt stellt Reiner Fricke (1991) dar, indem er feststellt, dass die üblichen Effektivitätsstudien „nur absoluten oder relativen Lernerfolg in summarischer Form auflisten“ (a.a.O., zit. nach Patrick Blumschein 2004:11). Auf diese Weise können „weder Aussagen über differentielle Methodeneffekte noch Erklärungsmodelle für das Zustandekommen dieser Effekte geliefert“ (ebd.) werden, was dazu führt, dass viele dieser Studien für die Praxis nicht hilfreich sind.

7.2 Meta-Evaluation

Meta-Evaluationen haben das Ziel, zu einer Verbesserung der Qualität von Evaluationen und der Praxis der Evaluationsforschung beizutragen (vgl. Thomas Widmer 1996:3f.).

„Etwas ausführlicher können exemplarisch folgende mögliche Zielsetzungen angeführt werden: die Einschätzung der Qualität, der Wirkung und der Verwendung von Evaluationsstudien zuhanden der Beteiligten oder interessierter Dritter; die Untersuchung der Charakteristik eines evaluativen Prozesses; die Vermeidung möglichen Missbrauchs; die Begutachtung evaluativer Arbeit; die Aufdeckung und Kontrolle von Verzerrungen in Evaluationen; die Einschätzung der Nützlichkeit neu entwickelter Evaluationsansätze“ (a.a.O.).

Unter dem Begriff Meta-Evaluation wird also die Bewertung von Evaluationsstudien verstanden, genauer – die Evaluation von Evaluationen. Meta-Evaluationen sind also Konstruktionen dritter Ordnung⁴⁶ (vgl. a.a.O.:9).

Im Gegensatz zu Metaanalysen, die Primärstudien inhaltlich miteinander vergleichen, haben Meta-Evaluationen eher das Ziel, die Evaluationsverfahren der Primärstudien zu bewerten und somit die Qualität dieser Studien sicherzustellen.

„Nach Cook und Gruder (1978) haben Metaevaluationen zum Ziel, inhaltliche Fragestellungen auf der Basis einer Reanalyse von Evaluationsstudien zu beantworten, die Qualität von Evaluationen und damit deren Aussagekraft zu bewerten und Evaluationsstudien zu untersuchen, um Fragestellungen zum Forschungsprozess zu beantworten“ (Bernhard Bührlen 1999:4).

Thomas D. Cook und Charles L. Gruder (1978) weisen jedoch ausdrücklich darauf hin, dass sich der Begriff Meta-Evaluation ausschließlich auf die Evaluation von empirisch-summativen Evaluationen bezieht – „studies where the data are collected directly from program participants within a systematic design framework“ (a.a.O.:6).

Bei der Durchführung einer Meta-Evaluation können verschiedene Aspekte eine Rolle spielen, wie beispielsweise „die wissenschaftliche Qualität, die Praxisrelevanz, die Nutzung, die Wirkungen usw. einer oder

⁴⁶ Vor dem Hintergrund, dass es sich bei Evaluationen um Konstruktionen zweiter Ordnung handelt, die sich mit Konstruktionen der ersten Ordnung (der sozialen Realität) auseinandersetzen (vgl. a.a.O.).

mehrerer Evaluationen“ (Thomas Widmer 1996:4). Bei der Meta-Evaluation sind sowohl summative als auch formative Verfahren üblich.

7.2.1 Evaluationsmodelle

Metaevaluationen haben sich – so scheint es in der Forschungslandschaft der Fall zu sein – bisher kaum als eigene Forschungsdisziplin etablieren können.

„Meta-evaluation has many definitions. Some use it to describe aggregating information from several individual evaluations, others define it as a systematic tool for the quality control of evaluation studies“ (Petrei Uusikylä & Petri Virtanen 2000:50).

Thomas D. Cook & Charles L. Gruder (1978) leiten aus der Notwendigkeit, die Qualität von Primärstudien zu sichern, Modelle von Meta-Evaluationen ab. Dabei legen sie acht generelle Probleme von (Primär-)Evaluationen zu Grunde:

„(1) In deciding on research questions, evaluators and their sponsors rarely consult formal decision makers, let alone a wide range of interested constituencies. Instead, general program goals are outlined that are either operationally vague or of limited importance to many potential users. (2) It is still the case today that empirical summative evaluations are conducted by the research staff of the program being evaluated, while other evaluations are monitored by the same office that provided funds for the program development. Neither of these arrangements is satisfactory, because the evaluation sponsors have a vested interest in the results (Scriven, 1975). (3) Most evaluation research is conducted by profit-making, or non-for-profit, contract research agencies that are supposed to be more efficient than the available alternatives and to have fewer hidden agendas (i.e., publishing in disciplinary journals, teaching, or serving in professional organizations). [...]. (4) Many program and project administrators know that evaluations usually reach negative conclusions about impact. But whether they do or do not know this, they all seem to resent the additional tasks and complexities that can arise in cooperating with an evaluation. It is not surprising, therefore, that many administrators do not want to provide the level of cooperation that is necessary for a high-quality evaluation. (5) Inadequate reviews of evaluation research proposals mean that much research goes into the field with less than adequate sampling, design or measurement [...]. Moreover, it is questionable whether enough initial thoughts put into anticipating problems that might arise in the field and into developing practical means whereby these problems can be detected at a date early enough so that realistic fallback positions can be adopted. (6) Whatever the level of initial planning, unforeseen problems will arise and threaten the methodological integrity of the research. Therefore, a monitoring system is required to detect problems early so that constructive responses can be made promptly. (7) Deadline pressures, which are endemic to evaluation research, hinder careful planning and checking at all stages of the research [...]. (8) It is difficult to maintain the planned experimental contrasts. This can be because the program changes during the evaluation, because clients are free to choose how much treatment they want to receive, and because control group members often receive some of the treatment that they were not supposed to get“ (a.a.O.:15f.).

Sie betonen jedoch, dass die Modelle der Meta-Evaluation dabei helfen sollen, einige dieser Schwierigkeiten zu verringern, diese aber keine universellen Lösungen darstellen (vgl. a.a.O.:16).

Grundsätzlich lassen sich nach Thomas D. Cook & Charles L. Gruder (1978:17) sieben Modelle von Meta-Evaluationen unterscheiden, die in Tabelle 2 dargestellt sind.

zeitlich gesehen nach einer primären Evaluation	Daten nicht nochmals neu analysiert	Einzelner Datensatz	1. Essay-Review eines Evaluationsberichts
		Mehrere Datensätze	2. Review der Literatur über ein spezifisches Programm
	Daten nochmals neu analysiert	Einzelner Datensatz	3. Empirische Re- Evaluation einer Evaluationsstudie oder eines Programms
		Mehrere Datensätze	4. Empirische Re- Evaluation mehrerer Datensätze über das gleiche Programm
zeitlich gesehen simultan mit der primären Evaluation	Daten nicht parallel, neu analysiert	Einzelner oder mehrere Datensätze	5. Beratende Meta- Evaluation
	Daten parallel, neu und unabhängig analysiert	Einzelner Datensatz	6. Simultane Sekundäranalysen der Rohdaten
		Mehrere Datensätze	7. Mehrere unabhängige Replikationen

Tabelle 2: Modelle der Meta-Evaluation nach Thomas D. Cook & Charles L. Gruder (1978:17, nach der Übersetzung von Werner Wittmann 1985:404)

Die Modelle wurden mit Hilfe einer 3-faktoriellen Varianzanalyse generiert. Dabei ist ein Faktor, ob die Meta-Evaluation gleichzeitig mit der Primärevaluation durchgeführt wird oder im Anschluss daran. Der zweite Faktor beschreibt, ob die Primärdaten von den EvaluationsInnen manipuliert⁴⁷ wurden oder nicht, während sich der dritte Faktor auf die Anzahl von unabhängigen Datensätzen bezieht (vgl. ebd.⁴⁸).

⁴⁷ Im Sinne von „neu analysiert“

⁴⁸ Übersetzung AP

7.2.2 Evaluationstandards

Meta-Evaluationen liegen – wie anderen Evaluationen auch – Kriterien zu Grunde, nach denen die einzelnen Evaluanden bewertet werden. Die Deutsche Gesellschaft für Evaluation (2002) weist in ihren Richtlinien zur Meta-Evaluation darauf hin, dass

„an eine **Meta-Evaluation** [...] vom Grundsatz die gleichen Anforderungen zu richten [sind, Anm. AP] wie an eine primäre Evaluation: Sie sollen nützlich, durchführbar, fair und genau sein. Es ist bei der Planung einer Meta-Evaluation zu prüfen, welche Standards anzuwenden sind. Auch ist zwischen formativen und summativen Meta- Evaluationen zu unterscheiden. Zentral ist, dass Meta-Evaluationen auf einer empirischen Grundlage durchgeführt werden. In erster Linie kommen hierfür Inhaltsanalysen von Berichten und Dokumenten, Instrumenten und Materialien der zu beschreibenden und zu bewertenden Evaluation in Betracht. Darüber hinaus können auch Interviews, Gruppendiskussionen oder Fragebogen angemessen sein“ (a.a.O.:2, Hervorhebung im Original).

Mit anderen Worten: an eine Meta-Evaluation sind die gleichen Anforderungen zu richten, die auch bei primären Evaluationen Anwendung finden. Zur Bewertung der methodischen Qualität von Primärstudien bedienen sich also auch Meta-Evaluationen häufig Standards, die für die Durchführung von Evaluationen entwickelt wurden (vgl. Bernhard Bührlen 1999). Dies sind zum Beispiel

- die Standards der Evaluation Research Society (vgl. Evaluation Research Society Standards Committee 1982, vgl. Rossi 1982),
- des Joint Committee on Standards for Educational Evaluation (vgl. Committee on Standards for Educational Evaluation 1981; James R. Sanders & Wolfgang Beywl 2005),
- die Evaluation Principles der American Evaluation Association (AEA) (vgl. American Evaluation Association 1994) oder
- der Deutschen Gesellschaft für Evaluation (vgl. Deutsche Gesellschaft für Evaluation e.V. 2001).

So ist beispielsweise⁴⁹ in den Standards der American Evaluation Association die Rede von vier Hauptprinzipien bzw. Leitsätzen, die an Evaluationen anzulegen sind. Sie betreffen

1. die Systematik der Untersuchung (Evaluationen sollten systematisch und auf Grund von Daten erfolgen),
2. die Kompetenz (EvaluatorsInnen sollten vor der Interessengemeinschaft die nötige Kompetenz aufweisen), Integrität/ Ehrlichkeit (EvaluatorsInnen sollen die nötige Integrität

⁴⁹ Ich möchte an dieser Stelle darauf verzichten, sämtliche Evaluationsstandards aufzulisten. Zur Festigung der Argumentation und Diskussion werden nur eine englisch- und eine deutschsprachige Publikation näher beschrieben.

und Ehrlichkeit im gesamten Prozess der Evaluation sicherstellen),

3. den Respekt (EvaluatorsInnen sollen die Sicherheit, Würde und den Selbstwert der an der Evaluation Beteiligten sicherstellen) und
4. die Verantwortlichkeit (EvaluatorsInnen sollen die Vielfalt von Interessen und Werten bzw. Wertvorstellungen berücksichtigen, die mit dem generellen und öffentlichen Wohl verbunden sind) (vgl. American Evaluation Association 1994⁵⁰).

Diese Standards sind nicht als Regeln, sondern eher als Richtlinien für die Durchführung von Evaluationen zu verstehen. Ebenso hat die Deutsche Gesellschaft für Evaluation (2001) Standards veröffentlicht, nach denen Evaluationen vor allem Prinzipien der Nützlichkeit, Durchführbarkeit, Fairness und Genauigkeit erfüllen sollten.

„Die Nützlichkeitsstandards sollen sicherstellen, dass die Evaluation sich an den geklärten Evaluationszwecken sowie am Informationsbedarf der vorgesehenen Nutzer und Nutzerinnen ausrichtet“ (ebd.). Dazu gehören:

1. Identifizierung der Beteiligten und Betroffenen,
2. Klärung der Evaluationszwecke,
3. Glaubwürdigkeit und Kompetenz des Evaluators / der Evaluatorsin,
4. Auswahl und Umfang der Informationen,
5. Transparenz von Werten,
6. Vollständigkeit und Klarheit der Berichterstattung,
7. Rechtzeitigkeit der Evaluation,
8. Nutzung und Nutzen der Evaluation (ebd.).

Mit den Durchführbarkeitsstandards soll gewährleistet werden,

„dass eine Evaluation realistisch, gut durchdacht, diplomatisch und kostenbewusst geplant und ausgeführt wird“ (ebd.). Die Durchführbarkeitsstandards umfassen:

1. Angemessene Verfahren,
2. Diplomatisches Vorgehen,

⁵⁰ Übersetzung AP

3. Effizienz von Evaluation (ebd.).

Die Fairnessstandards, zu denen

1. Formale Vereinbarungen,
2. Schutz individueller Rechte,
3. Vollständige und faire Überprüfung,
4. Unparteiische Durchführung und Berichterstattung und
5. Offenlegung der Ergebnisse (ebd.)

gehören, sollen „sicherstellen, dass in einer Evaluation respektvoll und fair mit den betroffenen Personen und Gruppen umgegangen wird“ (ebd.). Als letzte Gruppe seien noch die Genauigkeitsstandards angeführt, die die Hervorbringung und Vermittlung von gültigen Informationen und Ergebnissen des entsprechenden Evaluationsgegenstandes bzw. Fragestellungen der Evaluation sichern sollen (vgl. ebd.). Diese umfassen:

1. Beschreibung des Evaluationsgegenstandes,
2. Kontextanalyse,
3. Beschreibung von Zwecken und Vorgehen,
4. Angabe von Informationsquellen,
5. Valide und reliable Informationen,
6. Systematische Fehlerprüfung,
7. Analyse qualitativer und quantitativer Informationen,
8. Begründete Schlussfolgerungen und
9. Meta-Evaluation (ebd.).

Vergleicht man die Standards der einzelnen Organisationen miteinander, so fällt auf, dass diese in einigen Bereichen nicht vollkommen trennscharf sind oder sich gegenseitig ergänzen. Festzuhalten bleibt, dass Standards, die generell an Evaluationen angelegt werden, auch für Meta-Evaluationen Gültigkeit haben.

Um Evaluationen zu evaluieren, ist allerdings von Kriterienkatalogen in Form von Checklisten abzusehen, weil sie oft nur eingeschränkte Problemstellungen bearbeiten oder keine breite Anwendung finden (vgl. Thomas Widmer 1996:10). Thomas Widmer (a.a.O.:9) schlägt daher folgende Funktionen eines Kriterienrasters vor:

1. Methodologische Offenheit: Die Kriterien müssen so ausgestaltet sein, dass sie in der Lage sind, Evaluationen unterschiedlichster methodologischer Ausrichtungen zu bewerten. Kriterien, die sich auf spezifische

- erkenntnistheoretische Positionen abstützen, sind nicht geeignet. Der Kriterienraster muss gegenüber den verwendeten Evaluationsansätzen möglichst offen sein.
2. Umfassende Bewertung: Weiterhin sollte die Bewertung einer Evaluation als Ganzes ermöglicht werden. Kriterien, die lediglich Teilaspekte (Nutzung, Wirkung, Methodik, Validität etc.) abdecken, reichen dazu nicht aus.
 3. Weite Verbreitung, Anerkennung und Akzeptanz: Die Kriterien sollten möglichst weit verbreitet sein und über eine breite Anerkennung durch Vertreter unterschiedlicher Evaluationsansätze verfügen. Sie sollten zudem im weiten Evaluationsfeld auf eine hohe Akzeptanz stossen. Zur Beurteilung stützen wir uns auf die Anerkennung durch relevante Fachorganisationen sowie auf die einschlägige Literatur.
 4. Ausreichender Detaillierungsgrad: Der zu verwendende Kriterienraster sollte die verschiedensten der zu bewertenden Aspekte einer Evaluation einzufangen vermögen. Dazu ist eine grössere Zahl von Kriterien und eine ausreichend differenzierte Formulierung der einzelnen Kriterien erforderlich (ebd.).

Die Heterogenität der Ansätze macht es erforderlich, spezifische Fragestellungen für die einzelnen Studien zu entwickeln und vorhandene Standards den eigenen Forschungszwecken anzupassen (vgl. Bernhard Bührlen 1999:4). Eine besondere Schwierigkeit besteht darin, die Einzelbefunde zu gewichten, da diese in hohem Maße von der Qualität der zu Grunde liegenden Primärstudien abhängig sind. Das Problem ist mit dem Garbage-in/Garbage-out-Problem vergleichbar, das ich bereits für Metaanalysen dargestellt habe (ebd.). Zusammenfassend lässt sich unter Bezugnahme auf Bernhard Bührlen-Armstrong (1998:113) festhalten, dass theoretische Konzepte zu metaevaluativer Forschung sehr rar sind. Normative Standards sind nur auf sehr allgemeinen Ebenen in der Lage, die Bedingungen der einzelnen Untersuchungen widergeben zu können, was eine Anpassung schwierig macht.

In Kapitel „Modelle der Meta-Evaluation“ ab Seite 77 werde ich darauf eingehen, welche Bewertungskriterien und welches Modell nach der Systematik von Thomas D. Cook & Charles L. Gruder (1978)⁵¹ der vorliegenden Arbeit zu Grunde liegen und wie die Kriterien abgeleitet wurden.

⁵¹ Vgl. Tabelle 2: Modelle der Meta-Evaluation nach Thomas D. Cook & Charles L. Gruder (1978:17, nach der Übersetzung von Werner Wittmann 1985:404) auf Seite 70.

8 Meta-evaluative Untersuchung

Dieses Kapitel verbindet die theoretischen Überlegungen mit der Forschungsfrage und stellt somit den empirisch-praktischen Teil der Arbeit dar. Anhand von 11 Primärevaluationen, die sich mit dem Zusammenhang von Lernerfolg und Lernsetting (Online- bzw. Präsenzkurs) beschäftigen, soll in einer Meta-Evaluation gezeigt werden, welche Stärken und Schwächen solche Evaluationen haben. Besonderes Augenmerk richtet sich dabei auf die Definition von Lernerfolg. Es ist an dieser Stelle wichtig zu betonen, dass es nicht darum geht, ein möglichst repräsentatives Resultat zu erzielen. Die vorgestellten Studien sprechen für sich und sollen nicht als Vertreterinnen einer breiten Masse von Evaluationsstudien stehen. Dennoch werden sie exemplarisch angeführt, um die Argumentation zu belegen und indifferente Sichtweisen von Lernerfolg und dessen Operationalisierung aufzuzeigen.

8.1 Fragestellung und Hypothesen

Die Hauptforschungsfrage der vorliegenden Dissertation lautet: Worin liegen die Probleme bei Vergleichsuntersuchungen in Bezug auf mediengestütztes Lernen? Wie können sie umgangen werden?

Aus der theoretischen Diskussion der vorherigen Kapitel leiten sich folgende Hypothesen für die meta-evaluative Untersuchung ab:

1. Lernerfolg ist nicht eindeutig operationalisierbar.
2. Unspezifische Vergleiche von Online- versus Präsenzlernen sind nicht sinnvoll anwendbar bzw. nicht zu verallgemeinern.
3. Die Methode der Metaanalyse ist nur bedingt geeignet, um Rückschlüsse auf den Lernerfolg zuzulassen.

Zur Klärung der Fragestellung werden 11 Primärstudien mittels meta-evaluativen Verfahren miteinander verglichen und ausgewertet. Indikatoren, die die Forschungsfrage beantworten sollen, leiten sich aus folgenden Überlegungen ab:

- Wie wird der Lernerfolg in den verschiedenen Studien definiert?
- Wird das Konstrukt Lernerfolg in Dimensionen aufgegliedert?
- Wie wird Lernerfolg operationalisiert?
- Werden Lernerfolg, Lernprozess und Lernleistung unterschieden?
- Welches Testverfahren wurde der Untersuchung zu Grunde gelegt?

- Kommen ältere und neuere Studien zu gleichen Ergebnissen?
- Gibt es Aspekte, die sich verallgemeinern lassen?

Darüber hinaus soll die bereits vorgestellte Lernzieltaxonomie nach Lorin W. Anderson & David R. Krathwohl (2001) herangezogen werden. Wie bereits im theoretischen Teil erläutert, ist die Feststellung der Leistung mit Hilfe desselben Testverfahrens nur bedingt zulässig. Wenn jedoch die Lernziele vergleichbar sind, ist auch ein Vergleich des Lernergebnisses möglich. Um einen sinnvollen Vergleich von Untersuchungs- und Kontrollgruppe in Vergleichsstudien zu ermöglichen, sollten nur die Faktoren miteinander verglichen werden, die die selbe Ausprägung von Lernerfolg zu Grunde legen. Daraus ergeben sich zusätzlich folgende Fragestellungen:

- Was sind formulierte und getestete Lernziele in den Primärstudien?
- Welcher Test wurde der Messung des Lernerfolgs zu Grunde gelegt?
- Werden die gleichen Ausprägungen von Lernerfolg miteinander verglichen?

Es sei jedoch an dieser Stelle noch folgendes festzuhalten:

Innerhalb der Studie sollen die Prozesse und Ergebnisse von Evaluationen bewertet werden, nicht jedoch die EvaluatorenInnen selbst. Wie Thomas Widmer (1996) bereits anmerkt, ist es „nicht zulässig, aufgrund der Bewertung einer Evaluation auf personelle Qualitäten einzelner Personen zu schliessen“ (a.a.O.:6). Er führt weiterhin an:

„Zweitens richtet sich die Meta-Evaluation nur auf je eine Evaluation eines Evaluators (oder eines Evaluationsteams). Erkenntnisse aus der Untersuchung eines Projektes lassen sich nicht auf andere Evaluationen übertragen. Ebenso wenig lässt sich daraus auf die individuellen Qualitäten der beteiligten Personen schliessen. Zu beachten ist, dass jede Evaluation in einem spezifischen Kontext erarbeitet wurde. So sind beispielsweise die Handlungsspielräume der Evaluatoren sehr oft stark und in unterschiedlicher Weise begrenzt. Fragestellungen einer Evaluation etwa, oder ihr Untersuchungsbereich werden sehr häufig vom Auftraggeber definiert. Die Gegenstandsbereiche von Evaluationsstudien, ebenso wie zum Beispiel die Datenlage, weisen stark unterschiedliche Charakteristiken auf. Weiter führen auch die finanziellen und zeitlichen Ressourcen, die zur Durchführung der Evaluation zur Verfügung stehen, zu deutlich divergierenden Rahmenbedingungen“ (ebd.)

Diese Aussagen möchte ich ausdrücklich unterstützen und auch für die vorliegende Arbeit betonen: Keine der einbezogenen Studien soll als schlechtes Beispiel herangezogen werden, es soll lediglich die Problematik

der Vergleichbarkeit thematisiert werden. Alle zu Grunde liegenden Primärstudien wurden in einem wissenschaftlichen Kontext ausgewählt.

8.2 Methodisches Vorgehen

Um die einzelnen Primärstudien miteinander zu vergleichen, wird eine Meta-Evaluation durchgeführt. Dabei ist anzumerken, dass es nicht darum geht, möglichst repräsentative Daten zu erhalten, sondern darum, bestehende Probleme argumentativ aufzuzeigen. Aus diesem Grund kann auch auf die Bewertung der einzelnen Primärstudien durch mehrere RaterInnen verzichtet werden. Überdies wird bei der Analyse von Evaluationen der Fokus auf die Forschungsfrage gelegt, welche Konsequenzen genau diese Evaluation für die Interpretation von Lernerfolg und Lernerfolgskonzepten hat. Es handelt sich also um eine spezifische Meta-Evaluation, die die Primärstudien in Bezug auf diese speziellen Kriterien miteinander vergleicht und keine „globale“ bzw. umfassende Meta-Evaluation, die detaillierte Aussagen zu Einzelbefunden treffen will. Ein Vergleich von Einzelbefunden ist darüber hinaus auch nur bedingt sinnvoll, da die Primärstudien selbst eine gewisse Heterogenität aufweisen. Zwar wurden alle nach denselben Merkmalen ausgewählt, doch handelt es sich bei dem zur Bewertung dieser Studien verfügbaren Material um verschiedenen Arten von Publikationen.

Die Ergebnisse dieser Arbeit können also bestenfalls Hinweise, „aber keine empirisch überprüften (erhärteten) Aussagen liefern“ (Thomas Widmer 1996:61).

8.2.1 Modelle der Meta-Evaluation

Es ist nicht ohne Weiteres möglich, Modelle vorhandener Meta-Evaluationen auf die vorliegende Arbeit zu übertragen. So stellt beispielsweise die Arbeit von Thomas Widmer (1996) eine Bewertung von kompletten Evaluationsstudien dar. Der Autor untersucht 10 Primärstudien, verfolgt jedoch nicht die Absicht, die Ergebnisse inhaltlich zu integrieren, sondern die, ein Kriterieninventar zu erstellen. Darüber hinaus sind in seiner Meta-Evaluation keine Studien aus dem Bildungsbereich vertreten (vgl. auch Bernhard Bührlen-Armstrong (1998:122).

Da die vorliegende Arbeit die Ergebnisse der Primärstudien auch inhaltlich zusammenzufassen versucht, kann die Methodik von Thomas Widmer (1996) nur in begrenzter Form als Modell genutzt werden.

Auch die Arbeit von Bernhard Bührlen-Armstrong (1998) kann nicht im vollen Umfang übertragen werden, obwohl methodisch eine größere Nähe vorliegt. In seiner Untersuchung werden 33 Primärstudien zum Thema AIDS-Prävention analysiert und integriert. Der Autor legt hierbei das „evaluation audit“-Verfahren nach Thomas A. Schwandt & Edward S. Halpern (1988) zu Grunde. Dazu ist jedoch eine genaue Betrachtung der Primärstudien inklusive aller verfügbaren Unterlagen

(Projektberichte, Protokolle etc.) erforderlich (vgl. Bernhard Bührlen-Armstrong (1998:114). Dies ist in diesem Fall allerdings nicht möglich.

Auch die vorgestellten Ansätze zur Meta-Evaluation können nur in modifizierter Form angewendet werden (vgl. a.a.O.:120f.). Wird die beschriebene Systematik nach Thomas D. Cook & Charles L. Gruder (1978, vgl. Tabelle 2 auf Seite 70) zu Grunde gelegt, eignen sich die Typen 1 und 2 für die vorliegende Arbeit.

„Diese Art von Reviews sind v.a. dann nützlich, wenn sie die Schlussfolgerungen der Primärstudien nacheinander in Frage stellen und die Gesamtheit der Evidenz über alle Primärstudien abwägen“ (Bernhard Bührlen-Armstrong 1998:121).

Im Folgenden werden diese beiden Typen etwas genauer betrachtet, um abschließend klären zu können, welches Meta-Evaluationsmodell zu Grunde gelegt werden kann.

Typ 1 beschreibt Reviews von Evaluationsberichten⁵².

„[...] Reviews are most useful when they break down the major conclusions one by one, and then separately focus on the totality of support for each conclusion“ (Thomas D. Cook & Charles L. Gruder 1978:19).

Ohne die Daten generell in Frage zu stellen, können durch Reviews von Evaluationsberichten neue Interpretationen einer Annahme hervorgebracht werden. Die Autoren betonen jedoch, dass es notwendig sei, alle genannten Anhaltspunkte für oder gegen eine vermutete Annahme aufzuzeigen. Ebenfalls sind solche Arten von Meta-Evaluationen dann sinnvoll, wenn spezifische Schwächen in der Primärevaluation mit Hilfe der gewonnenen Befunde überwunden werden können, oder wenn die Ergebnisse in einen politischen oder theoretischen Kontext gestellt werden sollen (vgl. a.a.O.:20ff).

Dennoch leiden Reviews unter der fundamentalen Schwäche, dass die Daten nicht reanalysiert werden können.

„Consequently, specific validity threads that escaped the attention of the primary evaluators cannot be empirically examined in order to estimate their impact or to conduct analyses free of their impact“ (a.a.O.:22).

Meta-Evaluationen können die Plausibilität der empirischen Daten in den Primärstudien lediglich abschätzen. Da diese eigenen Bewertungen die Validität nicht klären können, tragen sie eher zu größerer Verwirrung bei (ebd., vgl. auch Bernhard Bührlen-Armstrong 1998:121).

⁵² Thomas D. Cook und Charles L. Gruder (1978: 19) unterscheiden zwischen veröffentlichten und unveröffentlichten Evaluationsberichten. In dieser Arbeit werden nur veröffentlichte Studien analysiert, daher entfällt der zweite Punkt an dieser Stelle.

Typ 2 bezieht sich auf ein Review der Literatur zu einem spezifischen Programm. Thomas D. Cook und Charles L. Gruder (1978) führen als Beispiel eine Meta-Evaluation an, in die verschiedene Primärstudien zur Evaluation des Kinderfernsehprogramms „Sesame Street“ eingegangen sind (vgl. a.a.O.:24). Dennoch kann der Begriff „Programm“ an dieser Stelle auch weiter gefasst werden und sich auf bestimmte Aussagen beziehen. Werden beispielsweise mehrere Datenquellen zu einer bestimmten Forschungsfrage analysiert, handelt es sich ebenso um diesen Typ von Meta-Evaluation (vgl. Bernhard Bührlen-Armstrong (1998:121).

„Multiple evaluations permit more valid and credible research conclusions than does a single evaluation, especially when clear patterns of convergence are found between independent studies that used different versions of a treatment, different kinds of respondents, and different research settings“ (Thomas D. Cook & Charles L. Gruder 1978:24).

Die zur Verfügung stehenden Untersuchungen können sich jedoch aus einer verzerrten Stichprobe zusammensetzen, da manche Projektberichte nicht publiziert werden oder nur über Agenturen erhältlich sind, zu denen der Zugang nicht immer leicht ist. Überdies variiert die Auswahl, welche Studien auf Grund von inadäquater Dokumentation oder Methodologie von der Untersuchung ausgeschlossen werden. Ein weiteres Problem dieses Typs betrifft die Kriterien zur Integration und Gewichtung der Befunde aus den Einzelstudien. Die Autoren geben dafür jedoch selbst einen Lösungsvorschlag an:

„Our own position is that the reviewer should consider all studies meeting basic methodological standards as valid and should look for *a posteriori* hypotheses that fit the pattern of positive, negative, and null findings. This detective-like task will be all the easier if the reviewers ask: ‚What different kinds of effect do different kinds of projects have on different kinds of people?‘ and if they do not expect every single study to fit neatly into the interpretative framework they develop for the data (a.a.O.:25, Hervorhebungen im Original).

Als letztes Problem sei die Schwierigkeit der Interpretation von unterschiedlichen Ergebnissen der Einzelstudien genannt.

Dennoch ist diese Art von Review nach Meinung der Autoren mehr theoriegeleitet als die meisten anderen Aspekte der Evaluationsforschung, da sie einerseits die Fragen des Reviews selbst aus der Theorie ableiten, wie auch die Muster zur Interpretation und Bewertung der Befunde. Da Reviews Theorien erforschen und ausweiten können, ist eine hohe Attraktivität innerhalb der Sozialwissenschaften gegeben.

Zusammengefasst handelt es sich bei der vorliegenden Arbeit um eine Meta-Evaluation, die als Mischform der beiden vorgestellten Modelle anzusehen ist. Sie wird erst nach Abschluss der Primärevaluationen durchgeführt, wobei die Daten keiner erneuten empirischen Analyse unterzogen werden. Darüber hinaus werden zum Einen sowohl Evaluationsberichte als auch wissenschaftliche Publikationen (wie

beispielsweise Dissertationen und Zeitschriftenartikel) in die Analyse einbezogen, zum Anderen sind alle Primärstudien der Forschungsfrage entsprechend ausgewählt worden, d.h. alle in die Meta-Evaluation einbezogenen Studien sind Untersuchungen zur selben Art von Programm, da sie alle ein bestimmtes Lernsetting evaluieren.

8.2.2 Auswahl geeigneter Primärstudien

Bevor eine Recherche nach Primärstudien durchgeführt werden kann, ist es notwendig, Kriterien zu formulieren, die diese Studien aufweisen müssen, um in die Untersuchung einbezogen zu werden. In der vorliegenden Arbeit wurden folgende Abgrenzungen der Grundgesamtheit definiert:

- Es muss sich um eine evaluative Untersuchung handeln, die jedoch den Anspruch erhebt, mit sozialwissenschaftlichen Methoden zu arbeiten.
- Die Studie vergleicht mindestens zwei Lernsettings miteinander, i.d.R. Präsenzlernen vs. softwaregestütztes Lernen⁵³.
- Die Studie macht Aussagen zum Lernerfolg bzw. erhebt den Anspruch der Lernerfolgsmessung.
- Die Untersuchung weist ein Evaluationsdesign mit Experimental- und Kontrollgruppe auf.
- Die Untersuchung muss bereits abgeschlossen sein.
- Die Publikation muss in englischer oder deutscher Sprache verfasst worden sein.

Um über eine möglichst breite Auswahl an Evaluationsstudien verfügen zu können, wurde in folgenden Datenbanken recherchiert:

- Eric Database
- Subito
- AACE Digital Library
- Education Index
- No significant difference website⁵⁴
- Digitale Bibliothek
- Sonstige Quellen (z.B. Anfrage in der Mailingliste Forum Evaluation)

⁵³ Es muss sich nicht ausschließlich um Online-Lernen handeln, Untersuchungen, die sich mit CBT oder anderen Offline-Formen computergestützten Lernens beschäftigen, werden ebenfalls einbezogen.

⁵⁴ <http://www.nosignificantdifference.org/> letzter Aufruf 21.06.2007

Darüber hinaus wurde im Zeitraum von August 2005 bis August 2006 regelmäßig im Internet nach relevanten Studien recherchiert. Dies geschah in erster Line unter Zuhilfenahme der folgenden Suchmaschinen:

- Google
- Google Scholar
- Metager

Dabei wurden folgende Suchbegriffe verwendet:

evaluation unterricht E-Learning
evaluation unterricht online
evaluation online lernerfolg
evaluation E-Learning lernerfolg
studie online signifikant
studie E-Learning signifikant
evaluation vergleich online lernerfolg
Studie Online Lehrveranstaltung Vergleich
lernerfolg e-learning traditionell vergleich
signifikant kurs online
signifikant kurs E-Learning
vergleich online klassenraum
study online significant
study E-Learning significant
evaluation online classroom significant
evaluation E-Learning classroom significant
evaluation study lesson online
evaluation study lesson E-Learning
evaluation online "learning outcomes"
study online significant
study comparison learning online
evaluation online significant
evaluation comparison online "learning outcomes"
significant course online
comparison online classroom
"significant difference" virtual classroom
Effectiveness web based comparison

Darüber hinaus wurde mit aufwändigeren Suchstrategien experimentiert. Dennoch muss festgehalten werden, dass letzten Endes die einfachen Strategien die besten Ergebnisse lieferten⁵⁵. Zwar stellen weder die Quellen noch die Studien selbst eine vollständige Erfassung aller relevanten Arbeiten zu diesem Thema dar, um jedoch den aufgestellten Hypothesen argumentativ begegnen zu können, wird die Auswahl als ausreichend betrachtet (vgl. Thomas Widmer 1996:55).

Im ersten Schritt konnten 40 Studien herausgefiltert werden. Aus der Liste dieser Resultate wurden anschließend die Studien selektiert, die den formulierten Kriterien entsprachen. Die verbleibenden 12 Studien wurden in Textdateien transformiert und in das Programm Atlas.ti zur qualitativen Datenanalyse eingelesen, aus der die Beschreibungen entstanden sind. Eine Studie musste anschließend entfernt werden, da sie sich auf drei Untersuchungen gleichermaßen bezog, diese allerdings nicht detailliert genug beschrieb.

Folgende 11 Studien werden in der vorliegenden Untersuchung als Primärstudien berücksichtigt:

⁵⁵ Es bleibt anzumerken, dass sich die meisten Funde auf englischsprachige Publikationen bezogen. Für den deutschsprachigen Raum sind entsprechende Studien – wenn vorhanden – hauptsächlich für den medizinisch-technischen Bereich auszumachen.

Übersicht über die Primärstudien

ID	Quelle	Art der Publikation
P1	Jones, Edward R. (1999): A comparison of an all web-based class to a traditional class.	Paper Fachvortrag
P2	Wisan, Gail; Nazma, Shirin; Pscherer, Charles P., Jr. (2001): Comparing online and Face-to-Face Instruction at a Large Virtual University: Data and Issues in the Measurement of Quality.	Paper Fachvortrag
P3	Martens, Rob; Portier, Stanley J. & Valcke, Martin (1995): Comparing study outcomes from different learning environments for statistics.	Forschungsbericht
P4	Smolle, Josef; Staber, Reinhard; Neges, Heide; Reibnegger, Gilbert;Kerl, Helmut (2005): Computer-based Training in der Dermatoonkologie – erste Ergebnisse zum Vergleich elektronischer Lernprogramme mit Präsenzlehre.	Wissenschaftlicher Artikel
P5	Davies, Randall S. & Mendenhall, Robert (1998): Evaluation comparison of online and classroom instruction for HEPE 129—Fitness and lifestyle management course, Department of Instructional Psychology and Technology, Brigham Young University, Utah.	Forschungsbericht

ID	Quelle	Art der Publikation
P6	Hartzoulakis, Vassilis (2002): Investigating Computer Mediated Instruction in a Greek State Lykeio.	Master-Arbeit
P7	Ryan, William J. (2002): Online and in the classroom: The numbers and what they <i>might</i> mean.	Paper Fachvortrag
P8	Eriksson, Thommy; Goller, Alois; Muchin, Sergei (2001): A comparison of online communication in distance education and in conventional education.	Paper Fachvortrag
P9	Green, Ruth & Gentemann, Karen (2001). Technology in the curriculum: An assessment of the impact of on-line courses.	Evaluationsbericht
P10	David J. Pucel & Thomas F. Stertz (2005). Effectiveness of and Student Satisfaction with Web-Based Compared to Traditional In-service Teacher Education Courses.	Wissenschaftlicher Artikel
P11	Larry S. Tinnerman (2006). A Comparative Study Between Traditional and Distance Education Instructional Environments Involving Two Graduate Level Learning Disabilities Classes.	Wissenschaftlicher Artikel

Tabelle 3: Übersicht über die Primärstudien

8.2.3 Auswahl und Festlegung geeigneter Kriterien

Wie bereits erwähnt, können verschiedene Evaluationsstandards auf den Bereich von Meta-Evaluationen übertragen werden (siehe Kapitel 7.2.2 zu Evaluationsstandards ab Seite 71), so dienen sie auch in diesem Fall als Grundlage. Ich habe bereits dargestellt, dass eine genaue Verwendung bestehender Kriteriensets m.E. nicht sinnvoll ist, da das Ziel der Untersuchung nicht darin besteht, eine große Repräsentativität der Daten zu erlangen. Vielmehr sollen bestehende Schwierigkeiten argumentativ aufgezeigt werden.

Betrachten wir die Standards der Deutschen Gesellschaft für Evaluation (2001), so wird deutlich, dass nicht alle Kriterien für die „Mikroebene“ der vorliegenden Arbeit in Frage kommen. Zwar können die Oberkategorien Nützlichkeit, Durchführbarkeit und Genauigkeit⁵⁶ durchaus beibehalten werden, im Detail kann jedoch auf einzelne Kriterien verzichtet werden. Aus dem Bereich der Nützlichkeit können verwendet werden:

1. Identifizierung der Beteiligten und Betroffenen,
2. Klärung der Evaluationszwecke.

Alle anderen Kriterien aus diesem Bereich scheiden aus, da eine Erhebung mit Hilfe der vorliegenden Daten nicht möglich ist.

Aus dem Bereich der Durchführbarkeitsstandards kann das Kriterium „angemessene Verfahren“ übernommen werden. Nicht berücksichtigt werden können die Kriterien „diplomatisches Vorgehen“ und „Effizienz von Evaluation“.

Um die Fairnessstandards sinnvoll erheben zu können, reicht ein Review von Evaluationsberichten und Literatur nicht aus, sondern es müssten sämtliche Unterlagen der Evaluationen vorliegen. Anders ist eine Beurteilung durch ein nachträgliches Review nicht möglich. Aus diesem Grund werden die Fairnessstandards („Formale Vereinbarungen“, „Schutz individueller Rechte“, „Vollständige und faire Überprüfung“, „Unparteiische Durchführung und Berichterstattung“ und „Offenlegung der Ergebnisse“) nicht berücksichtigt.

Als letzte Gruppe – den Genauigkeitsstandards – können folgende Kriterien übernommen werden:

1. Beschreibung des Evaluationsgegenstandes,
2. Kontextanalyse,
3. Beschreibung von Zwecken und Vorgehen,

⁵⁶ Wie später aufgezeigt werden wird, können die Fairnessstandards nicht auf die vorliegende Arbeit übertragen werden.

- 5. Valide und reliable Informationen,
- 8. Begründete Schlussfolgerungen.

Der bereits erwähnten Arbeit von Thomas Widmer (1996) wird der Kriterienkatalog des amerikanischen „Joint Committee on Standards for Educational Evaluation“ (1981) bzw. dessen Überarbeitung aus dem Jahr 1984 zu Grunde gelegt, die aus 30 Kriterien besteht. Auch dieses Bewertungsinstrument ist in 4 Gruppen gegliedert: Nützlichkeit, Anwendbarkeit, Korrektheit und Genauigkeit (nach Thomas Widmer 1996:15ff):

- 1. Nützlichkeit
 - 1.1. Identifikation der beteiligten Gruppen
 - 1.2. Glaubwürdigkeit des Evaluators
 - 1.3. Informationsrahmen und -auswahl
 - 1.4. Bewertende Interpretation
 - 1.5. Klarheit des Berichts
 - 1.6. Verbreitung des Berichts
 - 1.7. Rechtzeitigkeit des Berichts
 - 1.8. Wirkung der Evaluation
- 2. Anwendbarkeit
 - 2.1. Praktische Verfahren
 - 2.2. Politische Überlebensfähigkeit
 - 2.3. Kostenwirksamkeit
- 3. Korrektheit
 - 3.1. Formale Verpflichtung
 - 3.2. Deklaration von Interessenkonflikten
 - 3.3. Vollständige und aufrichtige Offenlegung
 - 3.4. Herstellung von Öffentlichkeit
 - 3.5. Schutz individueller Menschenrechte
 - 3.6. Menschliche Interaktionen
 - 3.7. Ausgewogene Berichterstattung
 - 3.8. Finanzielle Verantwortlichkeit
- 4. Genauigkeit
 - 4.1. Identifikation des Gegenstands
 - 4.2. Kontextanalyse
 - 4.3. Beschreibung von Zielen und Vorgehen
 - 4.4. Verlässliche Informationsquellen
 - 4.5. Valide Messung
 - 4.6. Reliable Messung
 - 4.7. Systematische Datenüberprüfung
 - 4.8. Analyse quantitativer Information
 - 4.9. Analyse qualitativer Information
 - 4.10. Gerechtfertigte Folgerungen
 - 4.11. Objektive Berichterstattung

Die einzelnen Kriterien weisen keine hohe Trennschärfe zu den Standards der Deutschen Gesellschaft für Evaluation (2001) auf. Es wird jedoch deutlich, dass sowohl das Joint Committee (1981) als auch Thomas Widmer (1996) den Fokus der Bewertung vor allem auf die Transparenz

von Handlungen, Entscheidungen und Bewertungen sowie deren Interpretationen legen.

„Problematisch scheint bei der Arbeit von Widmer der Einstufungsprozess, mit dem die Erfüllung der Kriterien beurteilt wird. Die Beurteilung erfolgt mittels eines diskursiven Prozesses. Dabei wurde jedoch nicht explizit auf einheitliche, vorgegebene Indikatoren zurückgegriffen, so dass die Einstufung auf den Kriterien nicht transparent wird. Vereinzelt wird aus der Tatsache, dass keine Informationen über Verstöße gegen ein Kriterium vorliegen, darauf geschlossen, dass das Kriterium als erfüllt (oder nicht erfüllt) angesehen werden kann. Diese Praxis ist als kritisch zu bewerten, statt dessen sollte bei mangelhafter Bewertungsgrundlage auf eine Einschätzung verzichtet werden“ (Bernhard Bührlen-Armstrong 1998:117f.).

Für die vorliegende Arbeit können aus den einzelnen Bereichen folgende Kriterien übernommen werden:

1. Nützlichkeit
 - 1.1. Identifikation der beteiligten Gruppen
 - 1.5. Klarheit des Berichts
2. Anwendbarkeit
 - 2.1. Praktische Verfahren
4. Genauigkeit
 - 4.1. Identifikation des Gegenstands
 - 4.2. Kontextanalyse
 - 4.3. Beschreibung von Zielen und Vorgehen
 - 4.5. Valide Messung
 - 4.6. Reliable Messung
 - 4.10. Gerechtfertigte Folgerungen

Dabei decken sich einige Kriterien mit den bereits zur Verwendung bestimmten Kriterien der Deutschen Gesellschaft für Evaluation (2001), so dass letztlich nur die Kriterien „Klarheit des Berichts⁵⁷“ und „Praktische Verfahren“ als Ergänzung zu den bestehenden Kriterien übernommen werden.

Wie schon erwähnt, legt auch Bernhard Bührlen-Armstrong (1998) in seiner Meta-Evaluation zur personalkommunikativen AIDS-Prävention verschiedene Bewertungskriterien zu Grunde, die sich aus seinem theoretischen Kontext ableiten. Diese lauten⁵⁸:

- 1 Kriterien zur Überprüfbarkeit
 - 1.1 Begründung der Befunde in den Daten
 - 1.2 Logik der Schlussfolgerungen
- 2 Kriterien zur Zuverlässigkeit
 - 2.1 Angemessenheit der Forschungsentscheidungen
 - 2.2 Vermeidung von Versuchsleitereffekten

⁵⁷ Dieses Kriterium wird, da nicht nur Evaluationsberichte, sondern auch andere Publikationen untersucht werden, umbenannt in „Klarheit der Veröffentlichung“.

⁵⁸ Bernhard Bührlen-Armstrong selbst beschreibt die einzelnen Kriterien jedoch detaillierter als es mir an dieser Stelle möglich ist.

3 Kriterien zur Glaubwürdigkeit

3.1 Einsatz von Methoden zur Erhöhung der Glaubwürdigkeit

3.2 Einfluss dieser Methoden auf methodische Entscheidungen, Datenquellen und Befunde

4 Kriterien zur Übertragbarkeit

4.1 Bezug zum Programm

4.2 Nützlichkeit für das Programm (vgl. a.a.O.:125ff)

Einige der verwandten Kriterien können nur dann beurteilt werden, wenn sämtliche Unterlagen der Einzelstudien vorliegen, daher sind sie für diese Untersuchung nicht brauchbar. Sinnvoll scheinen mir jedoch die Bereiche der Zuverlässigkeit⁵⁹ und der Übertragbarkeit zu sein, weswegen die Kriterien „Angemessenheit der Forschungsentscheidungen“ und „Übertragbarkeit“ ebenfalls übernommen werden sollen.

Letztlich soll an dieser Stelle eine fünfte Kategorie eingefügt werden, die nicht aus bestehenden Standards übernommen wurde, nämlich der Bezug zu den Forschungsfragen dieser Meta-Evaluation. Diese umfasst das Kriterium der Lernerfolgsmessung.

In der vorliegenden Untersuchung soll zusammengefasst also von folgenden Kriterien ausgegangen werden:

1. Nützlichkeit

1.1. Identifizierung der Beteiligten und Betroffenen

1.2. Klärung der Evaluationszwecke

1.3. Klarheit der Veröffentlichung

2. Durchführbarkeit

2.1. Angemessene Verfahren

2.2. Praktische Verfahren

3. Genauigkeit

3.1. Beschreibung des Evaluationsgegenstandes

3.2. Kontextanalyse

3.3. Beschreibung von Zwecken und Vorgehen

3.4. Valide und reliable Informationen

3.5. Begründete Schlussfolgerungen

3.6. Übertragbarkeit

4. Zuverlässigkeit

4.1. Angemessenheit der Forschungsentscheidungen

5. Bezug zu Forschungsfragen

4.1. Lernerfolgsmessung

4.2. Signifikante Ergebnisse

⁵⁹ Durch ein Review von Literatur können zu Versuchsleitereffekten keine Aussagen gemacht werden, da diese sehr wahrscheinlich nicht publiziert wurden.

Durch die Untersuchung der Primärstudien nach diesen Kriterien wird überprüft, ob Evaluationsstandards eingehalten worden sind. Allerdings ist ersichtlich, dass einige dieser Kriterien nicht primär auf die Forschungsfrage abzielen und eine detaillierte Darstellung den Rahmen dieser Arbeit sprengen würde. Aus diesem Grund fließen letztlich nur die Kriterien in die Untersuchung ein, die relevante Befunde zur Forschungsfrage beisteuern können. Diese betreffen die Durchführbarkeit, die Genauigkeit, die Zuverlässigkeit sowie die Bezüge zu den Forschungsfragen. Auf eine ausführliche Bewertung der Primärstudien nach allen Kriterien soll jedoch nicht vollständig verzichtet werden – diese ist daher im Anhang zu finden.

Der meta-evaluativen Untersuchung werden zusammengefasst folgende Kriterien zu Grunde gelegt:

- 1. Durchführbarkeit**
 - 1.1. Angemessene Verfahren
 - 1.2. Praktische Verfahren
- 2. Genauigkeit**
 - 2.1. Valide und reliable Informationen
 - 2.2. Begründete Schlussfolgerungen
- 3. Zuverlässigkeit**
 - 3.1. Angemessenheit der Forschungsentscheidungen
- 4. Bezug zu Forschungsfragen**
 - 4.1. Lernerfolgsmessung
 - 4.2. Signifikante Ergebnisse

8.2.4 Erläuterung der einzelnen Kriterien

Was mit den aufgestellten Kriterien im Einzelnen gemeint ist, soll im Folgenden erläutert werden. Dies dient insbesondere einer präziseren Beschreibung der einzelnen Kriterien, was wiederum zu einer höheren Validität und Objektivität der Bewertung führt.

8.2.4.1 Kriterien der Durchführbarkeit

Mit den Kriterien zur Durchführbarkeit soll untersucht werden, ob eine Evaluation „realistisch, gut durchdacht, diplomatisch [...] geplant und ausgeführt wird“ (Deutsche Gesellschaft für Evaluation 2001).

Angemessene Verfahren

„Evaluationsverfahren, einschließlich der Verfahren zur Beschaffung notwendiger Informationen, sollen so gewählt werden, dass Belastungen des Evaluationsgegenstandes bzw. der Beteiligten und

Betroffenen in einem angemessenen Verhältnis zum erwarteten Nutzen der Evaluation stehen“ (ebd.).

Praktische Verfahren

„Die Evaluationsverfahren sollten praktisch sein, so dass Störungen minimal gehalten und die benötigten Informationen beschafft werden können“ (Thomas Widmer 1996:20).

8.2.4.2 Genauigkeit

„Die Genauigkeitsstandards sollen sicher stellen, dass eine Evaluation gültige Informationen und Ergebnisse zu dem jeweiligen Evaluationsgegenstand und den Evaluationsfragestellungen hervor bringt und vermittelt“ (Deutsche Gesellschaft für Evaluation 2001).

Valide und reliable Informationen

„Die Verfahren zur Gewinnung von Daten sollen so gewählt oder entwickelt und dann eingesetzt werden, dass die Zuverlässigkeit der gewonnenen Daten und ihre Gültigkeit bezogen auf die Beantwortung der Evaluationsfragestellungen nach fachlichen Maßstäben sichergestellt sind. Die fachlichen Maßstäbe sollen sich an den Gütekriterien quantitativer und qualitativer Sozialforschung orientieren“ (ebd.).

Begründete Schlussfolgerungen

„Die in einer Evaluation gezogenen Folgerungen sollen ausdrücklich begründet werden, damit die Adressatinnen und Adressaten diese einschätzen können“ (ebd.).

8.2.4.3 Zuverlässigkeit

„Das Vorgehen bei Datenerhebung und –auswertung und Interpretation der Befunde entspricht der allgemein akzeptierten Praxis“ (Bernhard Bührlen-Armstrong 1998:115).

Angemessenheit der Forschungsentscheidungen

Bei der Untersuchung muss der Zusammenhang zwischen Untersuchungshypothesen, Stichprobenziehung und Auswahl der Auswertungsmethode deutlich werden. Methodische Veränderungen müssen begründet werden (vgl. a.a.O.:126).

8.2.4.4 Bezug zu den Forschungsfragen

An dieser Stelle wird ein ergänzendes Kriterium eingefügt, das sich sehr speziell auf die Forschungsfrage der meta-evaluativen Untersuchung bezieht und daher nicht aus bestehenden Standards entnommen werden konnte:

Lernerfolgsmessung

Es muss in der Primärstudie deutlich werden, wie Lernerfolg definiert wird.

Signifikante Ergebnisse

Die Unterschiede von Test- und Kontrollgruppen sollten signifikant sein.

8.2.5 Operationalisierung der Kriterien

Die angelegten Kriterien können als „erfüllt“, „nahezu erfüllt“, „teilweise erfüllt“, „kaum erfüllt“ oder „nicht erfüllt“ bewertet werden. Wird in einer Primärstudie beispielsweise die Gruppenbildung nicht hinreichend beschrieben, kann das Kriterium „praktische Verfahren“ als nicht erfüllt betrachtet werden.

Die Kriterien dienen also in erster Linie der Bewertung der Evaluationsverfahren der einzelnen Primärstudien. Um die Kriterien auf die Forschungsfragen beziehen zu können, wird diese bei der Operationalisierung stark in den Blick genommen. Andernfalls bestünde die Gefahr, als Ergebnis der Meta-Evaluation beispielsweise lediglich festhalten zu können, dass alle Primärevaluationen methodischen Standards entsprachen. Dies ist jedoch weder Sinn noch Zweck der Untersuchung. Aus diesem Grund werden die Indikatoren, die dazu dienen sollen, ein Kriterium messbar zu machen, aus den theoretischen Überlegungen im vorderen Teil der Arbeit abgeleitet und beziehen sich auf die inhaltlichen Fragestellungen. Zusammengefasst bedeutet dies, dass generelle Kriterien für die Bewertung der Evaluationen zu Grunde gelegt werden. Die Aussagen darüber, ob ein Kriterium als „erfüllt“ oder „nicht erfüllt“ bewertet wird, ist jedoch von der Forschungsfrage abhängig.

8.2.5.1 Durchführbarkeit

Angemessene Verfahren

- Es wird angegeben, in welchem Rahmen die Evaluation stattfand.
- Belastungen des Evaluationsgegenstandes bzw. der Beteiligten und Betroffenen stehen in einem angemessenen Verhältnis zum erwarteten Nutzen der Evaluation.
- Gender-Aspekte sowie kulturelle und soziale Differenz wurden berücksichtigt.

Praktische Verfahren

- Die Messungen wurden mit Test- und Kontrollgruppen durchgeführt.
- Die Kriterien der Gruppenbildung werden erläutert.
- Die Kriterien der Gruppenbildung sind nachvollziehbar.

8.2.5.2 Genauigkeit

Valide und reliable Informationen

- Die Verfahren der Datengewinnung und -analyse werden präzise beschrieben.
- Die Verfahren der Datengewinnung und -analyse sind nachvollziehbar.
- Die Evaluation orientierte sich an Gütekriterien.
- Reliabilität und Validität der Daten werden angegeben.

Begründete Schlussfolgerungen

- Die Ergebnisse der Evaluation werden beschrieben.
- Es werden Angaben darüber gemacht, welche Daten diese Schlussfolgerungen zulassen.
- Es wird eigene Kritik an der Evaluation angegeben.

8.2.5.3 Zuverlässigkeit

Angemessenheit der Forschungsentscheidungen

- Im Vorfeld wurden Hypothesen aufgestellt.
- Die Wahl der Stichprobe wird begründet.
- Die Wahl der Auswertungsmethode wird begründet.
- Wenn bei dem Vergleich der Gruppen die unabhängigen Variablen konstant gehalten werden, bezieht sich auch die Interpretation der Ergebnisse nicht auf diese Variablen.

8.2.5.4 Bezug zu den Forschungsfragen

Lernerfolgsmessung

- Es wird deutlich, wie der Begriff Lernerfolg operationalisiert wird.
- Es wurden vor der Untersuchung Lernziele formuliert.
- Es wird deutlich, welche Lernziele getestet wurden.
- Das Konstrukt Lernerfolg wird in Dimensionen aufgegliedert.
- Es werden Lernerfolg, Lernprozess und Lernleistung unterschieden.
- Es werden nur solche Gruppen miteinander verglichen, deren Lernziele sich in die Taxonomie nach Lorin W. Anderson und David R. Krathwohl (2001) einordnen lassen.
- Beim Vergleich von Untersuchungs- und Kontrollgruppe werden nur die Faktoren miteinander verglichen, die dieselbe Ausprägung von Lernerfolg zu Grunde legen (Taxonomie nach Lorin W. Anderson und David R. Krathwohl 2001).

Signifikante Ergebnisse

- Unterschiede zwischen Test- und Kontrollgruppe werden auf Signifikanz getestet.

8.2.6 Codierung der Primärstudien

Die Primärstudien werden zunächst einzeln beschrieben. Die Beschreibung umfasst Fragestellungen und Zielsetzungen der Evaluationsstudie, Voraussetzungen, Methodik, Datengrundlage, Ergebnisse, die Definition von Lernerfolg, die Einordnung in die Taxonomie nach Lorin W. Anderson und David R. Krathwohl (2001) sowie eine zusammenfassende Bewertung.

Bei der Zuordnung der einzelnen Kriterien werden Verfahren der inhaltlichen Strukturierung einer deduktiven Inhaltsanalyse (vgl. Philipp Mayring 1993, vgl. Thomas Widmer 1996:64) angewendet.

Die Basis dafür liefern die Codierungen, die mit Hilfe der Software Atlas.ti⁶⁰ erstellt worden sind. Jede Primärstudie wird dabei in eine Textdatei transferiert und in das Programm eingelesen. Einzelne Textpassagen werden entsprechend des genannten Rasters codiert, so dass auf diese Weise die für die Meta-Evaluation relevanten Merkmale der Primärstudien abgebildet und zusammengefasst werden können. Mit Hilfe dieser Beschreibungen können anschließend Aussagen zu den bereits formulierten Bewertungskriterien getroffen werden.

⁶⁰ Bei der Software handelt es sich um ein Programm zur qualitativen Datenanalyse.

Es können nicht in jeder Einzelstudie Angaben zu allen genannten Indikatoren gemacht werden. Dieses Problem greift auch Bernhard Bührlen-Armstrong (1998:131) auf. Die Vergleichbarkeit soll vielmehr auf der Ebene der Bewertungskriterien erfolgen. Dabei wird jede Studie im Hinblick auf jedes Kriterium eingestuft bzw. bewertet. Der Grad der Erfüllung wird auf einer fünfstufigen Skala eingeschätzt, d.h. für jede Primärstudie wird jedem Kriterium ein Wert zugeordnet (vgl. ebd., vgl. Thomas Widmer 1996:242). Dabei sind unterschiedliche Verfahren möglich:

Bernhard Bührlen-Armstrong leitet die Bewertung der Primärstudien aus dem arithmetischen Mittel der Kriterienwerte ab (vgl. ebd.). Dabei gibt er jedoch selbst zu bedenken, dass Verzerrungen möglich seien,

„da die Höhe des Beitrags eines Kriteriums zum Gesamtscore des Bewertungsbereichs von Studie zu Studie variieren kann. Es muss von der Annahme ausgegangen werden, dass die Kriterien für alle Studien das gleiche Gewicht besitzen und die Abstände zwischen je zwei benachbarten Werten konstant sind“ (ebd.).

Ich halte dieses Verfahren nur für bedingt zulässig, da es von einer linearen Skala der Bewertung der Einzelkriterien in den Primärstudien ausgeht. Eine Summierung ist jedoch nur bei Intervall- oder Ratioskalen zulässig, nicht jedoch bei einer Ordinalskala (vgl. Peter Baumgartner & Annabell Preussler 2005:21).

In seiner Meta-Evaluation bedient sich Thomas Widmer (1996:242) dementsprechend einer Symbolik der Bewertungen und stellt die Ergebnisse qualitativ dar:

„Einerseits ist darauf hinzuweisen, dass die [...] angegebenen Bewertungen ein ordinales Messniveau aufweisen und dass es aus diesem Grund nicht zulässig ist, die Werte der einzelnen Kategorien zu einem Kriterium in irgendeiner Form zusammenzufassen. Ebenso wenig zweckmässig ist die Zusammenfassung der Bewertungen zu einer Studie über die Einzelkriterien hinweg, da dafür das relative Gewicht der einzelnen Kriterien zu bestimmen wäre“ (a.a.O.:244).

In der vorliegenden Arbeit möchte ich ein ähnliches Verfahren zu Grunde legen: Die Qualitative Gewichtung und Summierung (QGS). Diese Bewertungsmethode wurde von Michael Scriven (1991) entwickelt, um einerseits differenzierte Bewertungen zu ermöglichen, andererseits jedoch die Verwendung als nicht zulässige Skala auszuschließen. Dieses Verfahren hat sich bereits bei Projektevaluationen in hohem Maße bewährt (vgl. Peter Baumgartner & Stefan Frank 2000).

Anders als in herkömmlichen Verfahren werden bei der QGS-Methode nicht-ziffernartige Symbole verwendet. Für die Begutachungskriterien werden in diesem Fall fünf Gewichte vergeben: positiv (*), eher positiv (#), neutral (+), eher negativ (0) und negativ (-). Diese Symbolik entspricht beispielsweise der Verwendung bei der Bewertung von Einreichungen im Rahmen des Medidaprix⁶¹. Die Ergebnisse der Leistungsbewertung werden abschließend integriert,

⁶¹ <http://www.medidaprix.org>, letzter Aufruf: 21.06.2007

„indem jede Kategorie mit der gleichen Wertigkeit summiert wird“, d.h. man erhält mehrere Gesamtwerte für jeden einzelnen Evaluanden (Peter Baumgartner & Annabell Preussler 2005:21f.).

Das Raster zur Bewertung der Kriterien ist in Tabelle 4 dargestellt.

Wert	Bezeichnung	Erläuterung
*	positiv	Das Kriterium ist erfüllt
#	eher positiv	Das Kriterium ist nahezu erfüllt
+	neutral	Das Kriterium ist teilweise erfüllt
0	eher negativ	Das Kriterium ist kaum erfüllt
-	negativ	Das Kriterium wurde nicht erfüllt oder nicht beachtet

Tabelle 4: Raster zur Bewertung der Kriterien in den Primärstudien in Anlehnung an Thomas Widmer (1996)

Wurde ein Kriterium in idealer bzw. überdurchschnittlicher Weise erfüllt, wird eine positive Bewertung vergeben. Eher positiv wird gewertet, wenn ein Kriterium mit leichten Abstrichen, aber nahezu erfüllt wird. Sind positive und negative Aspekte gleichermaßen zu finden, wird das Kriterium als neutral eingestuft. Als eher negativ wird gewertet, wenn eine Studie zwar positive Aspekte hinsichtlich des Kriteriums aufweist, negative Aspekte jedoch überwiegen. Letztlich wird eine negative Bewertung vergeben, wenn das Kriterium nicht erfüllt oder nicht beachtet wurde (vgl. Thomas Widmer 1996:242, vgl. Bernhard Bührlen-Armstrong 1998:131).

Fehlen ausreichende Informationen, um das Kriterium überhaupt bewerten zu können, wird keine Wertung vergeben.

Das Verfahren soll in der vorliegenden Arbeit dazu dienen, eine Vergleichbarkeit der Primärstudien hinsichtlich der relevanten Kriterien zu ermöglichen. Es soll jedoch kein Ranking stattfinden, es geht am Ende nicht darum, eine „Siegerin“ zu küren.

8.3 Bewertung der Primärstudien

Im folgenden Kapitel möchte ich die Bewertung der einzelnen Primärstudien nach den aufgestellten Kriterien vornehmen. Zusammenfassende Beschreibungen der Studien sind im Anhang zu finden.

8.3.1 P 1: Edward R. Jones et al. (1999)

Die vorliegende Evaluation hatte zum Ziel, Vor- und Nachteile einer Online-Klasse gegenüber einer traditionellen Face-to-Face-Klasse aufzuzeigen. TeilnehmerInnen waren 89 Studierende in zwei Gruppen eines Statistikkurses. Dabei wurde die eine Gruppe in traditioneller Art und Weise unterrichtet, die andere über das Internet mit virtuellem Klassenzimmer.

Im Folgenden soll auf die einzelnen Kriterien detailliert eingegangen werden.

8.3.1.1 Durchführbarkeit

Angemessene Verfahren

Der Rahmen der Untersuchung wird erläutert – das Verhältnis des Aufwandes zum Nutzen der Evaluation schien angemessen zu sein. In der Publikation wird auf Differenzen innerhalb der Studierendenschaft hingewiesen, was bei der Bildung von Test- und Kontrollgruppen deutlich wurde.

Das Kriterium wurde überwiegend erfüllt und wird als eher positiv gewertet.

Praktische Verfahren

In der vorliegenden Studie bestand die Möglichkeit der freien Wahl bezüglich der Gruppenzuordnung seitens der Studierenden. Dieses Verfahren war laut der Angabe der Autoren beabsichtigt, um demographische Unterschiede zwischen Studierenden, die virtuelle Lehre traditionellem Unterricht vorzogen, ausmachen zu können. Darüber hinaus empfanden es die Autoren als unfair, Studierende in eine virtuelle Klasse zu zwingen, da eine gewisse Übung in der Handhabung – zum Beispiel eines Internetbrowsers – gewährleistet sein musste. Allerdings zeigten sich wichtige Unterschiede in den Gruppen: Diejenigen, die sich in die virtuelle Klasse einschrieben, waren durchschnittlich signifikant älter als diejenigen in der traditionellen Klasse und wiesen auch signifikant höhere Testergebnisse auf. Insofern werden die Kriterien zu Gruppenbildung zwar erläutert, sind aber nur bedingt nachvollziehbar, da ein Vergleich nicht uneingeschränkt möglich ist.

Die Untersuchung lief über 15 Wochen an zwei Kohorten von Studierenden in einem virtuellen bzw. traditionellen Statistikkurs. Dabei

bezeichnet der virtuelle Kurs die Experimentalgruppe, der traditionelle Kurs die Vergleichsgruppe.

Die Gruppenbildung wird erläutert, ist aus methodischen Gründen jedoch nicht nachvollziehbar. Das Kriterium wurde daher nur teilweise erfüllt und wird als neutral gewertet.

8.3.1.2 Genauigkeit

Valide und reliable Informationen

Im Vorfeld der Untersuchung wurde ein Algebratest durchgeführt. Die Ergebnisse dieses Tests dienten nicht nur der Zulassung zum Kurs, sondern auch als Pretest, der anschließend zum Vergleich herangezogen wurde.

Die Untersuchung basierte auf folgenden Daten:

1. „Student demographics - age, GPA⁶², completed hours, and pretest scores
2. Scores on the three, in-class, multiple-choice in-class examinations
3. Student withdrawal and passing rates“ (Edward R Jones et al. 1999:3).

Diese Daten wurden bei der Anmeldung zum Kurs sowie nach dessen Abschluss erhoben. Weitere Datenerhebungen während des Semesters fanden in Form von Multiple-Choice-Tests statt.

Die demographischen Angaben aus Experimental- und Vergleichsgruppe wurden mit Hilfe des Wilcoxon-Rangsummentests⁶³ miteinander verglichen. Anschließend wurden auch die Examensnoten auf der Grundlage von Kovarianzanalysen mit Messwiederholung einem Vergleich unterzogen. Diese Analyse wurde zum einen mit und zum Anderen ohne die Kovarianten (GPA und Algebravortest) durchgeführt. Mit dem „exakten Test auf Unabhängigkeit“ nach Fisher⁶⁴ wurden die Abbruchquoten analysiert.

Allerdings werden keine Angaben zu Gütekriterien gemacht. Aus diesem Grund kann das Kriterium nur mit neutral gewertet werden, da es nicht in vollem Umfang erfüllt wurde.

Begründete Schlussfolgerungen

In der Untersuchung wurden die Schlussfolgerungen auf Grund der Datenlage gezogen und es wurden keine Interpretationen geliefert, die nicht aus den Daten resultieren: Beide Kurse hatten die selbe Laufzeit und

⁶² Als GPA (Grade Point Average) wird im US-amerikanischen Bildungssystem eine Art Notendurchschnitt bezeichnet. Dabei wird jeder einzelnen Note ein Punktwert zugeordnet. Aus diesen wird anschließend ein gewichtetes arithmetisches Mittel berechnet (vgl. http://de.wikipedia.org/wiki/Grade_Point_Average, letzter Aufruf: 25.08.2006).

⁶³ Nichtparametrischer Test, um Mediane zu vergleichen.

⁶⁴ Hierbei handelt es sich um einen Test zur Signifikanzprüfung.

wurden von derselben Lehrperson durchgeführt. Beide deckten den gleichen Inhalt ab, nutzten das gleiche Kursmaterial und Hausaufgaben. Noten wurden auf Grund derselben Multiple-Choice-Tests vergeben. Auch die Prüfungen wurden laut Jones et al. unter gleichen Konditionen durchgeführt (vgl. ebd.). Die Untersuchungsbedingungen und unabhängigen Variablen wurden also konstant gehalten. Es ist jedoch als positiv hervorzuheben, dass kaum Interpretationen der Ergebnisse hinsichtlich dieser Variablen erfolgten. Allerdings geben die Autoren an, dass sie virtuelle Lehre weiterhin fördern wollten. Dies darf m.E. allerdings nicht als Schluss aus der Evaluation gezogen werden. Da es kein spezifisches Kursmaterial gab, sind die Möglichkeiten der neuen Medien nicht ausgeschöpft worden.

Die wichtigsten Ergebnisse werden übersichtlich dargestellt. Darüber hinaus äußern die Autoren auch Kritik an ihrer Untersuchung, indem sie feststellen, dass die Umgebungsvariablen nicht mit der Vorgängerstudie von Schutte vergleichbar gewesen seien:

„It could be that the students in Schutte's classes were more motivated since his course was probably not a service course. Likewise, his students may have been younger and less involved raising families and working. In this study, the average age of the students completing the web-based class was 30.7. On average they were enrolled in 11.8 semester hours and were employed 22.3 hours per week. It can also be speculated that Schutte's teaching style may be more suited to a web-based class than a traditional class“ (a.a.O.:96).

Als wichtigstes Ergebnis kann festgehalten werden, dass die Ergebnisse der Vorgängerstudie (Schutte 1997) nicht reproduziert werden konnten. Mögliche Gründe hierfür werden in der Publikation genannt. Das Kriterium wurde überwiegend erfüllt und kann als eher positiv gewertet werden.

8.3.1.3 Zuverlässigkeit

Angemessenheit der Forschungsentscheidungen

In der Untersuchung sollte unter anderem folgende Hypothese untersucht werden: Der Hintergrund der Studierenden, die einen Webkurs vorziehen, weise keine Unterschiede zu den Studierenden auf, die einen traditionellen Kurs bevorzugten. Die zur Untersuchung dieser Hypothese verwendeten Methoden werden erläutert. Allerdings scheint die Gruppenzusammensetzung nicht sinnvoll zu sein, da eine zu große Heterogenität vorliegt.

In Bezug auf die Erfassung des Lernerfolgs ist eine bloße Orientierung an Noten und Abschlussraten als nicht hinreichend zu bewerten, da damit nur die Performanceleistung, nicht aber der Lernprozess und noch weniger die Kompetenz bewertet werden kann.

Es wurde einerseits versucht, Rahmenbedingungen konstant zu halten. So wurden beide Kurse von derselben Lehrperson über die gleiche Zeit mit gleichem Inhalt durchgeführt. Die Autoren sagen zudem aus, dass

auch das gleiche Kursmaterial genutzt worden sei – es sei einfach in webfähige Form „übersetzt“ worden. Damit mag zwar eine größere Vergleichbarkeit der Gruppen gegeben sein, es können jedoch keine Aussagen zu Spezifika neuer Medien gemacht werden, da diese nicht ausgeschöpft worden sind.

Andererseits schien es nicht möglich zu sein, alle Fehlerquellen durch die Anlage des Untersuchungsdesigns gering zu halten. Dies zeigte sich gravierend in der Zusammensetzung der Experimental- und der Kontrollgruppe. Zwar geben die Autoren an, das praktizierte Verfahren aus Gründen der Fairness bewusst angewendet zu haben, dennoch ergaben sich hier Verzerrungen, wie ein höherer Altersdurchschnitt und eine größere Vertrautheit mit der Technik in der Experimentalgruppe. Diese Effekte werden in der Publikation jedoch als Ergebnisse und nicht als Voraussetzungen interpretiert. Insofern halte ich die Vergleichbarkeit von Experimental- und Kontrollgruppe für nicht sehr hoch und das Evaluationsdesign somit für problematisch.

Das Kriterium wurde mit Abstrichen erfüllt und ist als eher negativ zu werten.

8.3.1.4 Bezug zu den Forschungsfragen

Lernerfolgsmessung

Die unabhängige Variable Lernerfolg wurde in der vorliegenden Untersuchung ausschließlich als Performanceleistung an Hand einfach zu messender Werte definiert und zwar unter Bezugnahme der abhängigen Variablen „Note des Vortests“, „Noten von drei Multiple-Choice-Tests“ sowie „Abbruchs- bzw. Bestehensquoten“. Zusätzlich wurden demographische Angaben wie Alter, GPA und absolvierte Kursstunden erhoben. Einen höheren Lernerfolg erzielten dieser Definition nach also Studierende, die bessere Noten in den jeweiligen Tests erreichten. Vereinfacht gesagt: Der Lernerfolg wurde hier ausschließlich an Hand der Menge der richtig angekreuzten Felder in den Multiple-Choice-Tests definiert. Es wurde also keine Kompetenzmessung betrieben, auch der Lernprozess selbst blieb im Hintergrund⁶⁵ – es zählte einzig der Output an Hand von Testnoten.

Es werden keine expliziten Lernziele im Bericht erwähnt. Festzuhalten bleibt jedoch, dass die Studie darauf abzielte, Testergebnisse von Multiple-Choice-Tests abzufragen. Dies war der von Beginn an formulierte Anspruch, der auch wie beschrieben getestet wurde. Insofern ergeben sich keine Widersprüche zwischen formulierten und getesteten Lernzielen.

Auswirkungen von Lerneffekten zeigten sich durch den Vergleich des Pretests mit den Abschlussnoten und dem GPA mit dem Ergebnis,

⁶⁵ Die Multiple-Choice-Tests wurden zu drei Messzeitpunkten erhoben, so dass in gewisser Weise ein Prozess sichtbar wird. Diesen würde ich allerdings nicht als Lernprozess bezeichnen, denn was tatsächlich gelernt wurde, entzieht sich meiner Kenntnis.

dass die Studierenden in der virtuellen Klasse zum einen älter und zum anderen technikvertrauter waren als die Studierenden in der traditionellen Klasse. Dies fand jedoch in der Auswertung so gut wie keine Berücksichtigung.

Auch auf der Taxonomie der Lernziele scheinen beide Kurse auf der gleichen Ebene zu liegen, allerdings kann keine eindeutige Zuordnung vorgenommen werden, da die Kursinhalte und Arbeitsweisen an sich zu wenig bekannt sind. Die Erhebung der Daten bzw. Tests wurde in beiden Gruppen zum gleichen Messzeitpunkt mit demselben Messinstrument vorgenommen: In beiden Untersuchungsgruppen wurde der gleiche Multiple-Choice-Test durchgeführt, der darüber hinaus auf dem gleichen Kursmaterial basiert, so dass spezielle Eigenschaften neuer Medien nicht genutzt worden sind. Dies legt den Schluss nahe, dass in beiden Fällen nur Wissen auf der Stufe des Erinnerns abgefragt wurde⁶⁶.

DIE KOGNITIVE PROZESS-DIMENSION					
1. Erinnern	2. Verstehen	3. Anwenden	4. Analysieren	5. Bewerten	6. (Er)schaffen
Tests beider Gruppen					

Tabelle 5: Einordnung von P 1 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)

Allerdings werden keine Lernziele benannt, so dass nicht klar ist ob die Lernerfolgskontrolle tatsächlich an Lernzielen ausgerichtet war.

Das Kriterium wird kaum erfüllt und damit als eher negativ gewertet.

Signifikante Ergebnisse

Die erzielten Testergebnisse fielen in der virtuellen Klasse signifikant besser aus als in der traditionellen. Es zeigte sich aber, dass sich die Werte im Verlauf des Semesters einander näherten. Den Unterschied erklären sich die Verfasser durch den höheren Durchschnittswert der virtuellen Klasse im Vortest. Wurden diese Werte berücksichtigt, so betrug der Unterschied zwischen den Gruppen nur noch 3% und war nicht mehr signifikant. Wurde zusätzlich noch der GPA berücksichtigt, zeigte sich am Ende des Kurses sogar eine Überlegenheit der traditionellen Klasse.

Darüber hinaus wurden auch die Abbrecherquoten in der Untersuchung einkalkuliert. Diese waren zwar nicht signifikant, wiesen aber höhere Durchschnittswerte in der virtuellen Klasse auf.

Das Kriterium wurde voll erfüllt und ist als positiv zu werten.

⁶⁶ Allerdings kann diese Einordnung nur vermutet werden, da das Kursmaterial nicht spezifisch erläutert wird.

8.3.1.5 Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Durchführbarkeit						
1.1	Angemessene Verfahren		x			
1.2	Praktische Verfahren			x		
2. Genauigkeit						
2.1	Valide und reliable Informationen			x		
2.2	Begründete Schlussfolgerungen		x			
3. Zuverlässigkeit						
3.1	Angemessenheit der Forschungsentscheidungen				x	
4. Bezug zu Forschungsfragen						
4.1	Lernerfolgsmessung				x	
4.2	Signifikante Ergebnisse	x				
	Gesamtbewertung	1	2	2	2	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 6: Bewertungsraster P1

8.3.2 P 2: Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr. (2001)

In der vorliegenden Studie sollte die Zufriedenheit der Studierenden mit der akademischen Qualität von Online- und Face-to-Face-Kursen untersucht werden. Darüber hinaus sollte diese auf Zusammenhänge mit der Anzahl von Online-Kursen, in die die Studierenden eingeschrieben waren, überprüft werden.

8.3.2.1 Durchführbarkeit

Angemessene Verfahren

Der Rahmen der Evaluation wird beschrieben. Weiterhin schien das Verhältnis des Aufwandes zum Nutzen der Evaluation angemessen zu sein. Auf kulturelle oder soziale Differenzen sowie auf Gender-Aspekte wird nicht hingewiesen. Das Kriterium wurde überwiegend erfüllt und wird als eher positiv gewertet.

Praktische Verfahren

Verglichen wurde die Zufriedenheit der Studierenden mit Online- und Face-to-Face-Kursen. Ziel war es, signifikante Unterschiede zwischen diesen beiden Gruppen identifizieren zu können.

Dabei wurde die Hypothese aufgestellt, dass sich keine Unterschiede zwischen den Gruppen ergeben:

“Ho: true mean difference is zero (i.e., $\mu = 0$)

Ha: true mean difference is not equal to zero (i.e., $\mu \neq 0$)”

(Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr. 2001:9).

Es wurden 1802 ehemalige Studierende befragt, als Rücklauf konnten 567 ausgefüllte Fragebögen verzeichnet werden. Dies entspricht einer Rate von etwa 31%. Davon bewerteten 330 ProbandInnen sowohl Präsenz- als auch Online-Kurse. Zusätzlich gab es 232 Bewertungen, die nur auf Präsenzkurse bezogen waren. 5 Fragebögen wurden nicht vollständig beantwortet. Die Gruppenbildung vollzog sich also über die Kursbelegung der Studierenden.

Zusätzlich wurden die ProbandInnen, die einen oder mehrere Online-Kurse absolviert haben, in drei Gruppen eingeteilt:

1. Studierende, die nur einen Onlinekurs belegt hatten,
2. Studierende, die zwei oder drei Onlinekurse belegt hatten,
3. Studierende, die vier oder mehr Onlinekurse belegt hatten.

Das Kriterium wurde insofern erfüllt, als es zwei Gruppen gab, die miteinander verglichen werden. Die Zusammensetzung dieser Gruppen wurde nachvollziehbar begründet.

Das Kriterium ist als positiv zu bewerten.

8.3.2.2 Genauigkeit

Valide und reliable Informationen

Untersuchungsgegenstand war die AbsolventInnen-Befragung aller Bachelor-AbsolventInnen des Jahres 1999 (n=1802), die mit Fragen zur Bewertung des Colleges im Hinblick auf bestimmte Qualitätskriterien⁶⁷ versehen war. Zusätzlich wurde die Zufriedenheit mit der akademischen Qualität von Online- und Face-to-Face-Kursen und der Qualität des Services erhoben. Der Fragebogen enthielt 14 Items zur Messung der Exzellenz oder Qualität des Studiengangs (academic program) und 9 Items zur Messung der Qualität von Service und Support.

Zur Bewertung durch die Studierenden wurde eine 5-stufige Likert-Skala benutzt, deren niedrigster Wert „sehr unzufrieden“ und deren höchster Wert „sehr zufrieden“ entsprach. So stellen höhere Werte gleichsam eine höhere Zufriedenheit dar. Ob signifikante Unterschiede zwischen Online- und Face-to-Face-Kursen vorlagen, wurde mittels des T-Tests auf Signifikanz berechnet.

Der Fragebogen wurde von 567 Studierenden beantwortet, was einer Rücklaufquote von 31% entsprach. Davon bezogen sich 330 Rückmeldungen sowohl auf Online- als auch auf Face-to-Face-Kurse, 232 Studierende bewerteten nur Face-to-Face-Kurse⁶⁸.

Um die Zufriedenheit der Studierenden zu bewerten, wurden statistische Methoden präzise angewendet. Die AutorInnen legten Wert darauf, die Berechnung der Mittelwerte auch von der Zusammensetzung der Gruppen abhängig zu machen:

“For each assessment item, the respondent's mean satisfaction ratings for online classes was calculated separately from their mean satisfaction ratings for their face-to-face classes. In order to evaluate the research objective as to whether alumni reported any significant difference between their online and face-to-face ratings, a matched t-test was calculated. For each item, the mean of the differences between respondent's ratings of online and face-to-face classes was calculated.

To calculate the t-statistics for matched samples, mean of the differences between online and face-to-face ratings is divided by the standard error. The p value of the test was determined.

A decision was then made for each item as to whether the difference between the online and face-to-face classes was statistically significant and thus whether alumni rated online and face-to-face classes differently on that quality and effectiveness indicator” (a.a.O.: 8 f.).

⁶⁷ Z.B. Inhalt des Curriculums, Qualität der Ausbildung, Entwicklung von kritischem Denken, Karriererelevanz, Tests und Bewertungen.

⁶⁸ 5 Studierende haben nicht auf die Qualitätsfragen geantwortet.

Die Verfahren der Datengewinnung und –analyse werden im vorliegenden Forschungsbereich sehr genau beschrieben. Auf Gütekriterien wird jedoch nicht eingegangen; ebenso wird die Entwicklung des Testinstruments nicht erläutert.

Das Kriterium wurde teilweise erfüllt und ist somit als neutral zu werten.

Begründete Schlussfolgerungen

Als positiv ist hervorzuheben, dass die Auswertung nach Anzahl der belegten Kurse differenziert wurde. Dies bedeutet, dass in der Darstellung der Ergebnisse danach unterschieden wurde, wie viele Kurse jeweils belegt worden waren. Darüber hinaus wurden die Unterschiede zwischen diesen Gruppen auf Signifikanz geprüft.

Die Schlussfolgerungen wurden inhaltlich begründet. Allerdings müssen diese Aussagen vor dem Hintergrund relativiert werden, dass die Funktionalität der Online-Kurse nicht näher untersucht wurde.

Die AutorInnen äußern selbst ebenfalls Kritik an ihrer Untersuchung, indem sie feststellen:

“Differences between the online and face-to-face students' ratings of face-to-face academic quality raised questions about the significance of context for self-reported data [...]. Other measurement issues related to whether online and face-to-face classes should be compared are discussed” (a.a.O.:3).

Das Kriterium wurde überwiegend erfüllt und ist als eher positiv zu werten.

8.3.2.3 Zuverlässigkeit

Angemessenheit der Forschungsentscheidungen

In der vorliegenden Untersuchung wurde die Hypothese aufgestellt, dass sich keine Unterschiede in der Zufriedenheit mit Präsenz- und Onlinekursen zwischen den Gruppen ergeben:

“Ho: true mean difference is zero (i.e., $\mu = 0$)
Ha: true mean difference is not equal to zero (i.e., $\mu \neq 0$)”
(Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr. 2001:9).

Die Wahl der Stichprobe und der Auswertungsmethode wird begründet. Als kritisch ist zu bewerten, dass die Befragung erst ein Jahr nach Absolvierung der Kurse durchgeführt wurde. Die Untersuchung beruhte auf Selbsteinschätzungen der Studierenden. Dies ist insofern als positiv hervorzuheben, als damit kein absoluter Lernerfolg gemessen wurde, was mit dem angewendeten Evaluationsdesign nicht in akzeptabler Form möglich gewesen wäre.

Die Zufriedenheit mit den Online-Kursen wurde mit dem gleichen Fragebogen abgefragt, der auch die Zufriedenheit mit den Präsenzkursen maß. Dadurch wurden die Funktionalitäten der neuen Medien nicht

explizit erfasst. Da die AutorInnen auf diesen Umstand jedoch hinweisen, ist das Kriterium letztlich als eher positiv zu bewerten.

8.3.2.4 Bezug zu den Forschungsfragen

Lernerfolgsmessung

Es wird durch die Zitierung der „Middle States Commission on Higher Education“ (2001) darauf hingewiesen, dass das so genannte „outcome assessment“ zwei Absichten verfolge – Verantwortlichkeit („accountability“) und Effektivität („effectiveness“). Dieses könne einerseits auf der Mikroebene (z.B. studentisches Lernen) gemessen werden, andererseits auf der Makroebene (z.B. institutionale Effektivität). Ein Werkzeug zur Untersuchung dieser beiden Ebenen stelle der Alumni Survey dar (vgl. a.a.O.). Weiterführende Angaben zur Definition von Lernerfolg oder zu Lernzielen werden nicht gemacht.

In beiden Untersuchungsgruppen wurde der gleiche Fragebogen eingesetzt. Es ist dabei allerdings nicht bekannt, ob spezielle Eigenschaften neuer Medien genutzt worden sind. Da die Untersuchung aber ohnehin auf Selbstauskünfte und nicht auf Lernziele ausgerichtet war, ist eine Einordnung in die Taxonomie nach Lorin W. Anderson und David R. Krathwohl (2001) nicht möglich.

Das Kriterium wurde nicht erfüllt und ist damit als negativ zu werten.

Signifikante Ergebnisse

Die Resultate wurden detailliert dokumentiert. Signifikante Ergebnisse konnten gefunden werden - so wurde beispielsweise festgestellt, dass die Zufriedenheit mit den Online-Kursen signifikant zunahm, je mehr Kurse belegt wurden.

Allgemein kann festgehalten werden, dass die Zufriedenheit mit Präsenzkursen grundsätzlich höher war als mit Online-Kursen. Bei 12 von 14 Indikatoren waren diese Unterschiede statistisch signifikant. Bei einem Vergleich der Mittelwerte wurde jedoch deutlich, dass die Unterschiede, obwohl signifikant, von nur geringer praktischer Bedeutung waren, da sie zu klein waren, als dass eine sinnvolle, unterscheidbare Interpretation möglich gewesen wäre. Lediglich hinsichtlich der „Interaktion von StudentIn zu StudentIn“ betrug die Abweichung -0.646, was bedeutet, dass hier der Präsenzkurs auch in Bezug auf die zu Grunde liegende Skala deutlich besser abschnitt. In Bezug auf die Kategorien „develop critical thinking“ und „rigor and scholarship“ wurden keine signifikanten Unterschiede gefunden.

Betrachtet man die Auflistung nach Anzahl der belegten Kurse, so wird deutlich, dass die Mittelwerte in Bezug auf die Zufriedenheit mit Online-Kursen stiegen, je mehr dieser Kurse belegt werden. Allerdings nahm dabei der Grad der Signifikanz ab: Unterschieden sich bei der Belegung von nur einem Kurs noch 12 von 14 Mittelwerten der

Indikatoren signifikant, so konnten nur noch 5 signifikante Unterschiede bei der Belegung von 3 oder mehr Kursen gefunden werden. Dabei ist ein Blick auf die einzelnen Indikatoren interessant: Beispielsweise bewerteten Studierende, die 3 oder mehr Online-Kurse besucht hatten, Präsenzkurse signifikant geringer als Studierende, die nur einen oder zwei Online-Kurse besucht hatten.

Das Kriterium wurde voll erfüllt und wird mit positiv bewertet.

8.3.2.5 Bewertungsraster

Die vergebenen Bewertungen werden in Tabelle 7 zusammenfassend dargestellt. Es wird deutlich, dass die Primärstudie im eher positiven bis neutralen Bereich angesiedelt werden kann.

Nr.	Kriterium	*	#	+	0	–
1. Durchführbarkeit						
1.1	Angemessene Verfahren		x			
1.2	Praktische Verfahren		x			
2. Genauigkeit						
2.1	Valide und reliable Informationen			x		
2.2	Begründete Schlussfolgerungen		x			
3. Zuverlässigkeit						
3.1	Angemessenheit der Forschungsentscheidungen		x			
4. Bezug zu Forschungsfragen						
4.1	Lernerfolgsmessung					x
4.2	Signifikante Ergebnisse	x				
	Gesamtbewertung	1	4	1	0	1
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 7: Bewertungsraster P2

8.3.3 P 3: Rob Martens, Stanley J. Portier & Martin Valcke (1995)

In der vorliegenden Studie wurde die Effizienz von Lernen in einer interaktiven Lernumgebung im Vergleich zu Lernen mit Printmaterialien und Face-to-Face Kursen untersucht. Evaluiert wurden Prototypen von interaktiven Lernumgebungen. Dabei dienten die Studierenden, die den Stoff durch eine Präsenzvorlesung vermittelt bekommen hatten, als Kontrollgruppe. Das Thema aller Kurse war „Inferenzstatistik“ im ersten Studienjahr im sozialwissenschaftlichen Studiengang „Psychology and Educational Sciences“.

Eine Schlüsselrolle in der Untersuchung spielten die sogenannten „embedded support devices (ESD)“, eine Art didaktische Zugabe zum Basiskursmaterial mit Lösungshinweisen, Glossaren, Selbsttests, Beispielen usw. Die ESDs waren sowohl in den online vermittelten als auch in den Printkursen vorhanden, allerdings entschied in beiden Fällen der Autor bzw. die Autorin der jeweiligen Kurse über ihre Verwendung.

Von der Lernumgebung existierten verschiedene Versionen: In der ersten Version waren alle Möglichkeiten der Interaktivität geblockt, so dass Studierende die Umgebung als eine Art elektronisches Lernbuch ohne weitere Funktionen verwenden konnten. In zwei weiteren Versionen wurden diese angeboten, die Umgebungen unterschieden sich hier nur in der Art und Weise der angebotenen ESD. Diese waren entweder mehr oder weniger sichtbar.

Das Printmaterial entsprach dem Standard von Lehrbriefen. Es handelte sich hierbei um didaktisch aufbereitete Kurse, die von den Studierenden eigenständig bearbeitet wurden. Zu Forschungszwecken wurden zwei verschiedene Versionen von Printmaterial eingesetzt – einmal mit und einmal ohne Hilfestellungen bzw. didaktischen Anweisungen. Als Kontrollgruppe diente der Untersuchung eine traditionelle Face-to-Face-Vorlesung der Inferenzstatistik.

8.3.3.1 Durchführbarkeit

Angemessene Verfahren

Der Rahmen der Untersuchung wurde beschrieben. Es schien ein ausgewogenes Verhältnis des Aufwandes zum Nutzen der Evaluation vorzuliegen. Auf kulturelle oder soziale Aspekte wird nicht eingegangen, jedoch wurden Genderaspekte, Motivation und Studiergewohnheiten abgefragt. Das Kriterium wurde erfüllt und wird als positiv gewertet.

Praktische Verfahren

Um verschiedene Prozesse und Variablen kontrollieren zu können, wurden im Vorfeld ein Leseverständnistest sowie ein Vorwissenstest durchgeführt, der gleichsam als Pretest anzusehen war.

Darüber hinaus wurde auch der Einfluss der Lernmaterialien und der Lernumgebung selbst auf das Lernen untersucht. Als individuelle Variablen wurden Vorwissen und Lesekompetenz in die Studie einbezogen. Darüber hinaus sollte der Gebrauch und der Einfluss der ESDs in Bezug auf das Vorwissen der Lernenden mittels Logfile-Analyse überprüft werden: Aus der Information, was genau benutzt wurde, sollte geschlossen werden, welche Hilfen jemand benötigte und welches Vorwissen vorhanden war.

Um Vergleiche von allen Versionarten und der Kontrollgruppe anstellen zu können, wurden also verschiedene Settings bzw. Experimentalgruppen miteinander verglichen:

1. Lernumgebung mit sichtbarem Hilfesystem (n=40)
2. Lernumgebung ohne sichtbares Hilfesystem⁶⁹ (n=20)
3. Lernumgebung mit sichtbarem Hilfesystem (nicht-interaktiv)⁷⁰ (n=20)
4. Printkurse mit Hilfesystem (n=40)
5. Printkurse ohne Hilfesystem (n=40)
6. Vorlesung (Kontrollgruppe) (n=342)

Die Zuweisung der ProbandInnen zu den einzelnen Lerngruppen erfolgte mittels randomisierter Verfahren. Testergebnisse von Studierenden mit höherem Ausbildungsniveau sowie Studierende, die den Kurs zum zweiten Mal besuchten, wurden von der Untersuchung ausgeschlossen. Die Kriterien der Gruppenbildung werden erläutert.

Die Evaluation hatte einen formativen Charakter. Das Kriterium wurde voll erfüllt und wird mit positiv bewertet.

8.3.3.2 Genauigkeit

Valide und reliable Informationen

Bei der Beschreibung der Instrumente wird zunächst auf die Einflussvariablen eingegangen und es wird beschrieben, wie diese isoliert werden können:

“Some variables have to be explicitly measured or evaluated by making use of the following instruments:

- a reading comprehension test;
- a prior-knowledge-state test used as a pre-test;
- a subject-oriented mastery test used as a post-test and to determine study outcome; a three-part questionnaire:
 - part A includes questions about motivation, age, study habits, educational level and gender;

⁶⁹ Die Arbeitshilfen sind in diesem Fall in den Content integriert und nicht als solche ausgewiesen.

⁷⁰ Die Arbeitshilfen sind explizit als solche ausgewiesen.

- part B asks students to judge the degree of accessibility of the learning materials studied;
- part C puts forward questions about the use and appreciation of each specific embedded support device.
- an attitude measuring scale in relation to learning with computers" (Rob Martens, Stanley J. Portier & Martin Valcke 1995:16).

Die einzelnen Schritte der Untersuchung werden ebenfalls ausführlich erläutert:

- "Administration of the reading comprehension test, prior knowledge state test and Ou-questionnaire part A
- Assignment of students to experimental or control groups. We controlled for differences in prior knowledge.
- Introduction session to learn how to use the ILE
- Taking the course: the alternative approaches
 - Students in the experimental conditions study during three sessions of two hours each with the specific learning environment. After each session they fill in the Ou-questionnaire, part C.
 - Students in the control condition follow the regular lecture-format. Three lectures were given, each lecture focusing on one of the selected chapters on inferential statistics.
- Administration of the subject-oriented mastery test and Ou-questionnaire part B Test administration was organised one week after the research sessions were finished.
- The final examination. At the end of the academic year, students have to take a statistics examination as a part of their final assessment of the first year. The results of this regular examinations are considered as long-term study outcome effects and can be compared partly with the scores of the subject-mastery test" (ebd.).

KursteilnehmerInnen in den Experimentalgruppen arbeiteten drei Lernabschnitte von jeweils zwei Stunden mit der spezifischen Lernumgebung. Nach jedem Lernabschnitt füllten sie den Fragebogen (Teil C) aus. Teilnehmende der Kontrollgruppen folgten dem regelmäßigen Vortragsformat. Drei Vorträge wurden gehalten, jeweils bezogen auf eins der gewählten Kapitel im Bereich der Inferenzstatistik.

In den weiteren Analysen sind die folgenden Techniken angewendet worden: Regressionsanalyse, ANOVA, MANOVA und Profilanalyse, was von den AutorInnen als geeignet beschrieben wurde, um Effekte zwischen unabhängigen Mehrebenenvariablen, Effekte der Maßnahme und abhängigen Variablen zu isolieren. Die Verfahren der Datengewinnung und -analyse werden beschrieben. Auf Gütekriterien und die Entwicklung des Testinstruments wird jedoch nicht eingegangen.

Das Kriterium wurde somit teilweise erfüllt und ist als neutral zu werten.

Begründete Schlussfolgerungen

Werden die erwähnten Hypothesen betrachtet, so kann in Bezug auf die hier relevanten Hypothesen festgehalten werden, dass Hypothese 1 akzeptiert werden konnte bzw. die vermuteten Zusammenhänge nicht widerlegt wurden. Hypothesen 3 wurde dagegen abgelehnt. Auch Hypothese 6 musste den Ergebnissen der Studie nach verworfen werden.

In den meisten getesteten Bereichen brachte die Studie keine signifikanten Ergebnisse zu Tage: Keines der Forschungssettings führte zu signifikant besserem Lernerfolg der Studierenden. Ebenfalls keinen signifikanten Unterschied zeigte der Vergleich zwischen interaktiven und nicht-interaktiven Lernmaterialien, wobei die Experimentalgruppe, die mit der voll-interaktiven Lernumgebung gearbeitet hatte, die besten Testergebnisse erzielte. In Bezug auf die Verfügbarkeit der ESDs kann festgehalten werden, dass auch hier keine signifikanten Ergebnisse erzielt wurden: Keine der Lerngruppen erreichte signifikant höhere Testwerte. Allerdings zeigte sich hier ein interessanter Aspekt: Studierende, die oft mit ESDs arbeiteten, erreichten signifikant höhere Werte, wenn diese gerade nicht verfügbar waren. Umgekehrt hatten auch Studierende, die die ESDs wenig benutzten, ein höheres Testergebnis, wenn ESDs vorhanden waren:

„However, if the two-way interactions are taken into consideration, the analysis uncovers a significant interaction between discernability and the use of processing ESD ($F(1,42)=5.66$, $p<.05$) on study outcome. Additionally we find a trend interaction between discernability and use of testing ESD ($F(1,41)=3.67$, $p=.06$). The features of both interactions are analogous: students who use much processing and testing ESD benefit from a condition (treatment) in which ESD are non-discernable, i.e., integrated into the basic content. In contrast, low ESD-users seem to benefit from a condition in which ESD can be clearly identified“ (a.a.O.:17.)

In Bezug auf den Lernerfolg wurde ein signifikanter Zusammenhang zwischen dem Vorwissen der Studierenden und der Benutzung der ESDs festgestellt, ebenso ein (nicht verwunderlicher bzw. überraschender) Zusammenhang zwischen der Einstellung zum Lernen mit Computern und des Darreichungsmodus der Inhalte.

„The interaction implies that the choice of a certain delivery mode depends on a student's attitude towards learning with computers. Students who like learning with computers will achieve better study outcomes in a computer condition, but will achieve worse study outcomes if they have to study from printed learning materials“ (ebd.).

Die Messungen zeigten ebenfalls keinen Einfluss der Motivation auf verschiedene ESD-Typen. Allerdings war ein leichter Einfluss von Einstellungen gegenüber Computern im Allgemeinen und Nutzung von

ESDs erkennbar. Eine Post-Hoc-Analyse zeigte einen positiven signifikanten Zusammenhang zwischen Einstellung und Motivation.

Des Weiteren wurden keine Effekte in Bezug auf den Einfluss der Motivation auf die Lernergebnisse gefunden. Schaut man sich für die interaktive Lernumgebung jedoch erneut die Variable „Einstellung gegenüber Computern“ an, so kann auch hier ein signifikanter Zusammenhang festgestellt werden:

„In the interactive learning environment, a students' attitude towards learning with computers is a significant, predictive variable for the variable study outcome“ (ebd).

In der Untersuchung sind verschiedene Einflussvariablen auf das Lernen berücksichtigt worden: kognitive Fähigkeiten, Lesekompetenzen, Motivation und Vorwissen. Da mit einer Logfile-Analyse gearbeitet wurde, ist davon auszugehen, dass auch Effekte, die während der Benutzung der Lernumgebung aufgetreten sind, berücksichtigt wurden. Inwiefern dies auch für die Printkurse möglich war, wird nicht erwähnt.

Eigene Kritik an der Studie äußern die AutorInnen insofern, als sie die Stichprobe als gering ansehen:

„A possible explanation for the non-significant differences can be found in the relatively small number of students who studied in the non-interactive learning environment setting“ (a.a.O.:18).

Das Kriterium wurde somit überwiegend erfüllt und wird mit eher positiv bewertet.

8.3.3.3 Zuverlässigkeit

Angemessenheit der Forschungsentscheidungen

Es wurden sechs Hypothesen aufgestellt. Die ersten drei zielten dabei auf die Effektivität der Lernumgebung ab, die letzten drei auf die Interaktion zwischen Studierenden, den Gebrauch der Lernumgebung und den Lernerfolg. Sie lauteten:

1. Studierende, die mit der interaktiven Lernumgebung arbeiten, werden zumindest gleiche Ergebnisse (study outcomes) erzielen, wie diejenigen, die „face-to-face“ studieren.
2. Studierende, die mit den ESDs arbeiten, profitieren von Lernumgebungen, in welchen die ESDs weniger sichtbar sind. Andererseits profitieren auch Studierende, die keine ESDs benutzen, von Lernumgebungen, in denen ESDs sichtbar sind⁷¹.

⁷¹ “This interaction is expected because low-users may want to skip certain ESD. When ESD are nondiscernable this might hinder these learners because they have to read the ESD before they know what ESD it is. On the other hand high-users who always tend to read ESD might find the different lay-out of ESD obstructive”(Rob Martens, Stanley J. Portier & Martin Valcke 1995:8).

3. Studierende, die in der interaktiven Lernumgebung studieren, erzielen höhere Lernerfolge als Studierende in einer nicht-interaktiven Lernumgebung.
4. Vorwissen macht sich im positiven Sinn bei der Benutzung der Funktionalitäten der interaktiven Lernumgebung bemerkbar und stellt eine signifikante Interaktionsvariable im Zusammenhang von Maßnahme und Lernerfolg dar.
5. Gute LeserInnen benutzen weniger ESDs, um die eigenen Aktivitäten zu unterstützen.
6. Es gibt eine positive Korrelation zwischen Motivation und ESD-Benutzung, ihre Interaktion beeinflusst den Lernerfolg (Übersetzung AP).

Für die vorliegende Meta-Evaluation sind nur Hypothese 1, 3 und 6 von Bedeutung.

Ebenfalls wird die Wahl der Auswertungsmethode begründet (vgl. die Beschreibung von Zwecken und Vorgehen). Auf die Auswahl der Stichprobe wird jedoch nicht gesondert eingegangen.

Bei der abschließenden Einschätzung fallen mehrere Aspekte auf: Zunächst stellt sich die Frage, inwiefern sich die Materialien tatsächlich miteinander vergleichen ließen. Es gab auf der einen Seite verschiedene Versionen der Lernumgebung und der Printmaterialien, auf der anderen Seite aber auch verschiedene Untersuchungsgruppen. Dabei stellte die Kontrollgruppe eine Gruppe dar, die mit keiner der vorgestellten Materialien arbeitete, sondern „nur“ eine Vorlesung besuchte.

Alle Gruppen wurden mit dem gleichen Wissenstest geprüft, obwohl bereits vorher deutlich gemacht wurde, inwiefern sich die Gruppen und das Lernen innerhalb der Gruppen voneinander unterscheiden. In diesem Fall kann die Forschungsentscheidung nicht als angemessen bezeichnet werden.

Das Kriterium wird teilweise erfüllt und kann mit neutral bewertet werden.

8.3.3.4 Bezug zu den Forschungsfragen

Lernerfolgsmessung

Kognitive Fähigkeiten, Lesekompetenzen, Motivation und Vorwissen zählten in der vorliegenden Studie als Einflussfaktoren in Bezug auf das Lernen. Zur Feststellung des Lernerfolgs wurde neben den Abschlussnoten ein so genannter „Mastery Test“ durchgeführt, der aus einem dreiteiligen Fragebogen bestand:

- A) „includes questions about motivation, age, study habits, educational level and gender;
- B) asks students to judge the degree of accessibility of the learning materials studied;

- C) puts forward questions about the ease and appreciation of each specific embedded support device" (Rob Martens, Stanley J. Portier & Martin Valcke 1995:16).

Dieser beinhaltet Fragen zu den Rahmenbedingungen.

Betrachtet man beispielsweise, wie der Lernerfolg operationalisiert wurde, fällt auf, dass ein besonderes Interesse auf den Ergebnissen des Mastery-Tests und des regulären Abschlusstests der Hochschule lag. Insofern wurde nicht der Lernprozess, sondern eher die Performance untersucht: Je höher das Testergebnis, desto höher der Lernerfolg.

Die Lernziele wurden bei der Betrachtung der Hypothesen deutlich, die für die Untersuchung formuliert wurden. Diese wurden ausführlich analysiert. Allerdings wird nicht klar, welche inhaltlichen Lernziele gestellt wurden. Es ist jedoch davon auszugehen, dass es sich um einen Standardkurs der OU handelte, der mit einer standardisierten Klausur abschloss. Da der Inhalt des Kursmaterials nicht bekannt ist, ist es an dieser Stelle schwierig, eine Einordnung in die Taxonomie vorzunehmen. Betrachtet man jedoch die Art und Weise des Lernens und des Lernprozesses sowie die Methodik der Messung des Lernerfolgs, kann folgende Vermutung geäußert werden:

Als kognitive Dimension lässt sich für Testgruppe 1 und 2 am ehesten Level 3 „anwenden“ angeben, da die Studierenden mit der interaktiven Lernumgebung gearbeitet haben und direktes Feedback nutzen konnten, während die Testgruppen 4 und 5 sowie die Kontrollgruppe diese Möglichkeit nicht hatten. Daher möchte ich sie an dieser Stelle im Level 2 „verstehen“ einordnen.

DIE KOGNITIVE PROZESS-DIMENSION					
1. Erinnern	2. Verstehen	3. Anwenden	4. Analysieren	5. Bewerten	6. (Er)schaffen
	Lernziele EG3,4,5 KG1	Lernziele EG1+2			

Tabelle 8: Einordnung von P 3 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)

Dass abschließend alle Gruppen mit dem gleichen Wissenstest geprüft wurden, halte ich an dieser Stelle also nur bedingt für richtig. In Bezug auf die Einordnung in die Taxonomie der Lernziele wäre es meiner Auffassung nach nur zulässig gewesen, die nicht-interaktiven Gruppen (also EG 3-5 und die Kontrollgruppe) miteinander zu vergleichen. Allerdings ist es auf Grund der Unkenntnis des genauen Studienmaterials nicht möglich, begründete Aussagen zu treffen.

Das Kriterium wird mit eher negativ bewertet.

Signifikante Ergebnisse

Auf die Ergebnisse der Untersuchung wurde bereits ausführlich eingegangen. Diese werden einem Test auf Signifikanz unterzogen, weshalb das Kriterium als erfüllt und damit positiv bewertet werden kann.

8.3.3.5 Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Durchführbarkeit						
1.1	Angemessene Verfahren	x				
1.2	Praktische Verfahren	x				
2. Genauigkeit						
2.1	Valide und reliable Informationen			x		
2.2	Begründete Schlussfolgerungen		x			
3. Zuverlässigkeit						
3.1	Angemessenheit der Forschungsentscheidungen			x		
4. Bezug zu Forschungsfragen						
4.1	Lernerfolgsmessung				x	
4.2	Signifikante Ergebnisse	x				
	Gesamtbewertung	3	1	2	1	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 9: Bewertungsraster P3

8.3.4 P 4: Josef Smolle, Reinhard Staber, Heide Neges, Gilbert Reibnegger & Helmut Kerl (2005)

In der im Folgenden beschriebenen Evaluationsstudie wurde Präsenz- mit Online-Lernen verglichen, wobei beide Varianten in ein und demselben 30-stündigen Seminar durchgeführt und auch beide von den Studierenden belegt werden mussten. 24 Unterrichtsstunden wurden in Präsenzform vermittelt, 6 virtuell.

8.3.4.1 Durchführbarkeit

Angemessene Verfahren

Der Rahmen der Evaluation sowie deren Zeitraum werden beschrieben. Aufwand und Nutzen standen m.E. jedoch in einem wenig ausgewogenen Verhältnis: So wurde einerseits eine Untersuchung über 5 Wochen mit relativ aufwändigen Verfahren durchgeführt, andererseits waren nur 26 Untersuchungspersonen beteiligt, weshalb zur Anwendbarkeit der Ergebnisse keine allgemeinen Aussagen getroffen werden können. Die Verfahren waren daher nur teilweise angemessen – das Kriterium wird mit neutral bewertet.

Praktische Verfahren

Zum Vergleich von Präsenz- und Online-Lernen wurde ein Seminar im Studienmodul Dermatoonkologie mit 30 Stunden Dauer (24 in Präsenzform, 6 virtuell) untersucht. Dieses schloss mit einem Multiple-Choice-Test ab, der zu gleichen Teilen Fragen zum Inhalt von Präsenz- bzw. Online-Unterricht enthielt.

Da die Studierenden jeweils den Präsenz- und den Online-Teil des Kurses absolvieren mussten, bildeten sich dementsprechend die zwei Untersuchungsgruppen. Es wurden jedoch keine expliziten Kontrollgruppen gebildet.

Die Kriterien der Gruppenbildung sind somit klar erläutert und aus der Sicht der AutorInnen nachvollziehbar. Das Kriterium wird mit eher positiv gewertet, da es überwiegend erfüllt wurde.

8.3.4.2 Genauigkeit

Valide und reliable Informationen

Um zu überprüfen, welches Lernsetting den höheren Lernerfolg erzielte, wurden die Prüfungsergebnisse untersucht, „die mit einem elektronischen Format im Vergleich zu einem Format der Präsenzlehre erzielbar sind“ (a.a.O.:884). Diese Leistungsüberprüfung fand in Form eines Multiple-Choice-Tests statt, der am Ende des gesamten Kurses durchgeführt wurde.

„Es stand ein Fragenpool von 80 Fragen zur Verfügung, von denen je 40 aus dem Team-Learning und 40 aus dem CBT stammten. Jeder Studierende erhielt eine Zufallsauswahl von 20 Fragen aus

dem Team-Learning und 20 aus dem CBT zur Prüfung vorgelegt“ (Josef Smolle, Reinhard Staber, Heide Neges, Gilbert Reibnegger & Helmut Kerl 2005:885).

Die durch diese Prüfung erzielten Ergebnisse sowie die Leistungszahl⁷² dienten als Datenmaterial für die statistische Analyse. Angegeben ist, dass deskriptive Auswertungsverfahren, Student's T-Test, der Wilcoxon-Test für gepaarte Stichproben und der Korrelationstest nach Pearson angewendet wurden. Darüber hinaus wurden die Kommentare der Studierenden untersucht, zu denen nach jeder Absolvierung eines Lernobjekts Gelegenheit gegeben wurde. Am Ende des Moduls fand eine Befragung mittels formalisiertem Fragebogen statt.

Es fehlen Angaben zu Validität und Reliabilität. Die Datengrundlage ist dagegen gut beschrieben worden.

Das Kriterium wird als neutral gewertet, da es teilweise erfüllt wird.

Begründete Schlussfolgerungen

Die Ergebnisse der Evaluation werden detailliert beschrieben. Zusammengefasst kann festgehalten werden:

„Die Prüfungsergebnisse waren bei jenen Fragen, die sich auf Computer-based Trainings-Inhalte bezogen, mit $90,4 \pm 12,5\%$ richtigen Antworten signifikant besser als bei jenen, die sich aus dem Stoff des Team-Learning rekrutiert hatten ($80,6 \pm 13,1\%$; $p < 0,001$). Die Ergebnisse beider Prüfungsteile zeigten innerhalb der Einzelpersonen eine starke Korrelation ($r=0,52$, $p<0,01$), jedoch keinen statistischen Zusammenhang mit den übrigen Prüfungsleistungen der Studierenden“ (a.a.O.:883).

Die AutorInnen äußern in der vorliegenden Publikation Interpretationen und Schlussfolgerungen der Ergebnisse, deren Begründungen sich auf die Daten beziehen. Es ist allerdings fraglich, ob die geringe Fallzahl überhaupt zu interpretierbaren Ergebnissen führt. Weiterhin wird angegeben, dass in der Studie auch Einschränkungen deutlich würden⁷³ und dass damit „über die Anwendbarkeit des Gelernten in einem erweiterten Kontext keine Aussage getroffen werden [könnten]“ (a.a.O.:887f.).

Das Kriterium wird in einigen Bereichen nicht erfüllt und wird somit als neutral gewertet.

⁷² allgemeines Maß für den bisherigen Studienerfolg

⁷³ „Die Studie zeigt allerdings zahlreiche Einschränkungen. Eine davon ist, dass nicht gleiche Lerninhalte auf zwei verschiedene Arten vermittelt worden sind, sondern dass es sich zwar um vergleichbare, aber doch nicht identische Lerninhalte gehandelt hat. Dies ist darin begründet, dass die Studie nicht unter Labor-, sondern unter Alltagsbedingungen des Medizinstudiums durchgeführt wurde“ (a.a.O.:887)

8.3.4.3 Zuverlässigkeit

Angemessenheit der Forschungsentscheidungen

Die Forschungsfrage lautete, ob der Erfolg beim Lernen mit Computerlehrprogrammen in Form von Computer Based Training geringer oder höher sei als der Lernerfolg, der mit „innovativen didaktischen Konzepten im Präsenzunterricht“ (a.a.O.:2005:883) erzielt würde. Die AutorInnen bezogen sich bei dieser Fragestellung auf den wachsenden Einfluss neuer Medien im Bildungsbereich, speziell in der medizinischen Ausbildung.

In der Publikation werden keine Hypothesen benannt. Da es sich jedoch hierbei nicht um einen reinen Forschungsbericht handelt, bedeutet das Fehlen von Hypothesen in diesem Kontext nicht, dass generell darauf verzichtet wurde.

Die Wahl der Stichprobe wird erläutert, Begründungen für diese Auswahl werden aus dem Zusammenhang deutlich. Nicht erläutert werden hingegen die Gründe für die Wahl der Auswertungsmethode.

Problematisch ist m.E., dass hier kollaboratives Lernen im Präsenzunterricht mit selbständigem Lernen mittels CBT verglichen wurde, was jeweils einen anderen Lernbegriff impliziert. Dennoch wurde die Prüfung nach gleichem Verfahren, einem Multiple-Choice-Test, der also lediglich Wissen abfragt, durchgeführt. Dies bedeutet, dass beim Vergleich der Gruppen die unabhängigen Variablen konstant gehalten wurden. Die Interpretationen der Ergebnisse berücksichtigten dies jedoch nicht, sondern bezogen sich auf genau diese Variablen. Auf die geringe Fallzahl wurde an anderer Stelle bereits eingegangen. Weiterhin ist zu kritisieren, dass der Abschlusstest am Ende des Moduls und nicht unmittelbar nach Absolvieren der einzelnen Kursteile durchgeführt wurde. Auf diese Weise wurde in Kauf genommen, dass Wissen aus gerade bearbeiteten Aufgaben präsenter war als Wissen, das bereits vor einigen Wochen angeeignet worden war.

Das Kriterium ist damit in weiten Bereichen nicht erfüllt und wird mit eher negativ gewertet.

8.3.4.4 Bezug zu den Forschungsfragen

Lernerfolgsmessung

In der vorliegenden Studie wurde Lernerfolg mit Behaltensleistung gleichgesetzt und mit Wissensabfragen abgeprüft. Lernprozess und Lernleistung wurden nicht unterschieden. Der zu Grunde liegende Lernbegriff wurde sehr allgemein formuliert: Kollaboratives Team-Lernen und Selbstlernen am Computer wurden miteinander vermischt und in einem einzigen Lernbegriff subsumiert.

Zwar waren keine Lernziele angegeben, jedoch lässt allein die Art und Weise der jeweiligen Settings darauf schließen, dass im Präsenzkurs Wissen auf der Ebene des Analysierens erlernt werden sollte. Im Online-Kurs wurde das Gelernte weniger reflektiert – auch, wenn AutorInnen

eine Bezeichnung als behavioristische Lernform ablehnen. Die Lernziele können also maximal auf der Ebene des Anwendens eingeordnet werden.

DIE KOGNITIVE PROZESS-DIMENSION					
1. Erinnern	2. Verstehen	3. Anwenden	4. Analysieren	5. Bewerten	6. (Er)schaffen
Test beider Gruppen		Lernziel Online	Lernziel Präsenz		

Tabelle 10: Einordnung von P 4 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)

Auch an dieser Stelle zeigt sich, dass Faktoren miteinander verglichen werden, die unterschiedliche Ausprägungen von Lernerfolg zu Grunde legen. Das Kriterium wird kaum erfüllt und somit als eher negativ gewertet.

Signifikante Ergebnisse

Die Ergebnisse wurden dokumentiert und auf Signifikanz untersucht. Der nicht-parametrische paarweise Vergleich zeigte dabei ein signifikant besseres Abschneiden des CBT-Kurses in Bezug auf die Prüfungsergebnisse.

Ebenfalls wurde die Trefferquote der eingesandten Rückmeldungen mit dem CBT mit den Prüfungsleistungen der einzelnen Studierenden auf Signifikanz getestet. Hier ist ein hoch signifikanter Zusammenhang festzustellen ($r=0,791$, $p<0,001$). Auch der Vergleich der Rückmeldungen aus dem Team-Lernen mit den Prüfungsleistungen weist einen signifikanten Zusammenhang auf ($r=0,40$; $p<0,05$).

Das Kriterium wurde erfüllt und wird damit als positiv gewertet.

8.3.4.5 Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Durchführbarkeit						
1.1	Angemessene Verfahren			x		
1.2	Praktische Verfahren		x			
2. Genauigkeit						
2.1	Valide und reliable Informationen			x		
2.2	Begründete Schlussfolgerungen			x		
3. Zuverlässigkeit						
3.1	Angemessenheit der Forschungsentscheidungen				x	
4. Bezug zu Forschungsfragen						
4.1	Lernerfolgsmessung				x	
4.2	Signifikante Ergebnisse	x				
	Gesamtbewertung	1	1	3	2	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 11: Bewertungsraster P4

8.3.5 P 5: Randall S. Davies & Robert Mendenhall (1998)

Bei der vorliegenden Primärstudie aus dem Jahr 1998 handelt es sich um einen Forschungsbericht zu einer Evaluation, die WWW-basiertes Online-Lernen mit Instruktion im Klassenraum verglich. Beide Darbietungsmodi bezogen sich auf den Kurs „HEPE⁷⁴ 129 - Fitness and Lifestyle Management“, der im Studiengang „Health Education/ Physical Education“ an der Brigham Young University im US-Bundesstaat Utah durchgeführt wurde. Ziel der Untersuchung war, Daten über die Qualität des Unterrichts in die Diskussion um den Wert webbasierter Kurse einzubringen.

8.3.5.1 Durchführbarkeit

Angemessene Verfahren

Der Rahmen der Untersuchung wird erläutert, schien allerdings nur bedingt angemessen zu sein. So ist m.E. die Verpflichtung, eine Blutprobe abzugeben, eine relativ große Belastung für die beteiligten Studierenden, deren Sinn noch nicht einmal deutlich wird. Bei der Kritik, die die Autoren an ihrer eigenen Studie üben, wird deutlich, dass der betriebene Aufwand in keinem angemessenen Verhältnis zum Nutzen der Untersuchung stand:

„We did not use the online lesson exams, fitness assessment, the exercise log or the online final examination. Thus we could not assess the effectiveness of these aspects of the course. We were, however, able to make comparisons in student performance on specific lessons and student preferences for instructional methods. A study using the fully implemented online course is something that will have to take place when the numbers of students choosing that option increases“ (Randall S. Davies & Robert Mendenhall (1998:10).

Darüber hinaus war die Evaluation sehr breit angelegt, untersuchte letztlich aber doch nur einen kleinen Teil des Kursprogramms. Das Kriterium wurde nur teilweise erfüllt und wird als eher negativ gewertet.

Praktische Verfahren

Um die Zielgruppe für die Untersuchung zu rekrutieren, wurden die Studierenden, die die Stunden online absolvieren sollten, um eine freiwillige Beteiligung gebeten. Es stellten sich letztlich 39 Studierende zur Verfügung, von denen 15 die Inhalte zu Hause bearbeiteten, 24 im Computerraum der Universität. Diese Freiwilligen wurden daraufhin zufällig auf drei Gruppen aufgeteilt: Zwei Gruppen studierten die Inhalte nur unter Zuhilfenahme des Online-Materials, eine Gruppe wurde im Klassenraum unterrichtet.

⁷⁴ Health Education/ Physical Education

Während des Kurses waren drei Prüfungen erforderlich. Zwischen diesen „exams“ wurden die Gruppen rotiert, damit jede Gruppe mit jeder Lernform Erfahrung machen konnte.

Die Kriterien der Gruppenbildung wurden nachvollziehbar erläutert. Das Kriterium wird mit positiv bewertet.

8.3.5.2 Genauigkeit

Valide und reliable Informationen

Im Präsenzkurs wurden die Prüfungen zu 6 Zeiträumen während des laufenden Kurses durchgeführt und bezogen sich jeweils auf 2 bzw. 3 Module. Jede Prüfung konnte bis zu 2x gegen eine Gebühr wiederholt werden. Die Endnote des Kurses wurde zu 75% durch diese Modulprüfungen abgedeckt. Weitere 15% machten die Fitness-Übungen (Fitness-Vertrag, Fitness-Programm und Fitness-Check) aus. Für den Kurs im Klassenraum gab es 2 Lehrbücher und Leseaufgaben, die sich auf die einzelnen Module bezogen. Ferner wurde nach jeder Vorlesung auf ausgewähltes weiterführendes Material verwiesen, das in der Bibliothek beschafft werden konnte. Der Unterricht wurde im Klassenraum zusätzlich durch Beobachtungen untersucht.

Die Ergebnisse der Lektionentests des Online-Kurses flossen zu 65% in die Endnote ein. Ferner machten die Fitness-Übungen (Fitness-Vertrag, Fitness-Programm und Fitness-Check) 15% der Endnote aus. Weitere 20% der Endnote wurden durch einen protokollierten Paper-Pencil-Test abgedeckt.

In allen drei Untersuchungsgruppen wurden nach jeder Prüfung Interviews durchgeführt, die die Haltungen und Meinungen der Studierenden abfragten:

„Students responded to questions about how they prepared for the exam, their reaction to the instruction and whether they felt they had learned something from the lessons (a.a.O.:9).

Der Fragebogen umfasste Fragen zur Vorbereitung auf die Prüfungen, die als Indikator dafür dienen sollten, in welchem Umfang sich die Studierenden mit dem zur Verfügung gestellten Material beschäftigt hatten. Jede Untersuchungsgruppe wurde wiederum in 2 Untergruppen aufgeteilt:

„Students from each group were then divided into two subgroups - students who participated fully in the instruction and those who only partially made use of the instruction provided. We expected that students would do better if they actually used the instruction provided. If no difference was noticed then that method of instruction may be ineffective“ (ebd.).

Die zu analysierenden Daten wurden aus den Testergebnissen in den Untersuchungsgruppen gewonnen. Für jede Gruppe wurden statistische Berechnungen durchgeführt:

„Data was gathered from test results for both in-class and online students. Mean scores and standard deviations were calculated for each group. A measure of statistical significance was used to calculate the effect size. In addition, survey results were tabulated; classroom observations and interview information was organized by topic in order to identify issues and trends“ (ebd.).

Wie bereits erwähnt, wurden die Präsenzkurse darüber hinaus durch Beobachtungen untersucht. Ein Fragebogen wurde in allen Untersuchungsgruppen zusätzlich am Ende der Evaluation eingesetzt.

Die Verfahren der Datensammlung und –analyse werden beschrieben, sind jedoch nur teilweise nachvollziehbar. Die Durchführung von Signifikanztests zur Messung der Effektgröße ist beispielsweise kein geeignetes Verfahren. Sowohl die Messung von Signifikanz als auch die Messung der Effektgröße erfordern eigenständige Verfahren, wenngleich sie beide auf die praktische Relevanz der Wirkung einer Maßnahme abzielen. Studien mit großer Fallzahl und damit hoher Teststärke führen darüber hinaus oft zu hoch statistischen Ergebnissen. Dennoch kann der Effekt so gering sein, dass sich keine praktischen Interpretationen daraus ableiten lassen. Signifikanz ist also kein hinreichendes Kriterium für praktische Relevanz.

Darüber hinaus wurde bei der Entwicklung der Instrumente nicht auf Gütekriterien geachtet:

„In addition, we make the assumption that the assessment instruments being used by classroom teachers are valid for this course. Other than an item analysis to determine the degree of difficulty for each item, we did not complete an evaluation of the testing instruments used“ (a.a.O.:10).

Das Kriterium wird mit eher negativ bewertet, da es nur ansatzweise erfüllt werden konnte.

Begründete Schlussfolgerungen

Die Ergebnisse der Evaluation werden ausführlich dokumentiert. So ist festzuhalten, dass ein Vergleich der Untersuchungsgruppen keine signifikanten Unterschiede hervorbringen konnte. Die Autoren schließen daraus, dass beide Lehr- /Lernmethoden gleichermaßen geeignet gewesen seien. Die Befragung der Studierenden zeigte jedoch, dass das Lernen im Klassenraum die bevorzugte Methode war. Es stellt sich jedoch die Frage, ob diese Schlussfolgerung zulässig ist, da die Studierenden entweder das eine oder das andere Lernsetting belegt hatten und daher m.E. keine Aussagen über die jeweils andere Methode hätten treffen können

Unter Bezugnahme auf die Forschungsfragen stellen sich die Ergebnisse wie folgt dar:

1. Erreichen Studierende, die mit unterschiedlichen Instruktionsmodi lernen, gleiche Ergebnisse wie die Studierenden, die auf dem Campus lernen?

Diese Frage war mit „ja“ zu beantworten, da keine signifikanten Unterschiede gefunden werden konnten.

2. Sind die Studierenden der Ansicht, dass sich ihre Ansichten und ihr Verhalten nach Absolvieren des Kurses geändert hat?

Zwei Drittel der Studierenden gaben an, etwas gelernt zu haben, das sie in besonderem Maße motivierte, den eigenen Lebenswandel dauerhaft zu ändern.

„Most were optimistic that the changes they were making in their lifestyle would be permanent. This optimism was found in both groups but due to the limitations of the study's design, this optimistic attitude may speak more to the course in general and not to differences in instructional delivery methods (a.a.O.:16)“.

3. Benutzen die Studierenden die verfügbaren instruktionalen Ressourcen?

Nicht alle Studierenden nutzten die Ressourcen. Die Beteiligung im Klassenraum war vergleichsweise gering und die Online-Studierenden beschränkten ihre Beschäftigung auf das Lesen des Materials.

4. Präferieren die Studierenden eine Darreichungsform gegenüber der anderen?

Wie bereits erwähnt, wurde das Lernen im Klassenraum bevorzugt.

5. Finden die Studierenden eine Darbietungsform nützlicher als die andere?

Es wurden keine nennenswerten Unterschiede festgestellt. Die Studierenden konnten in ihren Aussagen zu jeder Methode Nachteile angeben.

Die Autoren geben im Forschungsbericht die Limitationen der Untersuchung an. Dabei stellen sie fest, dass persönliche Umstände der Studierenden wichtige Faktoren bei der Bewertung der Lernsettings darstellten, die jedoch nicht erfasst worden waren. Darüber hinaus werden widrige Umstände beschrieben, die während der Untersuchung auftraten:

„There were several restricting circumstances that affected how we conducted this study. Only two students were enrolled for the full online course due to the fact that the course had just been made available to students. The low enrollment meant we were not able to obtain adequate data from students taking the online course as the sole means of instruction. Because of this, this evaluation did not study a full implementation of the online course. This means that we were not able to make a comparison of course completion rates and of the overall online learning experience“ (a.a.O.:10).

Wie bereits erwähnt, wurde in der vorliegenden Studie deutlich, dass der betriebene Aufwand in keinem angemessenen Verhältnis zum Nutzen der Untersuchung stand. Dass die Messinstrumente nicht an Gütekriterien orientiert waren, ist ebenfalls bereits erwähnt worden.

Dennoch ist das Kriterium als mindestens neutral zu werten, da die Ergebnisse nicht nur beschrieben werden, sondern auch darauf eingegangen wird, welche Daten diese Schlussfolgerungen zuließen. Ebenso wird eigene Kritik angegeben. Trotz dieser Einschränkung werden bei der Darstellung der Ergebnisse dennoch Aussagen getroffen, die sich auf letztlich unzulässige Daten beziehen. Das Kriterium wird daher mit neutral bewertet.

8.3.5.3 Zuverlässigkeit

Angemessenheit der Forschungsentscheidungen

Die Evaluation zielte ab auf:

- Die Effektivität von Instruktion im Online- und im Klassenraummodus in Bezug auf die Vermittlung der Grundlagen des Kurses HEPE 129,
- den Grad, inwieweit die Instruktion die Haltung der Studierenden in Bezug auf Gesundheit und Fitness beeinflusst,
- Vor- und Nachteile der beiden Instruktionsmethoden,
- Signifikante Unterschiede der Testergebnisse und Auswertungen des studentischen Feedbacks (vgl. a.a.O.:8, Übersetzung AP).

Darüber hinaus wurden Forschungsfragen aufgestellt:

1. Erreichen Studierende, die mit unterschiedlichen Instruktionsmodi lernen, gleiche Ergebnisse wie die Studierenden, die auf dem Campus lernen?
2. Sind die Studierenden der Ansicht, dass sich ihre Ansichten und ihr Verhalten nach Absolvieren des Kurses geändert hat?
3. Benutzen die Studierenden die verfügbaren instruktionalen Ressourcen?
4. Präferieren die Studierenden eine Darreichungsform gegenüber der anderen?

5. Finden die Studierenden eine Darbietungsform nützlicher als die andere?

Die Wahl der Stichprobe wird jedoch nicht näher erläutert. Auch die Gründe für die Wahl der Auswertungsmethode sind nicht bekannt.

Für wesentlich bei der Interpretation und Bewertung der vorliegenden Primärstudie halte ich, dass zwei Lehr- /Lernmethoden miteinander verglichen wurden, die jeweils auf andere Arten von Lernen ausgerichtet waren. Positiv hervorzuheben ist jedoch, dass die Prüfungen in den einzelnen Gruppen auf unterschiedliche Art und Weise durchgeführt wurden. Während in der Online-Gruppe „Speedback“-Tests mit sofortigem Feedback durchgeführt worden sind, wurden im Präsenzkurs klassische Tests geschrieben. In der Online-Klasse wurde ferner ein Abschlusstest als „open book“ geschrieben, der Präsenzkurs schloss mit einem „closed book“-Test ab.

Das Kriterium ist damit teilweise erfüllt und wird mit neutral gewertet.

8.3.5.4 Bezug zu den Forschungsfragen

Lernerfolgsmessung

Der dieser Studie zu Grunde liegende Lernbegriff wurde sehr allgemein gehalten und vermischte Präsenz- und Online-Lernen. Zwischen Lernprozess und Lernleistung wurde nicht unterschieden. Die Messung des Lernerfolgs orientierte sich ausschließlich an den Ergebnissen der schriftlichen Tests.

Als Lernziele waren in beiden Kursmodi das Erlernen von Grundlagen zu Gesundheit und Fitness festzuhalten. Ein begleitendes Fitness-Programm sollte dies verstärken. Es sollte also Wissen auf der Ebene des Anwendens (durch den Fitnesskurs) erlernt werden. Das bedeutet: Es wurden in beiden Gruppen die gleichen Lernziele intendiert, also die gleichen Ausprägungen von Lernerfolg zu Grunde gelegt. Auf dieser Ebene ist ein Vergleich der Gruppen also sinnvoll. Das Prüfungsdesign wurde entsprechend variiert, um nicht unterschiedliche Bedingungen mit gleichen Tests abzufragen, was ebenfalls positiv ist. Allerdings hätten sich die unterschiedlichen Prüfverfahren auch auf die intendierten Lernziele beziehen müssen. Dies ist jedoch nicht erfolgt: Während der Open-Book-Test in der Online-Gruppe darauf hinweist, dass Wissen auf der Ebene des Verstehens geprüft wurde, ist ein Closed-Book-Test eher ein Zeichen dafür, dass Wissen hauptsächlich wiedergegeben werden sollte. Um einen sinnvollen Vergleich der Gruppen anzustellen, hätte sich auch die Prüfung in beiden Gruppen auf begriffliches Wissen auf Anwendungsebene beziehen müssen – lediglich das Prüfungsdesign hätte unterschiedliche Varianten zulassen müssen (beispielsweise summative versus formative Prüfungsverfahren, vgl. Lorin W. Anderson & David R. Krathwohl 2001:101ff.).

Da sich die Evaluation in erster Linie auf die Daten der Prüfung bezog, zeigt sich, dass Faktoren miteinander verglichen werden, die doch unterschiedliche Ausprägungen von Lernerfolg zu Grunde legen.

DIE KOGNITIVE PROZESS-DIMENSION					
1. Erinnern	2. Verstehen	3. Anwenden	4. Analysieren	5. Bewerten	6. (Er)schaffen
Test Präsenz	Test Online	Lernziele			

Tabelle 12: Einordnung von P 5 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)

Das Kriterium wird in einigen Punkten erfüllt und somit als eher positiv gewertet.

Signifikante Ergebnisse

Die Ergebnisse wurden detailliert dokumentiert. Ein Vergleich der Untersuchungsgruppen konnte allerdings keine signifikanten Unterschiede hervorbringen. Da die Unterschiede zwischen den Gruppen jedoch auf Signifikanz getestet wurden, wird das Kriterium mit positiv bewertet.

8.3.5.5 Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1 Durchführbarkeit						
1.1	Angemessene Verfahren				x	
1.2	Praktische Verfahren	x				
2. Genauigkeit						
2.1	Valide und reliable Informationen				x	
2.2	Begründete Schlussfolgerungen			x		
3. Zuverlässigkeit						
3.1	Angemessenheit der Forschungsentscheidungen			x		
4. Bezug zu Forschungsfragen						
4.1	Lernerfolgsmessung		x			
4.2	Signifikante Ergebnisse	x				
	Gesamtbewertung	2	1	2	2	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 13: Bewertungsraster P5

8.3.6 P 6: Vassilis Hartzoulakis (2002)

Die Hauptfragestellung der Untersuchung zielte darauf ab, ob sich die Effektivität von computerunterstütztem Sprachenlernen (CALL⁷⁵) mit den Ergebnissen der traditionellen Form des Unterrichts vergleichen ließe. Die Untersuchung fand in einem Lykeion⁷⁶ in der griechischen Stadt Peristeri statt.

8.3.6.1 Durchführbarkeit

Angemessene Verfahren

Der Rahmen der Evaluation wird beschrieben. Dabei wird darauf hingewiesen, dass die Lehrperson, die die analysierten Unterrichtsstunden abhielt, keine unerfahrene Person gewesen sei und sich beim Lehren nicht von den Ergebnissen der vorliegenden Studie habe beeinflussen lassen. Es sei von ihr darüber hinaus keine Methode besonders favorisiert worden. Die Lernenden zeigten keine negative Einstellung im Hinblick auf Technologie und seien darüber hinaus nicht technikfeindlich.

Weiterhin schien das Verhältnis des Aufwandes zum Nutzen der Evaluation angemessen gewesen zu sein. Auf kulturelle oder soziale Differenzen wurde nicht näher eingegangen, jedoch wurden Gender-Aspekte thematisiert. Das Kriterium wurde überwiegend erfüllt und wird als eher positiv gewertet.

Praktische Verfahren

Die Stichprobe der Untersuchung stellten 103 Lernende in der zweiten und dritten Jahrgangsstufe der Schule dar. Vor der Ziehung der Stichprobe wurde ein Oxford Placement Test zur Feststellung des Wissensstandes im Fach Englisch durchgeführt, der von 83 Lernenden bearbeitet wurde. Ausgehend von den Ergebnissen dieses Tests wurden 47 Schülerinnen und Schüler einer „B“- und 36 Schülerinnen und Schüler einer „C“-Klasse zugeordnet:

Da jedoch die Punkteverteilung in beiden Klassen annähernd gleich war, wurde entschieden, die Daten beider Klassen zu einer Gruppe zusammenzufassen.

„This distribution is almost equal for both classes and the average grade for C class is higher than the average grade for B class. Specifically, a range of 45 points and an average of 63.12% was noted for B (second) class, while a range of 42 points and an average of 74.44% was observed for C (third) class“ (Vassilis Hartzoulakis (2002:23).

⁷⁵ Computer Assisted Language Learning

⁷⁶ Das Lykeion kann nach dem griechischen Schulsystem nach Abschluss des Gymnasiums (Schulpflicht: 9 Jahre) besucht werden. Nach 3 Jahren werden Abschlusstests durchgeführt, mit deren Bestehen man eine Hochschulzugangsberechtigung erwirbt (vgl. <http://de.wikipedia.org/wiki/Lykeion>, letzter Aufruf: 18.08.2007).

Um die Homogenität der Gruppe sicherzustellen, wurde ferner der Gesamtdurchschnitt der Testergebnisse berechnet. Werte unter 50% und über 80% wurden als Extremwerte ausgeklammert, so dass letztendlich 54 repräsentative Lernende in die Untersuchungsgruppe einbezogen wurden. Diesen wurde ein erneuter Prätest zur Messung des Vorwissens vorgelegt, auf Grund dessen Ergebnis die 40 leistungsschwächsten Schülerinnen und Schüler ausgewählt wurden.

„Based on the results of the pretest, the 40 students with the lowest scores ranging from 0% to 37.5%, were selected to compile the final list from which the students who would participate in the experiment would be randomly selected“ (a.a.O.:26).

Da für das Experiment genau 24 Lernende benötigt wurden, wurde diese Anzahl zufällig aus der Gruppe der verbliebenen 40 Schülerinnen und Schüler ausgewählt. Dabei wurden 12 Lernende der Experimental- und 12 Lernende der Kontrollgruppe zugeteilt. Die erste Gruppe (the lab group) bekam ihren Platz im Computerraum, in dem jede/jeder einen eigenen PC zugeteilt bekam. Die zweite Gruppe (the traditional group) wurde in einem regulären Klassenraum unterrichtet.

Die Bildung von Test- und Kontrollgruppe wird nachvollziehbar und detailliert beschrieben. Das Kriterium ist voll erfüllt und wird mit positiv bewertet.

8.3.6.2 Genauigkeit

Valide und reliable Informationen

Um das Vorwissen der Studierenden abzufragen wurde ein Prätest durchgeführt. Dieser wurde zunächst von zwei Experten in Bezug auf die Validität des Inhalts überprüft, woraufhin Anpassungen vorgenommen wurden. Vor dem Einsatz des Tests wurde dieser in einer anderen Kohorte desselben Lykeions getestet und auf Schwächen überprüft.

„Before giving the pretest to the students in the final shortlist, we piloted it with a different group of students in another Lykeio. The piloting did not reveal any weaknesses in the wording of the rubrics or the structure of the exercises and the test seemed clear to the students. The time they required to complete the test was measured and it coincided with our estimate (about 35 minutes on average)“ (a.a.O.:26).

Nach Durchführung des Experiments wurden die Ergebnisse der Leistungsmessung (Punktzahlen) der Prä-, Post-, und Placement-Tests statistisch ausgewertet:

„A T-Test investigating overall differences between the pretest and the posttest findings.

An ANCOVA⁷⁷ investigating whether there were differences in the outcomes of the two teaching methods, and

⁷⁷ analysis of covariance, Anm. AP

An ANOVA⁷⁸ investigating whether there were differences in the performance between males and females" (a.a.O.:32).

Im Vorfeld der Datenanalyse wurde eine Nullhypothese formuliert, um den T-Test durchführen zu können. Diese lautete:

„There were no significant differences between the pre- and the posttest scores" (a.a.O.:36).

Die Alternativhypothese lautete dementsprechend:

„The differences between the pre- and posttest scores were significant" (ebd.).

Auch für die Analyse der Kovarianz wurden ebenfalls Hypothesen aufgestellt:

„‘There is no difference between the two teaching methods’ (H0: $\mu_1 - \mu_2 = 0$). The alternative hypothesis is: ‘The two methods yield different results’ (H1: $\mu_1 \neq \mu_2$)" (a.a.o.:36).

Untersucht wurden im Wesentlichen fünf Variablen: die drei numerischen Variablen der eingesetzten Tests (Prä-, Post- und Placement-Test) und zwei Dummy-Variablen⁷⁹ für das Geschlecht und die Lehrmethode.

Die Verfahren der Datengewinnung und -analyse werden detailliert und nachvollziehbar beschrieben. Ebenso wird auf Gütekriterien eingegangen. Allerdings wird auf die Dokumentation der Skalenbildung verzichtet, weshalb die Reliabilität der Skalen nicht deutlich wird. Das Kriterium wird mit eher positiv bewertet, da es überwiegend erfüllt wurde.

Begründete Schlussfolgerungen

Die Studie kommt zu dem Ergebnis, dass schlechtere Lernergebnisse in der Computerklasse – der Experimentalgruppe – erzielt worden sind, deren Mittelwert 54.86 Punkte betrug. Die Kontrollgruppe im Klassenzimmer wies einen Mittelwert von 73.61 Punkten auf. Die Ergebnisse werden in der Arbeit sehr detailliert dargestellt.

Dabei wurde u.a. der Frage nachgegangen, ob die computervermittelte Instruktion für Jungen und Mädchen gleich effektiv ist. Dem kann auf Grund der Datenlage zugestimmt werden.

„All students improved significantly due to the two teaching methods. The students in the ‘traditional’ group did significantly better than the students in the ‘lab’ group" (a.a.O.:42).

Der Autor zieht daraus den Schluss, dass griechische SchülerInnen den traditionellen Methoden eher verbunden waren. Dieser Schlussfolgerung kann nur bedingt zugestimmt werden (siehe: Angemessenheit der Forschungsentscheidungen).

⁷⁸ analysis of variance, Anm. AP

⁷⁹ Unter einer Dummy-Variable wird eine binäre Variable verstanden, die die Ausprägungen 0 und 1 aufweist. Diese steht als Platzhalter (also als Dummy) für eine Ausprägung einer mehrstufigen Variablen.

Darüber hinaus gibt der Autor an, aus informalen, nicht aufgezeichneten Gesprächen mit den Lernenden in Erfahrung gebracht zu haben, dass diese den Computerunterricht interessant gefunden und ihn genossen hätten. Da diese Aussagen nicht wissenschaftlich belegbar sind, sind sie an dieser Stelle ebenfalls nicht zulässig.

Es bleibt anzumerken, dass eine Stichprobe von 24 Lernenden grundsätzlich keine besonders aussagekräftigen Schlussfolgerungen zulässt. Dies äußert der Autor selbst:

„The number of the students who were sampled for the experiment that is presented in the study was quite small in proportion to the population the author wanted to generalize to. Furthermore, the experiment was conducted within the framework of the regular teaching hours in the author's school; therefore, the sample is also limited in terms of the geographical area covered“ (a.a.O.:8).

Das Kriterium konnte also nur teilweise erfüllt werden. Die Ergebnisse der Studie werden detailliert beschrieben und es wird eigene Kritik geäußert. Die Schlussfolgerungen beziehen sich zwar auf die erhobenen Daten, werden aber nicht ausreichend begründet und sind nicht in allen Punkten nachvollziehbar. Das Kriterium wird mit neutral bewertet.

8.3.6.3 Zuverlässigkeit

Angemessenheit der Forschungsentscheidungen

Im Vorfeld der Untersuchung wurden Hypothesen aufgestellt. Auf die Wahl der Stichprobe sowie die Auswahl der Auswertungsmethode wird detailliert eingegangen, jedoch fehlen an vielen Stellen Begründungen.

Es wird an dieser Stelle deutlich, dass – wie bereits in einigen anderen Primärstudien auch – der gleiche Lernbegriff für zwei unterschiedliche Dinge verwendet wurde. Während im Präsenzunterricht gemeinsam gelernt und auf Fragen Bezug genommen wurde, war die Computerklasse weitgehend auf sich selbst gestellt.

„The differentiating characteristic of the lab group was that the students worked on their own and tried to grasp the particulars of inversion by going back and forth to the web pages presenting the rules and the examples. The traditional group on the other hand, had the chance to exchange ideas and listen to answers as a group rather than individually“ (ebd.).

Die gezogenen Schlüsse – dass beispielsweise griechische SchülerInnen eher dem Präsenzunterricht verbunden sind – kann demnach nicht vollständig unterstützt werden. Beim Vergleich der Gruppen wurden die unabhängigen Variablen bewusst konstant gehalten, die Interpretationen der Ergebnisse beziehen sich aber auf diese Variablen.

Das Kriterium wurde teilweise erfüllt und ist mit neutral zu bewerten.

8.3.6.4 Bezug zu den Forschungsfragen

Lernerfolgsmessung

Die unabhängige Variable Lernerfolg wurde in der vorliegenden Untersuchung ausschließlich als Performanceleistung definiert und zwar an Hand der Ergebnisse von Prä-, Post- und Placement-Test als abhängige Variablen. Einen höheren Lernerfolg erzielten dieser Definition nach also Lernende, die bessere Noten erreichten.

Es wurden keine expliziten Lernziele in der Publikation erwähnt. Lediglich das Ziel der Unterrichtsreihe „Inversion“ wurde deutlich gemacht. In beiden Untersuchungsgruppen wurden die gleichen Wissensabfragen durchgeführt, woraus geschlossen werden kann, dass spezielle Eigenschaften neuer Medien nicht genutzt worden sind. In den Tests beider Gruppen sollte erworbenes Wissen hauptsächlich auf der Ebene des Erinnerns wiedergegeben werden.

Allerdings kann durch die Art des Unterrichts davon ausgegangen werden, dass Lernende in der Präsenzkasse durch Diskussionen und kooperativere Lernformen eher auf der Ebene des Verstehens erlernt haben. Insofern sind die Gruppen nicht miteinander vergleichbar, da sie unterschiedliche Ausprägungen von Lernerfolg zu Grunde legen.

DIE KOGNITIVE PROZESS-DIMENSION					
1. Erinnern	2. Verstehen	3. Anwenden	4. Analysieren	5. Bewerten	6. (Er)schaffen
Test beider Gruppen Lernziel PC	Lernziel Präsenz				

Tabelle 14: Einordnung von P 6 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)

Das Kriterium wurde eher nicht erfüllt und wird mit eher negativ bewertet.

Signifikante Ergebnisse

Statistisch signifikante Unterschiede konnten im Vergleich von Prä- und Posttest sowie in der Favorisierung einer Lehrmethode ausgemacht werden konnten. Dabei wurden schlechtere Lernergebnisse in der Computerklasse – der Experimentalgruppe – erzielt, deren Mittelwert 54.86 Punkte betrug. Die Kontrollgruppe im Klassenzimmer wies einen Mittelwert von 73.61 Punkten auf.

„The results of the ANCOVA indicated that there were significant differences between the two teaching methods. The descriptive statistics in TABLE 10 show that the control group had a higher gain than the treatment group. The difference between the group

means after covarying the pretest scores, is significant, $F = 5.09$, $p = .035$ (TABLE 11). Thus, the null hypothesis is rejected in favour of the alternative hypothesis that the two methods yielded different results" (a.a.O.:37).

Keine Differenzen wurden jedoch beim Vergleich von Mädchen und Jungen deutlich. Zwar schnitten die Jungen geringfügig besser ab, diese Unterschiede waren jedoch nicht statistisch signifikant.

Das Kriterium wird erfüllt und mit positiv bewertet.

8.3.6.5 Bewertungsraster

Nr.	Kriterium	*	#	+	0	-
1. Durchführbarkeit						
1.1	Angemessene Verfahren		x			
1.2	Praktische Verfahren	x				
2. Genauigkeit						
2.1	Valide und reliable Informationen		x			
2.2	Begründete Schlussfolgerungen			x		
3. Zuverlässigkeit						
3.1	Angemessenheit der Forschungsentscheidungen			x		
4. Bezug zu Forschungsfragen						
4.1	Lernerfolgsmessung				x	
4.2	Signifikante Ergebnisse	x				
	Gesamtbewertung	2	2	2	1	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 15: Bewertungsraster P6

8.3.7 P 7: William J. Ryan (2002)

Bei der im Folgenden beschriebenen Primärstudie handelt es sich um einen Tagungsbeitrag⁸⁰, der anschließend in der ERIC-Database veröffentlicht wurde. Anders als bei einem Forschungsbericht kann dabei nicht auf alle relevanten Kriterien gleichermaßen eingegangen werden, so dass dieser Aspekt bei der Bewertung der Primärstudie zu berücksichtigen sein wird. Die vorliegende Publikation beschreibt eine Untersuchung, die am Lakeland Community College durchgeführt wurde. Diese hatte zum Ziel, die Auswirkungen webbasierter Angebote auf das Lernen im Vergleich mit äquivalenten Angeboten im Klassenraum zu vergleichen.

Die hier beschriebene Primärstudie wurde zwischen 1999 und 2000 durchgeführt. Sie war als Pilotstudie anzusehen, aus der sich anschließend ein eigenständiges Forschungsprojekt entwickelte. In den Jahren 2000 und 2001 wurden im Anschluss an die Pilotstudie alle angebotenen Fernlehrcurse untersucht, allerdings wurden dabei nur die Abgangsraten und die GPAs⁸¹ miteinander verglichen. Die Ergebnisse werden in der vorliegenden Publikation mit den Ergebnissen der Pilotstudie in Beziehung gesetzt. Da jedoch nur die erste Erhebungsphase ausführlich beschrieben wird, wird in der folgenden Analyse auch nur auf diesen Teil der Untersuchung eingegangen.

8.3.7.1 Durchführbarkeit

Angemessene Verfahren

Der Rahmen der Untersuchung wird erläutert – das Verhältnis des Aufwandes zum Nutzen der Evaluation schien angemessen zu sein. Auf Differenzen innerhalb der untersuchten Kohorte wird nicht eingegangen. Das Kriterium wurde überwiegend erfüllt und wird als eher positiv gewertet.

Praktische Verfahren

Zur Gruppenaufteilung werden keine weiteren Angaben gemacht. Der Kurs „Math 155“ wurde in drei Darbietungsformen angeboten: als traditioneller Vorlesungskurs, als Telekurs und als webbasierter Kurs. Es ist anzunehmen, dass die Studierenden jeweils einer Kursart zugeordnet waren. Wie diese Zuordnung vorgenommen wurde, ist nicht bekannt.

Das Kriterium wurde somit kaum erfüllt und wird mit eher negativ bewertet.

⁸⁰ Vortrag im Rahmen der „League for Innovation in the Community College Innovations Conference“ (Boston, MA, March 19-20, 2002).

⁸¹ Als GPA (Grade Point Average) wird im US-amerikanischen Bildungssystem eine Art Notendurchschnitt bezeichnet. Dabei wird jeder einzelnen Note ein Punktwert zugeordnet. Aus diesen wird anschließend ein gewichtetes arithmetisches Mittel berechnet (vgl. http://de.wikipedia.org/wiki/Grade_Point_Average, letzter Aufruf: 25.08. 2006).

8.3.7.2 Genauigkeit

Valide und reliable Informationen

Wie bereits erwähnt, wurde die Untersuchung an denjenigen Studierenden durchgeführt, die 1999/2000 im Kurs „Math 155“ eingeschrieben waren und einen Vortest sowie das Abschlussexamen erfolgreich absolviert hatten. 73% dieser Studierenden schlossen darüber hinaus einen ergänzenden Einstellungstest ab. Dieser beinhaltete Fragen zur Einschätzung ihrer Erfahrungen:

„Section 2 in the attitudinal survey asked the students to identify their perception of their level of technological sophistication [...]. Section 3 in the attitudinal survey asked the students to evaluate the course in terms of mode of delivery and presentation concerns. [...] Section 4 in the attitudinal survey asked the students to identify their perception of the resources available to them ranging from library resources to student services“ (vgl. William J. Ryan (2002:9).

Es werden keine Aussagen zu Validität und Reliabilität getroffen, obwohl davon auszugehen ist, dass diese Werte bei der Untersuchung nicht vernachlässigt worden sind.

Das Kriterium wurde somit kaum erfüllt und wird mit eher negativ bewertet.

Begründete Schlussfolgerungen

Die Ergebnisse der Evaluation werden beschrieben. Dabei ist festzustellen, dass keine der erwähnten Forschungsfragen signifikante Ergebnisse lieferte. In der Leistung der Studierenden gab es keine Unterschiede in Bezug auf die Darbietungsmodi. Der Autor schließt daraus, dass es den Studierenden nichts ausmachte, mit alternativen Methoden unterrichtet zu werden.

„This research indicates that there is also no harm to the learner, as measured in final course grade, by institutions using alternative instructional delivery systems. The feedback obtained in the attitudinal survey suggests that alternative delivery systems such as telecourses and web-based instruction may work to the student's advantage in terms of time and access gained. Technology offers the student the ability to shift time and increase access to instructional materials compared to traditional lecture-based delivery“ (a.a.O.:8).

Es sind darüber hinaus keine signifikanten Unterschiede in Bezug auf Alter und Lernleistung festgestellt worden. Ebenfalls führte die Messung von Studienleistungen in Bezug auf die Einstellung zu webbasiertem Lernen zu keinen signifikanten Ergebnissen.

„There was no significant difference in the interaction of the delivery methods with the pre-test used as a covariate on the final course grade with traditional and non-traditional age students. The results showed no significant difference among the three instructional delivery methods on the overall final course grade (dependent variable) when comparing traditional age students to non-traditional age students. The results also indicated no significant statistical difference in student achievement with regard

to gender. Therefore, it can be concluded that for this research project there is no significant statistical difference in student achievement, as measured by overall final course grade, using the three different instructional delivery systems when comparing traditional and non-traditional age students regardless of gender” (a.a.O.:9).

Studierende in einem webbasierten Kurs oder einem Telekurs erbrachten also keine schlechteren Leistungen als in einem traditionellen Kurs. Auch Alter und Geschlecht übten keinen Einfluss auf die Abschlussnote aus. Die Zufriedenheit der Studierenden mit der Darbietungsmethode hatte ebenfalls keine Auswirkungen auf das Lernergebnis, ebenso wenig wie Zeit- und Ortsabhängigkeit. Es wurde überdies festgestellt, dass äußere oder mentale Hindernisse ebenso wie technische Hindernisse keine Effekte hatten.

Der Autor resümiert, dass Institutionen auf Grund dieser Ergebnisse in ihren Bestrebungen unterstützt würden, alternative Lernmethoden einzusetzen und Studierende auf erfolgreiche Fernstudien Erfahrungen vorzubereiten. Eigene Kritik an der Studie äußert der Autor nicht.

Das Kriterium ist überwiegend erfüllt, da sich die Schlussfolgerungen auf die Daten beziehen und begründet werden. Es wird mit eher positiv bewertet.

8.3.7.3 Zuverlässigkeit

Angemessenheit der Forschungsentscheidungen

Da angenommen wurde, dass es in Bezug auf die erbrachte Punktzahl nach Abschluss des Kurses keine Unterschiede zwischen den Darbietungsmodi geben würde, wurden verschiedene Forschungsfragen geprüft:

- Gibt es signifikante Unterschiede in der durch die Abschlussarbeit gemessenen Leistung der Studierenden bezogen auf die Instruktionsmethoden?
- Wird das Lernergebnis von Geschlecht und Alter der Lernenden beeinflusst? Spielt der Darbietungsmodus eine Rolle?
- Beeinflusst die Zufriedenheit der Studierenden mit der Darbietungsmethode das Lernergebnis?
- Werden die Lernergebnisse je nach Darbietungsmethode von Reise- und Zeitoptionen beeinflusst?⁸²

⁸² Hintergrund ist die größere Unabhängigkeit von Ort und Zeit bei webbasierten Kursen oder Telekursen.

- Werden die Lernergebnisse von äußerlichen, mentalen, umfeldbedingten oder technischen Hindernissen bezüglich ihrer Leistung beeinflusst?

Die Wahl der Stichprobe wird nicht begründet, ebenfalls werden keine Begründungen zur Wahl der Auswertungsmethode geliefert.

Über das Studienmaterial ist zu wenig bekannt, als dass hier detaillierte Aussagen getroffen werden könnten. Dennoch bleibt festzustellen, dass beim Vergleich der Gruppen die unabhängigen Variablen konstant gehalten wurden, indem alle Gruppen mit demselben Test geprüft wurden. Die Spezifika der alternativen Darbietungsmethoden wurden also nicht genutzt. Dennoch beziehen sich die Interpretation der Ergebnisse auf diese Variablen.

Das Kriterium wird mit neutral bewertet, da es teilweise erfüllt wurde.

8.3.7.4 Bezug zu den Forschungsfragen

Lernerfolgsmessung

Die Messung des Lernerfolgs orientierte sich an Hand der abhängigen Variable der Noten des Abschlusstests. Es wurde also keine Kompetenzmessung betrieben, auch der Lernprozess selbst blieb im Hintergrund – es zählte einzig der Output an Hand von Testnoten.

Bei der Auswertung des Einstellungstests wurde die Lernleistung unter anderem an Hand eigener Einschätzungen der Studierenden bewertet.

Es werden keine Lernziele im Bericht erwähnt. Auswirkungen von Lerneffekten zeigten sich nur durch den Vergleich des Pretests mit den Abschlussnoten. In allen Untersuchungsgruppen wurde der gleiche Test durchgeführt. Dies legt den Schluss nahe, dass in beiden Fällen nur Wissen auf der Stufe des Erinnerns abgefragt wurde⁸³.

⁸³ Allerdings kann diese Einordnung nur vermutet werden, da das Kursmaterial nicht spezifisch erläutert wird.

DIE KOGNITIVE PROZESS-DIMENSION					
1. Erinnern	2. Verstehen	3. Anwenden	4. Analysieren	5. Bewerten	6. (Er)schaffen
Test aller Gruppen					

Tabelle 16: Einordnung von P 7 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)

Allerdings hätten hier unterschiedliche Prüfungsformen eingesetzt werden müssen, um auch die Spezifika der neuen Medien zu untersuchen.

Das Kriterium wird kaum erfüllt und damit als eher negativ gewertet.

Signifikante Ergebnisse

Wie bereits dargestellt, wurden die Unterschiede in den Testgruppen auf Signifikanz getestet. Es konnten jedoch keine signifikanten Differenzen gefunden werden.

Das Kriterium wird mit positiv bewertet.

8.3.7.5 Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Durchführbarkeit						
1.1	Angemessene Verfahren		x			
1.2	Praktische Verfahren				x	
2. Genauigkeit						
2.1	Valide und reliable Informationen				x	
2.2	Begründete Schlussfolgerungen		x			
3. Zuverlässigkeit						
3.1	Angemessenheit der Forschungsentscheidungen			x		
4. Bezug zu Forschungsfragen						
4.1	Lernerfolgsmessung				x	
4.2	Signifikante Ergebnisse	x				
	Gesamtbewertung	1	2	1	3	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 17: Bewertungsraster P7

8.3.8 P 8: Thommy Eriksson, Alois Goller & Sergei Muchin (2001)

Die vorliegende Studie wurde 2001 an Chalmers Tekniska Högskola in Göteborg, Schweden durchgeführt. In erster Linie ging es dabei um die Untersuchung der Online-Kommunikation, dennoch werden auch Angaben zu den Lernergebnissen gemacht, weshalb diese Evaluation als Primärstudie berücksichtigt wird.

Es wurde ein Kurs evaluiert, der in zwei Versionen verfügbar war – als konventioneller Kurs mit Workshops und Vorlesungen sowie als reiner Online-Kurs mit Kursarbeiten, Übungen und Newsgroups.

Der Kurs wurde in Teilzeit angeboten und hatte das Thema „Planning and design of web pages“, weshalb eine gewisse Affinität zum Lernen mit dem Internet vorausgesetzt werden konnte.

8.3.8.1 Durchführbarkeit

Angemessene Verfahren

Der Rahmen der Evaluation wird erläutert. Das Verhältnis des Aufwandes zum Nutzen der Evaluation schien angemessen zu sein, wenngleich in der vorliegenden Publikation nur sehr wenig über die Untersuchungsmethoden ausgesagt wird.

Auf Differenzen innerhalb der Zielgruppe wird insofern eingegangen, als berichtet wird, dass im Zuge der Seminararbeit eine Diskussion in Bezug auf Vorurteile stattgefunden habe:

„It was obvious during the interviews that the question about online student presentation was a bit controversial. During one semester we published portraits and personal presentation of the students on the course web site. Some students and several of the teachers had concerns about having student presentations. It was considered uninteresting, not relevant, and that it could be sensitive for some students to publish portraits. Other researchers have concluded that student presentations could give room for prejudices about gender, race, social status etc“ (Thommy Eriksson, Alois Goller & Sergei Muchin 2001:2).

Im Rahmen der Evaluation wurden kulturelle, soziale oder Genderaspekte jedoch nicht berücksichtigt.

Das Kriterium wird überwiegend erfüllt und wird mit eher positiv bewertet.

Praktische Verfahren

Die Untersuchung wurde mit zwei verschiedenen Lerngruppen durchgeführt. Während eine Gruppe von Studierenden den evaluierten Kurs als traditionelles Seminar mit Workshop- und Vorlesungsanteilen sowie Web-Unterstützung absolvierte, arbeitete die andere Gruppe komplett online.

„To investigate this we made a comparison and evaluation of two versions of the same short course. The course "Planning and design of web sites" is given at Chalmers University of Technology, Chalmers Lindholmen University College, both as a conventional course (ordinary lectures and workshops at location with online support) and a distance course (fully online). All students use the same web site, which have functionality such as coursework, exercises, newsgroup, chat, etc.“ (a.a.O.:1).

Wie die Zusammensetzung der Gruppen genau vollzogen wurde, wird nicht näher erläutert.

Das Kriterium wurde teilweise erfüllt und somit neutral bewertet.

8.3.8.2 Genauigkeit

Valide und reliable Informationen

In der Untersuchung wurden die Studierenden überwiegend zu ihrer Kommunikation befragt. Es wurde zunächst ein Review der relevanten Literatur durchgeführt. Kursevaluationen fanden in jedem Semester statt⁸⁴. Darüber hinaus wurden im Wintersemester 2000 5 Lehrpersonen und 7 Studierende interviewt. Auch Transkripte, das Webportal und die Ergebnisse von schriftlichen Übungen wurden analysiert. Die Kursevaluationen basierten dabei auf folgenden Fragestellungen:

- „A Are you satisfied with how much you have learned?
- B What's your general opinion about the course?
- C What's your opinion about teacher competence?
- D What's your opinion about web portal functionality?
- E What's your opinion about web portal content?
- F What's your opinion about communication with the teachers?
- G What's your opinion about the use of the newsgroup?
- H What's your opinion about peer-to-peer communication?
- I What's your opinion about teacher communication in the newsgroup?“ (a.a.O.:4).

Es werden keine Aussagen zu Validität und Reliabilität getroffen, obwohl davon auszugehen ist, dass diese Werte bei der Untersuchung nicht vernachlässigt worden sind. Auch über die Verfahren der Datengewinnung und -analyse ist nichts bekannt.

Das Kriterium wurde somit kaum erfüllt und wird mit eher negativ bewertet.

⁸⁴ In die Untersuchung fließen jedoch nur Ergebnisse aus dem Wintersemester 2000 ein.

Begründete Schlussfolgerungen

Neben den Ergebnissen, die zum Aspekt der Kommunikation vorgestellt werden⁸⁵, gibt es auch Resultate, die den Lernerfolg betreffen.

Demnach waren die Studierenden im Online-Kurs zufriedener mit den eigenen Lernergebnissen und der Funktionalität der Webseite. Daraus wird der Schluss gezogen, dass sich diese Studierenden besser an das Lernen mit einer Lernumgebung angepasst hätten:

„We have also found that the distance course students were more satisfied with what they had learned and the web site functionality. We thus concluded that the students in the fully online education adapted better to learning in the digital environment“ (ebd.).

Die Autoren äußern ihre Verwunderung darüber, dass diese Aussagen getroffen worden seien, obwohl der Online-Kurs mit mehr technischen Problemen konfrontiert gewesen sei als der Präsenzkurs:

„Could it be that the distance education students was more engaged in their own learning and had a more mature attitude towards their own learning and their own responsibility for it?“ (ebd).

Unterschiede zwischen den Lernergebnissen, die in den Übungen erhoben worden sind, wurden nicht beobachtet. Es gab keine Hinweise darauf, dass die Online-Studierenden mehr Schwierigkeiten hatten, die Übungen zu verstehen oder abzuschließen.

Die wichtigste Schlussfolgerung der Studie ist die Erkenntnis, dass die gefundenen Unterschiede zwischen den Gruppen – wenn überhaupt vorhanden – nur sehr klein gewesen seien.

Die Ergebnisse werden dargestellt. Es wird darüber hinaus erläutert, welche Daten diese Schlussfolgerungen zulassen – allerdings sind diese relativ vage. Es fehlen genauere Angaben, welches Ergebnis auf Grund von Interviews und welches auf Grund der Überprüfung der Lernresultate zu Stande kommt. Eigene Kritik an der Evaluation äußern die Autoren nicht.

Das Kriterium wurde teilweise erfüllt und wird mit neutral bewertet.

8.3.8.3 Zuverlässigkeit

Angemessenheit der Forschungsentscheidungen

Es ist nicht ersichtlich, ob im Vorfeld der Untersuchung Hypothesen aufgestellt wurden. Die Wahl der Stichprobe wird nicht begründet, ebenfalls werden keine Begründungen zur Wahl der Auswertungsmethode geliefert.

⁸⁵ Der wichtigste Vorteil der computervermittelten Kommunikation sind der Zeitaspekt im Allgemeinen und die Asynchronität im Speziellen. Dies wird sowohl von Lehrenden als auch von Lernenden kommuniziert. Die Präsenzgruppe war mit der Kommunikation zufriedener als die Online-Gruppe.

Über das Studienmaterial ist zu wenig bekannt, als dass hier detaillierte Aussagen getroffen werden können. Dennoch bleibt festzustellen, dass beim Vergleich der Gruppen die unabhängigen Variablen konstant gehalten wurden, indem alle Gruppen die gleichen Übungen zu absolvieren hatten. Die Spezifika der alternativen Darbietungsmethoden wurden zwar genutzt, jedoch nicht spezifisch evaluiert. Dennoch beziehen sich die Interpretationen der Ergebnisse auf diese Variablen.

Darüber hinaus ist es unklar, warum die Studierenden des Online-Kurses gefragt wurden, was sie im Vergleich mit einem Präsenzkurs vermisst hätten. Dies führte zu dem Ergebnis, dass vor allem das Fehlen von Vorlesungen („lectures“) als Nachteil empfunden wurde. Die Studierenden waren jedoch entweder der Online- oder der Präsenzgruppe zugeordnet. Direkte Vergleiche mit der jeweils anderen Gruppe waren ihnen in ihrem derzeitigen Kontext gar nicht möglich. Daher muss sich die Frage des Vermissens auf die Erfahrungen beziehen, die die Studierenden im Laufe ihres Studiums belegt hatten. Da diese sehr unterschiedlich sein können, ist die Frage an dieser Stelle als wenig glücklich anzusehen.

Das Kriterium wird mit eher negativ bewertet, da es kaum erfüllt wurde.

8.3.8.4 Bezug zu den Forschungsfragen

Lernerfolgsmessung

Es werden keine expliziten Aussagen über die Messung des Lernerfolgs getroffen, daher kann an dieser Stelle nur aus Angaben im Text geschlossen werden. So wird beispielsweise resümiert, dass es Unterschiede in der Zufriedenheit mit dem Kurs gegeben hat.

Ein anderer Indikator für den Lernerfolg schien in der vorliegenden Studie in der Abbruchquote gesehen zu werden. Die Autoren geben an, dass diese in beiden Kursen gleich hoch gewesen sei:

„The drop-out rate in the fall of 2000 was very similar between the conventional education (25%) and the distance education students (26%), showing that students performed similar in both versions of the course. The drop-outs in 1999 differed much more and was much higher (33% respectively 46%)“ (ebd.).

Die Übungen, die zur Messung des Lernerfolgs zu Rate gezogen wurden, waren in beiden Gruppen gleich. Sie werden nicht genauer spezifiziert, legen offenbar jedoch die gleiche Ausprägung von Lernerfolg zu Grunde. Die Art des Unterrichts wich von diesem Prinzip stark ab. Während der Präsenzkurs Vorlesungen und Workshops anbot, in denen das Gelernte reflektiert werden konnte, gab es im Online-Kurs diese Möglichkeit nicht. Dies zeigte sich allein an der eigentlich untersuchten Frage der Kommunikation, die im Online-Kurs weniger zufriedenstellend abgelaufen war.

Bei der Einordnung der Gruppen in die Prozess-Dimension der Taxonomie nach Lorin W. Anderson und David R. Krathwohl (2001) wird

der Präsenzkurs dementsprechend auf der Ebene des Verstehens platziert, der Online-Kurs lediglich auf der Ebene des Erinnerns.

DIE KOGNITIVE PROZESS-DIMENSION					
1. Erinnern	2. Verstehen	3. Anwenden	4. Analysieren	5. Bewerten	6. (Er)schaffen
Lernziel Präsenz Test beider Gruppen	Lernziel Online				

Tabelle 18: Einordnung von P 8 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)

Dass beide Gruppen die gleichen Übungen absolvierten, trägt an dieser Stelle also nicht zur Messung des Lernerfolgs bei.

Das Kriterium wurde kaum erfüllt und wird mit eher negativ bewertet.

Signifikante Ergebnisse

Die Ergebnisse wurden auf Signifikanz untersucht. Dabei wurden keine signifikanten Unterschiede zwischen den Gruppen festgestellt:

„No significant difference in student performance between conventional and distance education“ (ebd.).

Das Kriterium wurde erfüllt und wird als positiv bewertet.

8.3.8.5 Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Durchführbarkeit						
1.1	Angemessene Verfahren		x			
1.2	Praktische Verfahren			x		
2. Genauigkeit						
2.1	Valide und reliable Informationen				x	
2.2	Begründete Schlussfolgerungen			x		
3. Zuverlässigkeit						
3.1	Angemessenheit der Forschungsentscheidungen				x	
4. Bezug zu Forschungsfragen						
4.1	Lernerfolgsmessung				x	
4.2	Signifikante Ergebnisse	x				
	Gesamtbewertung	1	1	2	3	0
*positiv		# eher positiv		+ neutral		0 eher negativ - negativ

Tabelle 19: Bewertungsraster P8

8.3.9 P 9: Ruth Green & Karen Gentemann (2001)

Die vorliegende Untersuchung wurde im Wintersemester 1999/2000 an der George Mason University im US-Bundesstaat Virginia durchgeführt. Dabei ging es den Autorinnen in erster Linie darum, die Unklarheit der unterschiedlichen Schlussfolgerungen, die in der Literatur in Bezug auf die Effektivität von Online vs. Präsenzlernen gezogen wurden, durch weitere Forschung aufzulösen.

„It seems clear that the assessment of OLC⁸⁶ is complex and that findings can vary based on questions asked, level and type of course, level and type of student, student motivations and expectations, technology used, outcomes examined, etc.“ (Ruth Green & Karen Genteman 2001:2f.).

Die Autorinnen legen eine Längsschnittstudie vor, in der die Unterschiede zwischen Studierenden in einem Online- und einem Face-to-Face-Kurs zu zwei Zeitpunkten erhoben wurden. Dabei wurden sowohl die Einstellungen zum Kurs als auch die Leistungen der Lernenden betrachtet.

8.3.9.1 Durchführbarkeit

Angemessene Verfahren

Der Rahmen der Evaluation wird erläutert. Das Verhältnis des Aufwandes zum Nutzen der Evaluation schien angemessen zu sein.

Auf kulturelle, soziale oder Genderaspekte wird explizit eingegangen – diese Faktoren werden bei der Auswertung berücksichtigt:

„Finally, significance tests were conducted for course completion, course grade, gender, and race for students in the three OLC sections vs. students in the 12 FTF⁸⁷ sections of the English course“ (a.a.O.:5).

Das Kriterium wird voll erfüllt und mit positiv bewertet.

Praktische Verfahren

Als Untersuchungsgruppe diente der Kurs „English 302: Advanced Composition“, der als Online-Kurs in drei begrenzten Präsenzeinheiten unterrichtet wurde:

„The three sections were OLC with limited face-to-face contact. One class met only at the beginning of the course and two classes met at the beginning and again at the end“ (a.a.O.:3).

An diesem Kurs nahmen 57 Studierende teil, von denen 47 während des gesamten Semesters im Kurs verblieben. 52 Studierende beteiligten sich zu Beginn des Semesters an der Untersuchung, 30 Studierende am Ende des Kurses.

⁸⁶ OnLineCourse (Anm. AP)

⁸⁷ Face-To-Face (Anm. AP)

Die Kontrollgruppe bildeten 12 Einheiten des gleichen Kurses, der als traditioneller Präsenzkurs abgehalten wurde:

„In contrast, there were 12 sections of English 302 that were traditional FTF courses that met at different times during the week for varying periods of time as shown below“ (ebd.).

Über die Anzahl der Studierenden im Präsenzkurs sowie über die Verfahren der Gruppenzusammensetzung werden keine Angaben gemacht.

Das Kriterium wird mit neutral bewertet, da es teilweise erfüllt wurde.

8.3.9.2 Genauigkeit

Valide und reliable Informationen

Die Verfahren der Datengewinnung und -analyse werden nachvollziehbar beschrieben.

Die Studierenden wurden mittels Fragebogen zur Einschätzung der Instruktionsmethode und zum Vergleich der Leistungen getestet. Dabei beinhaltete der Fragebogen diverse Skalen (Importance of academic/social interactions, Expertise with technology, Student expectations for instructor priorities, Use of Seven Principles of Good Practice, Relationships and contacts with other students, Course characteristics, Course procedures, Course feedback, Benefits of OLC, and Benefits of FTF), von denen im Forschungsbericht jeweils die Reliabilität (Cronbach's Alpha) angegeben waren.

Insofern orientierte sich die Evaluation an Gütekriterien. Auf die Schwächen der Skalen wird nicht näher eingegangen (dieser Aspekt wird unter „Begründete Schlussfolgerungen“ näher thematisiert).

Das Kriterium wird mit eher positiv bewertet.

Begründete Schlussfolgerungen

Die Ergebnisse der Evaluation werden ausführlich und übersichtlich dargestellt. Dabei werden zwar Angaben darüber gemacht, welche Daten diese Schlussfolgerungen zuließen, jedoch wurden auch Aussagen über Skalen getroffen, deren Werte im nicht-zulässigen Bereich lagen:

Es ist verwunderlich, dass auch die Skalen beibehalten wurden, die eine niedrige Reliabilität aufwiesen. Beispielsweise hatte die Skala „Student expectations for instructor priorities“ (3 Items) nur einen Alpha-Wert von .41⁸⁸. Die Skala „Course characteristics“ (3 Items) wies einen Wert von $\alpha = .59$ auf, „Course procedures“ (2 Items) nur einen Wert von

⁸⁸ Der Alpha-Wert ist ein Maß für die interne Konsistenz der Skala und beschreibt damit deren Zuverlässigkeit. Der Alpha-Koeffizient weist Werte zwischen 0 und 1 auf. „In bestimmten Kontexten gelten Skalen mit Alphawerten zwischen 0,7 und 0,8 bereits als brauchbar, je höher der Alphawert wird, desto 'besser' ist die interne Konsistenz der Skala. Werte zwischen 0,9 und 0,99 gelten als gut“ (<http://psychologie.fernuni-hagen.de/Sem3/glossar.html>).

$\forall = .57$. Diese Werte sind auf Grund der geringen Itemanzahl nicht überraschend, jedoch hätte die Untersuchung in der Folge berücksichtigen müssen, dass diese Skalen nicht brauchbar sind und im Grunde keine Aussagekraft haben.

Eigene Kritik äußern die Autorinnen insofern, als sie feststellen:

„There is an absence of validated scales (and instruments) for many of the outcomes of interest for technology-based education --- student expectations for OLC, student perceptions of expertise with technology, benefits of OLC, etc. While the Flashlight Inventory is helpful because it provides a large number of individual items, there is a need for validated scales that facilitate the consistent measurement of important constructs/outcomes across studies. The absence of relevant validated scales increases the variability among operational definitions used in different studies and thus, lessens our ability to compare and understand the findings across different studies“ (a.a.O.:16).

Das Kriterium wird mit neutral gewertet, da sowohl positive als auch negative Aspekte festgestellt werden konnten.

8.3.9.3 Zuverlässigkeit

Angemessenheit der Forschungsentscheidungen

Es sind im Vorfeld der Untersuchung keine Hypothesen aufgestellt worden. Auch wurden keine Begründungen zur Auswahl der Stichprobe genannt.

Die Auswertungsmethode wird teilweise begründet. So geben die Autorinnen an, dass die Fragebögen keine Identifizierungscodes erhalten hätten, um ehrliche Antworten von den Studierenden zu bekommen. Aus diesen Gründen sei die Vergleichbarkeit im Längsschnitt nur auf Kursebene möglich.

Diese Aussage erscheint nicht vollständig nachvollziehbar, da es in den Sozialwissenschaften durchaus Werkzeuge gibt, wie Fragebögen eindeutig Personen zugeordnet werden können, ohne die Anonymität außer Kraft zu setzen⁸⁹.

Als problematisch wird generell der Vergleich der Gruppen erachtet: Es stellt sich die Frage, inwiefern sich der Unterricht tatsächlich miteinander vergleichen lässt. So wurde traditionelles Lernen im Präsenzunterricht (intensives Schreibtraining und Übungen zur Analyse von Essays) mit selbständigem Lernen mittels Web-CT verglichen, in welchem es vor allem um die Aufgabenbearbeitung mittels Fallstudien ging. Diese beiden Lernformen legen also einen anderen Lernbegriff zu Grunde.

⁸⁹ Bewährt hat sich beispielsweise eine Codierung, die sich aus diesen Angaben zusammensetzt: Dritter Buchstabe des eigenen Vornamens, Dritter Buchstabe des Vornamens der Mutter, Dritter Buchstabe des Nachnamens des Vaters, Zweiter Buchstabe des Vornamens der Mutter sowie Zweiter Buchstabe des Vornamens des Vaters.

Die genaue Prüfungsform ist hierbei nicht bekannt, jedoch werden im vorliegenden Bericht die Abschlussnoten miteinander verglichen. Hier ist die unterschiedliche Anzahl der Gruppen zu kritisieren, die keinen verallgemeinerbaren Vergleich zulässt (Online: 3 Kurse, Präsenz 12 Kurse).

Die Antworten der Fragebögen zur Einschätzung der Instruktionsmethode wurden vor allem in Bezug auf die Angaben zu Beginn bzw. zum Ende des Kurses untersucht und verglichen, weniger in Bezug auf die Unterschiede zwischen Online- und Präsenzlernen.

Das Kriterium wird mit eher negativ bewertet.

8.3.9.4 Bezug zu den Forschungsfragen

Lernerfolgsmessung

Es werden keine expliziten Aussagen über die Messung des Lernerfolgs und die entsprechenden Prüfungsformen getroffen, daher kann an dieser Stelle nur aus Angaben im Text geschlossen werden. Ein Indikator schien die Abbruchquote gewesen zu sein. Die Autorinnen geben an, dass diese in beiden Kursen gleich hoch gewesen sei. Überdies wurde die abhängige Variable der Noten des Abschlusstests (GPA) einbezogen. Es wurde also keine Kompetenzmessung betrieben, auch der Lernprozess selbst blieb im Hintergrund. Allerdings wurden Fragebögen zur Einschätzung der Instruktionsmethode eingesetzt.

Obwohl keine expliziten Lernziele benannt wurden, lässt sich aus der Beschreibung der Kurse darauf schließen, dass im Präsenzkurs („intensive practice in writing and analyzing such expository forms as the essay, article, proposal, and technical and scientific report“ (a.a.O.:3f.) vor allem Lernziele auf der Ebene des Analysierens und Erschaffens verfolgt wurden. Über den Onlinekurs wird berichtet:

„Assignments varied but were typically scenario/case studies that required students to provide on-line, research-based responses using the university data bases and other materials identified on the website. For example, one assignment provided students a mock memo with an identified problem from a work supervisor. Students were asked to respond via electronic memo with a research-based response and recommendations for the supervisor“ (ebd.).

Dies lässt darauf schließen, dass Lernziele vor allem in Bezug auf die kognitive Wissensdimension „Bewerten“ verfolgt wurden.

DIE KOGNITIVE PROZESS-DIMENSION					
1. Erinnern	2. Verstehen	3. Anwenden	4. Analysieren	5. Bewerten	6. (Er)schaffen
	Test beider Gruppen ⁹⁰		Lernziel I Präsenz	Lernziel Online	Lernziel II Präsenz

Tabelle 20: Einordnung von P 9 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)

Die Lernziele lassen sich also nicht miteinander im selben Abschlusstest vergleichen.

Das Kriterium wird mit eher negativ bewertet.

Signifikante Ergebnisse

Die Ergebnisse wurden auf Signifikanz untersucht. Dabei wurden in Bezug auf den Lernerfolg keine signifikanten Unterschiede zwischen den Gruppen festgestellt:

„There were no significant differences between students in the OLC compared to students in the FTF courses for course retention“ (a.a.O.:15).

Das Kriterium wurde erfüllt und wird als positiv bewertet.

⁹⁰ Es wird aus der Beschreibung der Studie nicht deutlich, welche Art von Test eingesetzt worden ist. Möglicherweise ist dieser auch auf einer anderen Prozessebene einzuordnen. Deutlich wird dennoch, dass ein Vergleich der Gruppen mit nur einem Test nicht sinnvoll ist.

8.3.9.5 Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Durchführbarkeit						
1.1	Angemessene Verfahren	x				
1.2	Praktische Verfahren			x		
2. Genauigkeit						
2.1	Valide und reliable Informationen		x			
2.2	Begründete Schlussfolgerungen			x		
3. Zuverlässigkeit						
3.1	Angemessenheit der Forschungsentscheidungen				x	
4. Bezug zu Forschungsfragen						
4.1	Lernerfolgsmessung				x	
4.2	Signifikante Ergebnisse	x				
	Gesamtbewertung	2	1	2	2	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 21: Bewertungsraster P9

8.3.10 P 10: David J. Pucel & Thomas F. Stertz (2005)

In dieser Studie wurde die Zufriedenheit der Studierenden mit zwei Online-Kursen und deren Leistung im Vergleich zu zwei gleichwertigen Präsenzkursen untersucht. Zielgruppe waren Lehramtsstudierende der University of Minnesota („in-service career and technical education teachers who were preparing to teach occupational programs at the high school or technical college level“ (David J. Pucel & Thomas F. Stertz 2005:11)).

Dabei wurde das Ziel verfolgt, ein generalisierbares Modell zur Evaluation von Online-Kursen⁹¹ im Vergleich zu traditionellen Kursen zu entwickeln und gleichzeitig festzustellen, ob diese gleiche Leistungen der Studierenden hervorbringen würden.

8.3.10.1 Durchführbarkeit

Angemessene Verfahren

Der Rahmen der Untersuchung wird erläutert. Dabei scheint das Verhältnis des Aufwandes zum Nutzen der Evaluation angemessen gewesen zu sein. Auf Differenzen innerhalb der untersuchten Kohorte wird allerdings nicht eingegangen.

Das Kriterium wurde überwiegend erfüllt und wird als eher positiv gewertet.

Praktische Verfahren

Über die Zusammensetzung von Test- und Kontrollgruppen werden keine Angaben gemacht. Es wird lediglich darüber berichtet, dass zwei Online- und zwei Präsenzkurse miteinander verglichen worden seien, von denen jeweils die Online- und Präsenzversion dasselbe Kursmaterial und denselben Kursinhalt verwendet hätten. Der Präsenzkurs in Philosophie wurde bereits während des Wintersemesters 2002 untersucht, der Präsenzkurs in Methoden im Sommer 2001. Die Online-Kurse fanden im Jahr 2003 statt. Es ist also anzunehmen, dass sich die Zusammensetzung der Kurse aus den Einschreibungen bzw. Belegungen ergeben hatte:

„[...] the students who took the earlier, traditional versions of the courses were similar to those who took the later, Web-based enhanced versions“ (a.a.O.:14).

Das Kriterium wird teilweise erfüllt und mit neutral bewertet.

8.3.10.2 Genauigkeit

Valide und reliable Informationen

Die Untersuchung basiert vorwiegend auf Test- oder Abschlussnoten, die im Laufe des Semesters von den Studierenden

⁹¹ Genauer: „WBI-courses“ – Kurse, die auf Web Based Instruction aufbauen.

erbracht wurden. Die Verfahren der Datengewinnung werden damit zwar beschrieben, aber nicht begründet.

„Grading information included student project scores, final test scores, and final grades gathered during each WBI course“ (ebd.).

Dabei wurden in jedem Kurs vier Erhebungen durchgeführt und zwar in Form von zwei Projekten (s. Abschnitt Lernerfolgsmessung), einem Abschlusstest und einer Gesamtnote. Darüber hinaus wurde die Zufriedenheit der Studierenden mit dem Kurs durch den universitären Fragebogen zur Kursevaluation erhoben. Als Vergleichsdaten wurden Ergebnisse früherer Jahrgänge in Präsenzkursen herangezogen. Alle Kursarten wurden mit dem gleichen Instrument bewertet, das im Wesentlichen aus sechs Dimensionen bestand:

- „1. Instructor’s overall ability?
2. Instructor’s knowledge of the subject matter?
3. How much would you say you learned from this course?
4. Overall quality of tests and handouts?
5. Helpfulness of feedback to you about your performance?
6. Degree to which evaluation procedures measured your knowledge and understanding?“ (a.a.O.:18)

Zusätzlich zu diesen standardisierten Verfahren wurden die Studierenden im Online-Kurs aufgefordert, ihre Zufriedenheit zu beschreiben, anzugeben, wieviel Zeit sie dafür verwendet hätten und die Schwierigkeit des Kurses im Vergleich mit ihren Erfahrungen im Präsenzkurs anzugeben.

Die Untersuchung wurde an insgesamt 84 Studierenden in (44 in Online-Kursen und 40 im Präsenzunterricht) durchgeführt. Bei der Datenanalyse wurden die üblichen statistischen Verfahren angewendet.

„The means of the percentage scores for each of the four assessments in both the WBI classes and comparable traditional classes were calculated, and the t-test for independent samples was applied to determine if the differences in the means of the two groups were statistically significant“ (a.a.O.:19).

Angaben zu Gütekriterien wurden nicht gemacht.

Das Kriterium wurde teilweise erfüllt und wird mit neutral bewertet.

Begründete Schlussfolgerungen

Bevor auf die Ergebnisse der Evaluation eingegangen wird, soll die Kritik in den Blick genommen werden, die die Autoren selbst an ihrer Untersuchung äußern. Diese bezieht sich auf die Vergleichbarkeit der Versuchsgruppen:

„A major limitation of this study is the assumption that the students who took the earlier, traditional versions of the courses were similar to those who took the later, Web-based enhanced versions. However, this assumption seemed reasonable since all of the participants in the study had been hired to teach in high schools or post secondary career and technical education programs and were required to take the courses as part of their continuing employment. There seems no reason to believe the two groups were significantly different“ (a.a.O.:14).

Darüber hinaus wird angemerkt, dass zwar die Lehrenden, die Übungen und die Bewertungskriterien die gleichen gewesen seien, dennoch bestehe keine Gewährleistung, dass die unterschiedlichen Kursversionen vollkommen identisch seien:

„Although the instructors, assignments and grading criteria were the same, a further limitation of the study is the assumption that the two versions of each course were equivalent. As much as possible, every care was taken to ensure parity“ (ebd.).

Die Ergebnisse der Evaluation werden beschrieben. Dabei wird u.a. deutlich, dass der Zeitaufwand für die Bearbeitung des Online-Kurses von den Studierenden höher eingeschätzt wurde als für den Präsenzkurs. Im Gegensatz dazu fiel die Arbeit jedoch leichter. Insgesamt konnte eine überwiegende Zufriedenheit der Studierenden mit den Online-Kursen festgestellt werden. Die Leistung der Studierenden war im Kurs Philosophie in der Online-Version signifikant besser, im Methodenkurs in der Präsenzversion. Die Autoren geben jedoch an, dass die praktische Bedeutsamkeit dieser Ergebnisse relativ gering gewesen sei:

„Although some statistically significant differences were found on some of the student performance measures, the practical differences between the WBI and traditional versions of the courses were not large. The courses were graded using a criterion referenced system in which students were not graded against one another but rather against a set of standards. Given this system, it is possible for all students to receive very high grades or very low grades depending upon their performance“ (a.a.O.:21).

Insgesamt kann als Ergebnis der Studie festgehalten werden, dass in allen 4 Kursen gute Leistungen erbracht werden konnten.

Das Kriterium wurde voll erfüllt und wird mit positiv bewertet.

8.3.10.3 Zuverlässigkeit

Angemessenheit der Forschungsentscheidungen

Es wurden keine spezifischen Hypothesen aufgestellt, jedoch wird die Hoffnung geäußert, dass das Angebot von Online-Kursen die studentische Flexibilität in Bezug auf ihre Zeitplanung erhöhen würde:

„It was hoped that offering the classes in a WBI format would reduce the amount of travel and in-classroom time and provide flexibility for students completing course requirements without compromising either student satisfaction or performance“ (a.a.O.:11).

In Bezug auf die Erfassung des Lernerfolgs ist eine bloße Orientierung an Noten und Abschlussraten als nicht hinreichend zu bewerten, da damit nur die Performanceleistung, nicht aber der Lernprozess und noch weniger die Kompetenz bewertet werden kann.

Es wurde versucht, Rahmenbedingungen konstant zu halten, indem beide Kurse von derselben Lehrperson über die gleiche Zeit mit gleichem Inhalt durchgeführt wurden und zudem das gleiche Kursmaterial genutzt wurde. Damit ist zwar eine größere Vergleichbarkeit der Gruppen gegeben, es können jedoch keine Aussagen zu Spezifika neuer Medien gemacht werden, da diese nicht ausgeschöpft worden sind. Dem widerspricht, dass alle Studierenden mit denselben Tests geprüft worden sind.

Ein weiteres Problem sehe ich darin, dass bei der Erfassung der studentischen Zufriedenheit auf der Grundlage von Selbsteinschätzungen angegeben werden musste, inwieweit sich der besuchte Online-Kurs von traditionellen Kursen unterschied. Dabei sollten die Studierenden auf ihre vergangenen Erfahrungen mit früheren Kursen zurückgreifen. Hierbei ist die Vergleichbarkeit zwischen den Kursen meinem Verständnis zu Folge noch weniger gegeben, als wäre nur zwischen Online- und Präsenzgruppen im Untersuchungszeitraum verglichen worden. Die Autoren gehen auf diesen Umstand jedoch nicht näher ein.

Die Wahl der Stichprobe wurde erläutert, jedoch werden die Gründe für die Wahl der Auswertungsmethode nicht spezifiziert.

Das Kriterium wurde kaum erfüllt und wird mit eher negativ bewertet.

8.3.10.4 Bezug zu den Forschungsfragen

Lernerfolgsmessung

In der Untersuchung wurde der Lernerfolg an Hand der Noten der zwei Projekte, Abschlusstests und der Gesamtnote definiert. Die Ergebnisse orientierten sich daran, wieviel Prozent der Leistungspunkte angesichts der maximal zu erreichenden Punktzahl erreicht wurden.

„A student's raw score on each performance assessment was divided by the total points possible for that assessment to arrive at a percentage score“ (a.a.O.:19).

In jedem Kurs wurden zwei Projekte durchgeführt. Im Philosophie-Kurs waren die Studierenden in Planspielen aufgefordert,

1. auf einen Einsatzauftrag ihrer Institution zu reagieren und Verbesserungsvorschläge einzureichen
2. eine Umfrage zu kritischen Fragen in dem jeweiligen Bereich der Studierenden anzufertigen, diesen den anderen Studierenden darzureichen und die Ergebnisse vorzutragen.

Im Methodenkurs waren die Projekte:

1. Führen eines Lerntagebuchs zum Thema Lehrmethoden, die im Kurs angewendet wurden. Dazu sollte überlegt werden,

wie diese in einem eigenen Kurs zur Anwendung kommen könnten.

2. Abliefern einer Micro-Teaching⁹²-Präsentation unter Zuhilfenahme verschiedener Methoden aus dem Umfeld der Studierenden.

Alle Projekte fanden in der Online- und der Präsenzversion der Kurse gleichermaßen statt. Sie lassen sich in der Taxonomie der Lernziele in den Bereichen „Bewerten“ und „Erschaffen“ einordnen.

DIE KOGNITIVE PROZESS-DIMENSION					
1. Erinnern	2. Verstehen	3. Anwenden	4. Analysieren	5. Bewerten	6. (Er)schaffen
	Test beider Gruppen ⁹³			Projekte beider Gruppen	Projekte beider Gruppen

Tabelle 22: Einordnung von P 10 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)

Da nicht bekannt ist, wie die Gesamt- und Abschlussnoten erhoben wurden, fällt eine Einordnung in diesem Fall schwer. Festzuhalten bleibt jedoch:

In beiden Untersuchungsgruppen wurden die gleichen Prüfungen durchgeführt, die darüber hinaus auf dem gleichen Kursmaterial basierten und die gleichen Lernziele zu Grunde legten. Insofern sind die Gruppen zumindest auf der Ebene der Taxonomie miteinander vergleichbar, da sie auf die selbe Ausprägung von Lernerfolg abzielen. Allerdings wurde das Prüfungsdesign nicht variiert, so dass unterschiedliche Bedingungen mit gleichen Tests abgefragt wurden. Hier hätten unterschiedliche Prüfungsformen eingesetzt werden müssen, um auch die Spezifika der neuen Medien zu untersuchen.

Das Kriterium wurde teilweise erfüllt und wird mit neutral bewertet.

Signifikante Ergebnisse

Die Zufriedenheit der Studierenden wurde an Hand der bereits erwähnten standardisierten Kursevaluationen erhoben. Anschließend wurden die Mittelwerte der Antworten berechnet und es wurde ein t-Test für unabhängige Stichproben zur Signifikanzprüfung durchgeführt. Dabei wurden keine signifikanten Unterschiede zwischen den Kursen gefunden.

⁹² Lehrmethode, bei der der Unterricht auf Video aufgenommen und anschließend mit den anderen teilnehmenden Personen gemeinsam diskutiert wird. Die Lehrperson erhält auf diese Weise ein Feedback zur absolvierten Stunde.

⁹³ Es wird aus der Beschreibung der Studie nicht deutlich, welche Art von Test eingesetzt worden ist. Möglicherweise ist dieser auch auf einer anderen Prozessebene einzuordnen. Deutlich wird dennoch, dass ein Vergleich der Gruppen mit nur einem Test nicht sinnvoll ist.

Ferner wurde die Leistung der Studierenden untersucht. Diese fiel in der Online-Version des Philosophie-Kurses schlechter aus als in der Präsenzform. Hier ergaben sich signifikante Unterschiede zwischen den Kursversionen:

„In two cases the differences were statistically significant: 1) performance on the mission statement project, and 2) final grades“ (a.a.O.:19).

Dem steht jedoch gegenüber, dass diese Ergebnisse für die Methodenkurse nicht bestätigt werden konnten. Hier zeigte sich eine signifikante Überlegenheit des Präsenzunterrichts gegenüber der Online-Version.

Das Kriterium wurde voll erfüllt und wird mit positiv bewertet.

8.3.10.5 Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Durchführbarkeit						
1.1	Angemessene Verfahren		x			
1.2	Praktische Verfahren			x		
2. Genauigkeit						
2.1	Valide und reliable Informationen			x		
2.2	Begründete Schlussfolgerungen	x				
3. Zuverlässigkeit						
3.1	Angemessenheit der Forschungsentscheidungen				x	
4. Bezug zu Forschungsfragen						
4.1	Lernerfolgsmessung			x		
4.2	Signifikante Ergebnisse	x				
	Gesamtbewertung	2	1	3	1	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 23: Bewertungsraster P10

8.3.11 P11: Larry S. Tinnerman (2006)

In dieser Primärstudie wurden zwei Kurse miteinander verglichen, die sich mit Lernbehinderungen befassen. Beide wurden von derselben Lehrperson gegeben, allerdings wurde ein Kurs als Fernlehrrangebot online unterrichtet, der andere in einem traditionellen Klassenraum-Szenario. Neben dem Vergleich des Lernerfolgs und der Effektivität von Online-Lernen wurden Faktoren wie Kommunikation, Zeit, Kontinuität sowie technische Erfahrungen untersucht.

8.3.11.1 Durchführbarkeit

Angemessene Verfahren

Der Rahmen der Untersuchung wird erläutert. Dabei scheint das Verhältnis des Aufwandes zum Nutzen der Evaluation angemessen gewesen zu sein. Differenzen innerhalb der untersuchten Kohorte werden ausführlich beschrieben. So wird berichtet, dass die Studierenden in den Gruppen geschlechts- und altersgemischt waren und darüber hinaus aus verschiedenen Familienständen und sozialen Hintergründen kamen. Überdies verfügten sie über unterschiedliche Erfahrungen im Bereich der Bildung.

Das Kriterium wurde voll erfüllt und wird mit positiv bewertet.

Praktische Verfahren

An der Studie nahmen insgesamt 46 Studierende teil, die ihrer Einschreibung entsprechend in die Online- oder die traditionelle Gruppe eines Kurses zu Lernbehinderungen aufgeteilt wurden.

„Students were selected for this study using a census approach as constituted by their enrollment in the course by either traditional or distance instruction. It was this researcher's intent to get as close to total participation from all the students in both settings as possible“
Larry S. Tinnerman 2006:o.S.).

Die Gruppenbildung wird nachvollziehbar erläutert. Das Kriterium wird mit positiv bewertet.

8.3.11.2 Genauigkeit

Valide und reliable Informationen

In der Studie wurden sowohl ein experimenteller als auch ein deskriptiver Ansatz verfolgt und integriert.

„Experimental - The independent variable was the method of instructional delivery, i.e. the traditional classroom instructional approach or the distance method of instruction. Enrollment in the class, the instructor and course content were the dependant variables. Since the class was offered by the same instructor to both

groups at the same time, this provided an excellent opportunity to examine the similarities and differences between the groups in the areas of content delivery, student performance and student satisfaction" (ebd.).

Bei der deskriptiven Vorgehensweise wurden vor allem Einstellungen und Meinungen der Studierenden untersucht. Hierzu wurden Likert-Skalen und offene Fragen benutzt.

Die Verfahren der Datengewinnung und -analyse werden präzise und nachvollziehbar beschrieben. Allerdings gab es keine formalen Validitäts- und Reliabilitätsprüfungen der Instrumente.

„There were no formal measures of validity or reliability, although there has been an attempt at developing construct validity based on the fact the questions were designed to directly study the major questions being investigated" (ebd.).

Folgende Instrumente wurden eingesetzt:

- Vortest: 20 Multiple-Choice-Fragen aus der Kursdatenbank für die Kapitel 1 bis 6. Dieser Test wurde mit beiden Untersuchungsgruppen durchgeführt.
- Demographischer Fragebogen: basierend auf den Forschungsfragen.
- Student Perception Survey: ebenfalls basierend auf den Forschungsfragen.
- Nachtest: 20 Multiple-Choice-Fragen aus der Kursdatenbank für die Kapitel 1 bis 6.

Der Autor gibt an, dass Validität und Reliabilität mit diesem Instrument zwar nicht formal getestet werden konnten, da sie jedoch den gleichen Inhalt in beiden Gruppen beinhalteten, sei eine Inhaltsvalidität⁹⁴ dennoch gegeben. Darüber hinaus wurde auf weitere Störfaktoren wie den Versuchsleitereffekt eingegangen. Dieser wurde versucht, mittels objektiver Datenerhebung durch Likert-Skalen zu verringern.

Folgende Bereiche wurden u.a. im Perception Survey abgefragt:

- Zeitmanagement
 - „Perceived time spent reading class text
 - Perceived time spent reading supplemental assigned materials
 - Perceived time spent writing assignments for class
 - Perceived time spent in communication with instructor
 - Perceived time spent in class research activities

⁹⁴ „Die Inhaltsvalidität (englisch: content validity) bedeutet, dass die Gültigkeit der Messung mehr oder weniger für jedermann einsichtig aus den einzelnen Teilen des Messinstruments hervorgeht. Letztlich beruht sie auf der Kenntnis von 'Experten' über den betreffenden Gegenstand [...]. Die Behauptung, ein Messinstrument habe Inhaltsvalidität, bedeutet in der Forschungspraxis aber oft nichts anderes, als dass der Entwickler des Instruments selbst glaubt, das Instrument sei valide" (ILMES - Internet-Lexikon der Methoden der empirischen Sozialforschung: http://www.lrz-muenchen.de/~wlm/ilm_v5.htm , letzter Aufruf 08.10.2007).

- Perceived time spent in dealing with technology issues
- Perceived time spent in communication with other class members
- Perceived time spent each week in total class activities“ (ebd.)
- **Kommunikation**
 - „Accessibility of other students in your class
 - Effectiveness of student to student discussions to progress or interest in class
 - Frequency of contacting other classmates in completion of assignments
 - Other classmates having a broad scope of knowledge that helps make the course topics relevant
 - Contacting other classmates during the week in a social or causal context“ (ebd.)
- **Hilfe und Feedback**
 - „Instructor accessibility
 - Prompt and helpful feedback from the instructor in a timely manner
 - Instructional methods learn and understandable
 - Instructor sympathetic to needs and situations of a graduate student
 - Instructor provides ample opportunity for questions and the voicing of concerns and opinions“ (ebd.)
- **Kurs-Effizienz und Relevanz**
 - „Factors beyond the control of the student, instructor or university that contribute to frustration and confusion
 - Relevance of material presented in regards to personal career goals
 - Generalizability of material to other professional and or personal life endeavors
 - Material found to be thought provoking and intellectually stimulating
 - Satisfaction as to how class is progressing
 - Perception of the difficulty of participation in the on-line version of this class
 - The course content is clear and understandable“ (ebd.)

Die Befragungen wurden anonym durchgeführt – den Studierenden wurden zur Vergleichbarkeit der Daten jedoch Identitätsnummern zugeteilt.

Das Kriterium wurde teilweise erfüllt und wird mit neutral bewertet.

Begründete Schlussfolgerungen

Die Ergebnisse der Evaluation werden beschrieben:

Es kann festgehalten werden, dass im traditionellen Unterricht mehr Interaktion stattgefunden hatte. Die Studierenden (in der traditionellen Klasse!) erachteten dabei die freie Zeiteinteilung beim Lernen in der Online-Klasse als günstig. Allerdings wurde auch die Möglichkeit, Fragen zu stellen in der Präsenzkasse begrüßt. Es wurden Verbesserungsvorschläge seitens der Studierenden gemacht, die sich vor allem auf inhaltliche Aspekte (z.B. weniger Zeit für Wiederholungen anzusetzen) bezogen.

Die Online-Studierenden erachteten ebenfalls die Möglichkeit der freien Zeiteinteilung als günstig, sahen jedoch auch die Vorteile der Face-to-Face-Interaktion im Klassenraum. Die Arbeit empfanden sie als generell zeitintensiver und anstrengender in Bezug auf die Recherche nach relevanten Informationen. Die Verbesserungsvorschläge der Studierenden bezogen sich in erster Linie auf die Erreichbarkeit des Lehrenden, die Menge der Arbeit und die Organisation des Kurses.

Die Leistungen der Studierenden fielen in beiden Klassen gleich aus. Die Analyse der Testdaten wurde mittels eines T-Tests durchgeführt. Es wurden die Leistungen von 15 Online-Studierenden und 22 Studierenden im Präsenzkurs miteinander verglichen⁹⁵.

Wenngleich in beiden Gruppen signifikante Anstiege der Lernleistung vor und nach der Unterrichtseinheit festgestellt werden konnten, ergab die Analyse keine signifikanten Unterschiede zwischen den Untersuchungsgruppen.

Der Autor äußert hierbei jedoch Kritik am Testverfahren:

„While students were able to perform equitably on pre and post test analysis, it is important to point out that the reproduction of fact on a test do not necessarily represent the fullness of the learning experience. This is a particular concern when examining the learning experience for the on-line students as compared to that of the face-to-face traditional class“ (ebd.).

Das Kriterium wurde voll erfüllt und wird mit positiv bewertet.

8.3.11.3 Zuverlässigkeit

Angemessenheit der Forschungsentscheidungen

Im Vorfeld der Untersuchung wurden verschiedene Forschungsfragen aufgestellt, die sich auf den Unterschied der Lernleistungen der Studierenden in den Untersuchungsgruppen bezogen:

„Is there a significant difference between students choosing to receive their degrees using a distance educational model as compared to a more traditional classroom approach? Further, are there differences in content delivery leading to variations in the quality and effectiveness of learning for students choosing to

⁹⁵ Die restlichen Studierenden beteiligten sich nur an den Fragebogenerhebungen.

receive their degrees via distance education as compared to the traditional classroom approach?

What educational similarities and differences exist for individuals enrolled in distance education classes versus the traditional classroom setting, in particular, educational background, professional status in the field of education and professional status in the area of special education in particular?" (ebd.).

Weitere Fragen waren an den persönlichen Hintergrund, demographische Differenzen, Zeitmanagement sowie Haltungen und Meinungen der Studierenden angelehnt.

Die Wahl der Stichprobe wird erläutert, Begründungen für diese Auswahl werden aus dem Zusammenhang deutlich. Nicht erläutert werden hingegen die Gründe für die Wahl der Auswertungsmethode.

In Bezug auf die Erfassung des Lernerfolgs ist eine bloße Orientierung an Noten und Abschlussraten als nicht hinreichend zu bewerten, da damit nur die Performanceleistung, nicht aber der Lernprozess und noch weniger die Kompetenz bewertet werden kann.

Es wurde versucht, Rahmenbedingungen konstant zu halten, indem beide Kurse von derselben Lehrperson über die gleiche Zeit mit gleichem Inhalt durchgeführt wurden. Damit ist zwar eine größere Vergleichbarkeit der Gruppen gegeben, es können jedoch keine Aussagen zu Spezifika neuer Medien gemacht werden, da diese nicht ausgeschöpft worden sind. Dem widerspricht, dass alle Studierenden mit denselben Tests geprüft worden sind. Weiterhin ist die geringe Fallzahl zu kritisieren, mit der keine generalisierbaren Aussagen möglich sind.

In der Untersuchung wurde eine demographische Analyse der Studierenden durchgeführt, die signifikante Unterschiede in Bezug auf deren Ausbildung hervorbrachte:

„Examining further educational background of students for both traditional and on-line classes, several factors offered significant differences. Credits taken towards the masters degrees between the two groups varied significantly. The data indicates that traditional students have taken significantly more graduate course work in preparation of their graduate degree with 11 members (50%) of the participants having taken more than 28 credit hours. This contrasts with the 14 members (93%) of the on-line students taking between 10 and 27 credit hours" (ebd.).

Auch die Erfahrungen in der Lehre waren zwischen den Gruppen sehr unterschiedlich:

„Years teaching also varied significantly between the two groups in this study. In the traditional group, only 4 (18%) of the participants had ever taught, while 14 (93%) of the on-line participants had formal teaching experience" (ebd.).

Es stellt sich die Frage, ob angesichts der hier beschriebenen Verschiedenheit der Gruppen ein Vergleich überhaupt als sinnvoll erachtet werden kann, zumal – wie bereits erwähnt – auch die Untersuchungsinstrumente nicht valide oder reliabel waren.

Das Kriterium wurde in weiten Bereichen nicht erfüllt und wird mit eher negativ bewertet.

8.3.11.4 Bezug zu den Forschungsfragen

Lernerfolgsmessung

Der Lernerfolg wurde durch Vor- und Nachtests ermittelt. Dieser bestand aus 20 Multiple-Choice-Fragen aus der Kursdatenbank für die Kapitel 1 bis 6 des zu lernenden Stoffes. Beide Kurse wurden den gleichen Tests unterzogen.

„Students in the on-line program were adhering to the specified conditions of both the pre and post tests, i.e. that both tests were to be taken without books or notes“ (ebd.).

Über die Lernziele ist nichts bekannt. Durch die Prüfungsform kann jedoch davon ausgegangen werden, dass nur „erinnern“ erforderlich war, um den Test zu bestehen. Allerdings wurde das Prüfungsdesign nicht variiert, so dass unterschiedliche Bedingungen mit gleichen Tests abgefragt wurden. Hier hätten unterschiedliche Prüfungsformen eingesetzt werden müssen, um auch die Spezifika der neuen Medien zu untersuchen.

DIE KOGNITIVE PROZESS-DIMENSION					
1. Erinnern	2. Verstehen	3. Anwenden	4. Analysieren	5. Bewerten	6. (Er)schaffen
Test beider Gruppen					

Tabelle 24: Einordnung von P 11 in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)

Auf das Konstrukt Lernerfolg wird nicht näher eingegangen.

Das Kriterium wird mit eher negativ bewertet.

Signifikante Ergebnisse

Die Ergebnisse wurden auf Signifikanz untersucht. Dabei wurden in Bezug auf den Lernerfolg keine signifikanten Unterschiede zwischen den Gruppen festgestellt:

„There was no significant difference between the performance on the pretest for the distance education group and the pretest for the traditional education group at least $p=.05$

There was no significant difference between the performance on the post-test for the distance education group and the post test for the traditional education group at least $p=.05$ “ (ebd.).

In Bezug auf den Anstieg der erreichten Punktzahl vor und nach der Unterrichtseinheit konnten signifikante Unterschiede festgestellt werden:

„There was a significant increase in performance from the pretest to the post-test for the distance education group at least $p=.05$

There was a significant increase in performance from the pretest to the post-test for the traditional education group at least $p=.05$ “ (ebd.).

Weitere signifikante Ergebnisse konnten in Bezug auf die Organisation des Kurses und der Kommunikation festgestellt werden. Auf diese wurde im Punkt „Begründete Schlussfolgerungen“ bereits eingegangen.

Das Kriterium wurde erfüllt und wird als positiv bewertet.

8.3.11.5 Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Durchführbarkeit						
1.1	Angemessene Verfahren	x				
1.2	Praktische Verfahren	x				
2. Genauigkeit						
2.1	Valide und reliable Informationen			x		
2.2	Begründete Schlussfolgerungen	x				
3. Zuverlässigkeit						
3.1	Angemessenheit der Forschungsentscheidungen				x	
4. Bezug zu Forschungsfragen						
4.1	Lernerfolgsmessung				x	
4.2	Signifikante Ergebnisse	x				
	Gesamtbewertung	4	0	1	2	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 25: Bewertungsraster P11

8.4 Vergleich der Primärstudien und Integration der Ergebnisse

In diesem Kapitel sollen die Primärstudien zusammengefasst und die Ergebnisse der meta-evaluativen Untersuchung dargestellt werden. Dabei werden zwei Aspekte verfolgt: Zum einen sollen die Einzelkriterien über alle Studien hinweg ausgewertet werden, zum anderen soll eine exemplarische Darstellung der überdurchschnittlich bewerteten Studien erfolgen.

Am Ende des Kapitels werden die zentralen Ergebnisse zusammenfassend dargestellt.

8.4.1 Auswertung der Einzelkriterien

Bevor auf die jeweiligen Kriterien gesondert eingegangen wird, erachte ich einen Überblick über die Gesamtbewertungen als hilfreich. In Tabelle 26 werden diese zusammenfassend dargestellt.

Nr.	Kriterium	*	#	+	0	–
1. Durchführbarkeit						
1.1	Angemessene Verfahren	3	6	1	1	0
1.2	Praktische Verfahren	4	2	4	1	0
2. Genauigkeit						
2.1	Valide und reliable Informationen	0	2	6	3	0
2.2	Begründete Schlussfolgerungen	2	4	5	0	0
3. Zuverlässigkeit						
3.1	Angemessenheit der Forschungsentscheidungen	0	1	4	6	0
4. Bezug zu Forschungsfragen						
4.1	Lernerfolgsmessung	0	1	1	8	1
4.2	Signifikante Ergebnisse	11	0	0	0	0
*positiv		# eher positiv		+ neutral		0 eher negativ – negativ

Tabelle 26: Zusammenfassung der Einzelkriterien⁹⁶

Es ist auffällig, dass die Verfahren in den meisten Studien angemessen und praktisch orientiert gewesen waren, was für eine gute Durchführbarkeit spricht.

⁹⁶ Markiert sind jeweils die meistgegebenen Bewertungen in der jeweiligen Kategorie.

In Bezug auf die Genauigkeit fallen die Ergebnisse schon deutlich schlechter aus. Hier bewegen sich die Bewertungen im neutralen Bereich, was bedeutet, dass hier sowohl positive als auch negative Aspekte aufgefallen sind.

Auch die Zuverlässigkeit der Studien ist nicht als sonderlich gut anzusehen. Nur 1 von 11 untersuchten Primärstudien ist mit „eher positiv“ bewertet worden. Dagegen ist das Kriterium der signifikanten Ergebnisse als durchweg äußerst positiv einzuschätzen.

Diese Übersicht legt die folgende Vermutung nahe: Am besten erfüllt werden die formalen Kriterien, die in den meisten Studien qua Evaluationsdesign festgelegt worden waren und deren Einhaltung relativ einfach ausführbar war. So werden beispielsweise die Kriterien der Gruppenbildung in den meisten Studien deutlich gemacht. Weniger gut beurteilt werden die Kriterien, deren Grad der Erfüllung sich erst im Laufe der jeweiligen Untersuchungen zeigte und die mehr auf die Innensicht der Evaluation und auf argumentative Fragen bezogen waren.

Um diese Vermutung zu überprüfen, werden im Folgenden die Kriterien einzeln zusammengefasst.

8.4.1.1 Durchführbarkeit

Angemessene Verfahren

9 von 11 untersuchten Studien erfüllen das Kriterium „sehr“ oder „eher“ und zeigen damit, dass angemessene Verfahren zu Grunde gelegt wurden. Der Rahmen der Evaluationen wird also größtenteils deutlich.

Grundsätzlich ist davon auszugehen, dass das Verhältnis des Aufwandes zum Nutzen der Evaluation in den meisten Fällen angemessen war. Nur in wenigen Fällen schien dies auf Grund der Größe der Stichprobe in Relation zur Anlage der Einzelstudie nicht ausgewogen gewesen zu sein. Vereinzelt werden die Belastungen des Evaluationsgegenstandes für die Studierenden als zu groß eingeschätzt, weshalb an dieser Stelle auch Bewertungen außerhalb des positiven Bereichs vergeben werden.

Weniger oft beachtet wurden Gender-Aspekte oder kulturelle und soziale Differenzen innerhalb der untersuchten Kohorte. Diese wurden – wenn überhaupt – am Rande erwähnt, waren i.d.R. aber nicht Teil der jeweiligen Forschungsfrage.

Praktische Verfahren

In allen untersuchten Studien wurden die Messungen mit Test- und Kontrollgruppen durchgeführt. In den Publikationen werden die Kriterien der Gruppenbildung in den meisten Fällen erläutert, allerdings mehr oder weniger detailliert. So wurde beispielsweise das Verfahren der Zuordnung zu den einzelnen Untersuchungsgruppen nicht immer transparent erläutert.

Die Kriterien der Gruppenbildung sind also nicht konstant nachvollziehbar. An dieser Stelle hätten noch zusätzliche Begründungen gegeben werden können.

8.4.1.2 Genauigkeit

Valide und reliable Informationen

Dieses Kriterium weist eine hohe Streuung auf, was darauf hindeutet, dass in den Primärstudien sehr unterschiedlich mit diesem Aspekt umgegangen wurde. In den meisten Fällen wurden die üblichen statistischen Verfahren zur Analyse der Daten angewendet. In den Veröffentlichungen werden die Verfahren der Datengewinnung und -analyse überwiegend detailliert beschrieben.

Andererseits werden Angaben zu Gütekriterien mehrfach vernachlässigt. Auf die Reliabilität der Messinstrumente und die Validität der Untersuchung wird nur in Einzelfällen direkt Bezug genommen. So werden die Verfahren der Instrumentenentwicklung oft nicht beschrieben. An dieser Stelle lässt sich jedoch vermuten, dass diese Angaben für die Publikation nicht als relevant betrachtet wurden, aber in der Studie dennoch berücksichtigt worden sind. Andernfalls würden die Verfahren der Datenanalyse keinen Sinn ergeben – was nützt eine Auswertung von Daten, die mittels einer Skala erhoben worden sind, wenn diese nicht das misst, was sie messen soll? Unter Bezugnahme auf diese Kritik sind die Verfahren der Datengewinnung und -analyse nicht durchgängig nachvollziehbar.

Begründete Schlussfolgerungen

Dieses Kriterium wird überwiegend erfüllt. Die Ergebnisse der Evaluation werden in allen Primärstudien ausführlich dokumentiert.

Unterschiede zwischen den einzelnen Veröffentlichungen gibt es in Bezug auf die Angaben darüber, welche Daten die jeweiligen Schlussfolgerungen zulassen. Zwar wurden diese in den meisten Fällen auf Grund der Datenlage gezogen und es wurden keine weitergehenden Interpretationen angestellt, dennoch sind die Schlussfolgerungen gelegentlich sehr vage. Beispielsweise fehlen genauere Angaben, welches Ergebnis auf Grund von Interviews und welches auf Grund der Überprüfung von Lernresultaten zu Stande kam, oder ob die Ergebnisse mit Hilfe von an Gütekriterien orientierten Skalen erzielt wurden.

In 9 von 11 untersuchten Studien werden darüber hinaus auch Limitationen deutlich, die die Autorinnen und Autoren selbst äußern. Hierbei wird beispielsweise auf die Vergleichbarkeit der Untersuchungsgruppen eingegangen. Andere AutorInnen geben an, dass keine Aussagen in einem erweiterten Kontext getroffen werden können.

Trotz der oft nicht erreichten vollständigen Erfüllung dieses Kriteriums wird in den Primärstudien relativ gut auf die an dieser Stelle relevanten Aspekte eingegangen.

8.4.1.3 Zuverlässigkeit

Angemessenheit der Forschungsentscheidungen

Dieses Kriterium gehört zu denjenigen, die in den Primärstudien am wenigsten erfüllt werden. Dies liegt sicherlich jedoch auch daran, dass es sich hierbei nicht um ein Kriterium handelt, welches in dieser Form zu den Evaluationsstandards gehört, sondern es wurde zum Zweck dieser Arbeit um weitere Indikatoren ergänzt.

Nur in 4 der untersuchten Studien werden Hypothesen angegeben. 3 andere Studien nennen immerhin explizite Forschungsfragen. Auch die Auswahl von Stichprobe und Auswertungsmethode wird nur in den wenigsten Fällen begründet. Überdies werden hier weitere Kritikpunkte subsumiert, wie zu geringe Fallzahlen der Einzelstudien, unterschiedliche oder zu späte Befragungszeitpunkte oder unangemessene Erhebungsmethoden.

Was an dieser Stelle jedoch am stärksten ins Gewicht fällt, ist die Beobachtung, dass sich die Interpretation der Ergebnisse in 7 der 11 untersuchten Studien auf die unabhängigen Variablen beziehen, die zum besseren Vergleich der Gruppen konstant gehalten wurden. So wurden beispielsweise die Prüfungen oftmals mit demselben Test in allen Untersuchungsgruppen abgehalten. Damit ist zwar eine größere Vergleichbarkeit der Gruppen gegeben, es können auf dieser Grundlage jedoch keine Angaben zu den jeweils spezifischen Funktionen und Eigenschaften der einzelnen Evaluanden gemacht werden. Dennoch werden Beurteilungen darüber abgegeben, welches Lernsetting besser geeignet sei.

Auf diesen Aspekt wird in Kapitel 8.5 (Darstellung der Ergebnisse in Bezug auf die Forschungsfragen), sowie im Kapitel 9.1 des Fazits ab Seite 181 detaillierter eingegangen.

8.4.1.4 Bezug zu den Forschungsfragen

Lernerfolgsmessung

An dieser Stelle möchte ich exemplarisch noch einmal auf die Einzelbewertungen dieses für die vorliegende Arbeit entscheidenden Kriteriums eingehen und an Hand jeder Primärstudie proaktiv aufzeigen, wie das Konstrukt Lernerfolg operationalisiert wird.

- P1. Die unabhängige Variable Lernerfolg wurde in der vorliegenden Untersuchung ausschließlich als Performanceleistung an Hand einfach zu messender Werte definiert und zwar unter Bezugnahme auf die abhängigen Variablen Note des Vortests, Noten von drei Multiple-Choice-Tests sowie Abbruchs- bzw. Bestehensquoten.
- P2. Es wurde durch die Zitierung der „Middle States Commission on Higher Education“ (2001) darauf hingewiesen, dass das sogenannte „outcome assessment“ zwei Absichten verfolge –

- Verantwortlichkeit („accountability“) und Effektivität („effectiveness“).
- P3. Zur Feststellung des Lernerfolgs wurde neben den Abschlussnoten ein so genannter „Mastery Test“ durchgeführt, der aus einem dreiteiligen Fragebogen besteht.
 - P4. In der vorliegenden Studie wurde Lernerfolg mit Behaltensleistung gleichgesetzt und durch Wissensabfragen abgeprüft.
 - P5. Die Messung des Lernerfolgs orientierte sich ausschließlich an den Ergebnissen der schriftlichen Tests.
 - P6. Die unabhängige Variable Lernerfolg wurde in der vorliegenden Untersuchung ausschließlich als Performanceleistung definiert und zwar an Hand der Ergebnisse von Prä-, Post- und Placement-Test als abhängige Variablen.
 - P7. Die Messung des Lernerfolgs orientierte sich an Hand der abhängigen Variable der Noten des Abschlusstests.
 - P8. Ein Indikator für den Lernerfolg schien in der vorliegenden Studie in der Abbruchquote gesehen worden zu sein. Die Autoren geben an, dass diese in beiden Kursen gleich hoch gewesen sei.
 - P9. Ein Indikator schien die Abbruchquote zu sein. Überdies wurde die abhängige Variable der Noten des Abschlusstests (GPA) einbezogen.
 - P10. In der Untersuchung wurde der Lernerfolg an Hand der Noten zweier Kursprojekte, des Abschlusstests sowie der Gesamtnote definiert.
 - P11. Der Lernerfolg wurde durch einen Vor- und einen Nachtest ermittelt. Diese Tests bestanden aus 20 Multiple-Choice-Fragen über den Kursinhalt.

Diese Übersicht macht deutlich, dass das Konstrukt Lernerfolg in 7 Studien allein durch die Noten verschiedener Tests operationalisiert wurde. Dazu kann ergänzend festgehalten werden, dass Lernerfolg, Lernprozess und Lernleistung nur in den Kursprojekten von P10 unterschieden wurden und hier auch nur während des Unterrichts und nicht in der Prüfung. Die Studien P2 und P3 gingen immerhin auf die verschiedenen Dimensionen des Konstrukts ein. 2 Studien – P8 und P9 legten der Operationalisierung von Lernerfolg die Abbruchquote zu Grunde.

Nur 4 Studien (P5, P6, P8, P10) gaben explizite Lernziele an. In 2 anderen (P4 und P9) ließen sich diese aus dem Kontext folgern.

Bereits bei der Beschreibung des vorherigen Kriteriums wurde deutlich, dass unterschiedliche Rahmenbedingungen oftmals mit den gleichen Tests abgefragt wurden. Für die Angemessenheit der Forschungsentscheidungen ist dies ein sehr wichtiger Aspekt, um zu zeigen, dass nicht alle Evaluanden gleichermaßen bewertet werden können. Die Argumentation muss an dieser Stelle in Bezug auf die

Prüfungen erneut aufgegriffen werden: In 9 Evaluationen⁹⁷ (P1, P3, P4, P6, P7, P8, P9, P10 und P11) lagen die getesteten Lernziele zwar auf derselben Ebene der kognitiven Prozesse (d.h. der Wissenstest bezog sich in beiden Gruppen auf den gleichen Prozess), jedoch bezog sich das Prüfungsdesign nicht auf die intendierten Lernziele bzw. variierte nicht. Wurde beispielsweise ein CBT zu Selbstlernzwecken, das nur eine Erinnerungsleistung erforderte, verglichen mit einem kollaborativen Präsenzkurs, der die Fähigkeit des Analysierens voraussetzte, dann hätte jedoch auch die Prüfungsform diesen unterschiedlichen Bedingungen angepasst sein müssen. Dies war i.d.R. jedoch nicht der Fall.

Signifikante Ergebnisse

Dieses Kriterium fiel durchweg sehr positiv aus, da die Unterschiede zwischen Test- und Kontrollgruppe in allen Studien auf Signifikanz getestet wurden.

8.4.2 Analyse der bestbewerteten Primärstudien

Im Folgenden sollen die Primärstudien mit den besten Bewertungen genauer in den Blick genommen werden. Damit kann gezeigt werden, dass die bereits aufgezeigten Schwierigkeiten auch auf die Untersuchungen zutreffen, die ansonsten überdurchschnittlich bewertet wurden.

Im Mittel erhielten die Primärstudien 1,8 positive (*) und 1,5 eher positive (#) Bewertungen. Als überdurchschnittlich von diesem Wert abweichend sind die folgende Studien auszumachen:

Studie	positive Bewertungen	eher positive Bewertungen
P11	****	
P3	***	#
P6	**	##
P10	**	#
P9	**	#

Tabelle 27: überdurchschnittlich bewertete Primärstudien (Sortierung nach Bewertung)

Betrachtet man diese Auswahl genauer, so fällt auf, dass die Streuung der Bewertungen innerhalb der Einzelstudien sehr hoch ist –

⁹⁷ Bei P2 war eine Einordnung nicht möglich.

auch die insgesamt bestbewerteten Studien konnten nicht zwangsläufig alle Kriterien voll erfüllen. Im Gegenteil – gerade die Studien, die die formalen Kriterien wie „angemessene Verfahren“, „Praktische Verfahren“, „Begründete Schlussfolgerungen“ und „Signifikante Ergebnisse“ „gut“ bis „sehr gut“ erfüllen, wiesen in Bezug auf die Angemessenheit der Forschungsentscheidungen oder der Lernerfolgsmessung oft nur „neutrale“ oder gar „eher negative“ Bewertungen auf. Dieser Eindruck verstärkt sich umso mehr, wenn man die Bewertungen der übrigen Kriterien hinzuzieht (siehe Anhang). Diese Kriterien zielten vor allem auf standardisierte Verfahren ab, die sehr gut erreicht werden konnten: Keine der hier angesprochenen Studien erzielte dabei Bewertungen im negativen Bereich.

Es lohnt sich darüber hinaus die Studien in den Blick zu nehmen, die in den abweichenden Kriterien („angemessene Verfahren“, „Lernerfolgsmessung“) die besten Bewertungen erzielten. Hier konnte beispielsweise P2 in Bezug auf die „Angemessenheit der Forschungsentscheidungen“ punkten – eine Studie, die auf Grund der sonstigen Bewertungen eher im mittleren Bereich anzusiedeln ist. Das Kriterium wurde vor allem deshalb als „eher positiv“ bewertet, da die Untersuchung auf Selbsteinschätzungen der Studierenden beruhte und nicht der Anspruch deutlich wurde, den „absoluten“ Lernerfolg messen zu wollen. Da jedoch ausschließlich die Zufriedenheit der Studierenden abgefragt wurde, hat die Studie nur bedingte praktische Relevanz für die Forschungsfrage.

In Bezug auf das Kriterium „Lernerfolgsmessung“ wurden die besten Bewertungen in P5 („eher positiv“) und P10 („neutral“) vergeben. Diese Studien fielen dadurch auf, dass Lernziele benannt oder abgeleitet werden konnten. Diese waren überdies in beiden Gruppen gleich. Das bedeutet, dass bei der Operationalisierung des Konstrukts Lernerfolg die gleichen Ausprägungen von Lernerfolg zu Grunde gelegt wurden. Allerdings hätten sich die Prüfverfahren auch auf diese Lernziele beziehen müssen.

8.5 Darstellung der Ergebnisse in Bezug auf die Forschungsfragen

Wie zu Beginn der meta-evaluativen Untersuchung dargestellt, verfolgt diese Arbeit das Ziel, die Möglichkeiten und Grenzen von Vergleichsuntersuchungen in Bezug auf mediengestütztes Lernen darzulegen. Zur Beantwortung dieser Fragestellung wurden folgende Überlegungen angestellt:

- Wie wird der Lernerfolg in den verschiedenen Studien definiert?
- Wird das Konstrukt Lernerfolg in Dimensionen aufgegliedert?
- Wie wird Lernerfolg operationalisiert?

- Werden Lernerfolg, Lernprozess und Lernleistung unterschieden?
- Welches Testverfahren wird der Untersuchung zu Grunde gelegt?
- Kommen ältere und neuere Studien zu gleichen Ergebnissen?
- Gibt es Aspekte, die sich verallgemeinern lassen?

Um in diesen Punkten Klarheit zu schaffen, sollen die Primärstudien in die folgende synoptische Übersicht eingeordnet werden.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
Jahr	1999	2001	1995	2005	1998	2002	2002	2001	2001	2005	2006
Definition Lernerfolg	Behaltensleistung	studentisches Lernen/ Effektivität	Behaltensleistung, Motivation	Behaltensleistung	Behaltensleistung	Behaltensleistung	Behaltensleistung	Zufriedenheit	Behaltensleistung, Zufriedenheit	Behaltensleistung	Behaltensleistung
Operationalisierung	Noten von Abschlusstests	Zufriedenheit der Studierenden	Noten von Abschlusstests	Punktwerte von Wissenstests	Ergebnisse schriftlicher Tests	Ergebnisse schriftlicher Tests	Noten von Abschlusstests	Abbruchquote	Noten von Abschlusstests, Abbruchquote	Noten von Projekten und Abschlusstests	Ergebnisse schriftlicher Tests
Unterscheidung von LE, LP und LL⁹⁸	nein	nein	nein	nein	nein	nein	nein	nein	nein	während des Kurses	nein
Testverfahren	Multiple-Choice, GPA-Werte, Abbruchquote	Fragebogen zur Selbsteinschätzung	Wissenstests, Fragebogen	Multiple-Choice	open book bzw. closed book	vermutlich Multiple-Choice	vermutlich Multiple-Choice und Selbsteinschätzung	Fragebogen zur Selbsteinschätzung	Fragebogen zur Selbsteinschätzung, Wissenstests	nicht bekannt	Multiple-Choice
Ergebnisse	Onlinekurs signifikant besser	Präsenzkurs signifikant besser	keine signifikanten Unterschiede	CBT signifikant besser	keine signifikanten Unterschiede	Präsenzkurs signifikant besser	keine signifikanten Unterschiede	keine signifikanten Unterschiede	keine signifikanten Unterschiede	keine signifikanten Unterschiede	keine signifikanten Unterschiede
formulierte Lernziele	erinnern	keine	verstehen anwenden	anwenden analysieren	anwenden	erinnern verstehen	nicht bekannt	erinnern verstehen	analysieren erschaffen bewerten	bewerten erschaffen	nicht bekannt
getestete Lernziele	erinnern	keine	nicht bekannt	erinnern	erinnern verstehen	erinnern	erinnern	erinnern	vermutlich verstehen	vermutlich verstehen	erinnern
Ausprägung Lernerfolg	Lernziele unbekannt, Tests gleich	nicht möglich	Tests gleich, Lernziele unterschiedlich	Tests gleich, Lernziele unterschiedlich	in den Gruppen unterschiedlich	Tests gleich, Lernziele unterschiedlich	Lernziele unbekannt, Tests gleich	Tests gleich, Lernziele unterschiedlich,	Tests gleich, Lernziele unterschiedlich	Tests gleich, Lernziele unterschiedlich	Lernziele unbekannt, Tests gleich

Tabelle 28: Synoptische Darstellung der Primärstudien

⁹⁸ Lernerfolg, Lernprozess und Lernleistung

An Hand der Übersicht wird ergänzend zu den bereits beschriebenen Ergebnissen deutlich, dass der Lernerfolg in den meisten Studien mit einem Multiple-Choice-Test überprüft und somit als reine Behaltensleistung definiert wurde: Je mehr Punkte im Abschlusstest erreicht wurden, desto höher war der Lernerfolg. In Einzelfällen wurde zusätzlich die Zufriedenheit der Studierenden und deren Motivation erhoben oder es wurden Abbruchquoten analysiert. Keine der hier evaluierten Studien nahm bei der Definition und Operationalisierung des Konstrukts Lernerfolg andere Dimensionen – wie z.B. Lernprozess, Kompetenz o.ä. in Betracht. Daran schließt sich an, dass nur in einem Fall überhaupt zwischen Lernerfolg, Lernprozess und Lernleistung unterschieden wurde. Dies erfolgte hierbei allerdings auch nur durch die konkrete Umsetzung im Unterricht, nicht durch das Prüfverfahren.

Angesichts der Heterogenität der Ansätze verwundert es nicht, dass auch die Ergebnisse sehr unterschiedlich ausfielen. In zwei Fällen schnitt die medial vermittelte Instruktion signifikant besser ab, in ebenfalls zwei Fällen der Präsenzkurs. In 7 Studien waren keine signifikanten Unterschiede zwischen den Gruppen erkennbar. Differenzen zwischen älteren und neueren Studien fielen an dieser Stelle ebenfalls nicht auf.

Interessant ist die Ausprägung des Lernerfolgs auf der Ebene der kognitiven Prozesse. Hier wird deutlich, dass große Schwierigkeiten darin bestanden, das Lernsetting mit der Prüfungsform zu vereinbaren. Wie bereits erwähnt, lag das Problem der meisten Studien darin, dass die Prüfung des Lernerfolgs mittels der gleichen Tests in beiden Untersuchungsgruppen erfolgte. Diese Problematik wurde bereits zu Beginn dieser Arbeit deutlich: Es ist zwar legitim, die unabhängigen Variablen zum besseren Vergleich der Gruppen konstant zu halten – beispielsweise durch gleiche Inhalte, gleiche Lehrpersonen usw. Allerdings sollte man sich bei einem solchen Vorgehen darüber bewusst sein, dass auf dieser Grundlage keine Angaben zu den jeweils spezifischen Funktionen und Eigenschaften der einzelnen Evaluanden gemacht werden können. Wenn beispielsweise ein textbasierter Wissenstest in beiden Untersuchungsgruppen eingesetzt wird, dann können in diesem Fall keine Aussagen über den Aspekt der Interaktivität neuer Medien getroffen werden. Ein Vergleich der Gruppen liefert also trotz der Reduktion der Einflussfaktoren keine zuverlässigen Ergebnisse, obwohl die getesteten Lernziele auf einer Ebene liegen.

Unter diesen Studien fielen weitere Schwierigkeiten auf: In 4 Studien (P3, P4, P9 und P10) wurde die Prüfung des Lernerfolgs in Bezug auf die Lernziele auf unterschiedlichen kognitiven Prozessebenen durchgeführt. Dies war beispielsweise dann der Fall, wenn ein Lernziel auf der Ebene des Anwendens formuliert wurde, der Test sich aber nur auf die Ebene des Erinnerns bezog. Das bedeutet, dass die Unterschiedlichkeit von Lernziel und Lernzielüberprüfung allein in einer der Versuchsgruppen so groß war, dass die Evaluation nur begrenzten Wert hatte. Dass diese beiden Gruppen auch noch zusätzlich miteinander verglichen wurden, erschwerte die Gültigkeit der Ergebnisse.

Trotz der geäußerten Kritik fielen viele Aspekte in den Primärstudien positiv auf. So entsprach in P6 und P8 mindestens ein Lernziel der Lernerfolgsprüfung, so dass zumindest eine Untersuchungsgruppe zuverlässig getestet wurde. Der Vergleich mit der jeweils anderen Gruppe war dennoch nicht aussagekräftig.

In P5 wurden in beiden Gruppen die gleichen Lernziele intendiert, also die gleichen Ausprägungen von Lernerfolg zu Grunde gelegt. Das Prüfungsdesign wurde entsprechend variiert, um nicht unterschiedliche Bedingungen mit gleichen Tests abzufragen. Allerdings hätten sich die unterschiedlichen Prüfverfahren (getestet wurden „erinnern“ und „verstehen“) auch auf die Lernziele (in beiden Gruppen „anwenden“) beziehen müssen.

Auf der Grundlage dieser Ergebnisse kann in Bezug auf die Hypothesen folgendes festgehalten werden:

1. Lernerfolg ist nicht eindeutig operationalisierbar.

Diese Hypothese ist – entgegen der Erwartungen – nicht zutreffend.

Zwar fielen allein in der relativ geringen Anzahl von untersuchten Primärstudien unterschiedliche Definitionen von Lernerfolg und verschiedene Operationalisierungen dieses Konstrukts auf. Ferner wurde der Lernerfolg in der Mehrzahl der Fälle als Behaltensleistung definiert. Dies widerspricht der differenzierten Aufzählung von Dimensionen, die in der Literatur gegeben werden. So muss gefolgert werden, dass der Lernerfolg in den Primärstudien auf sehr unterschiedliche Weise, jedoch nicht ausreichend detailliert operationalisiert wurde.

Dennoch zeigt die Untersuchung, dass auch eindeutige Operationalisierungen, die einen realistischen Vergleich der Untersuchungsgruppen gestatten, durchaus möglich sind. Im Fazit der Arbeit (Kapitel 9.1 ab Seite 181) wird aufgezeigt werden, welche Bedingungen zutreffen müssen, um Lerngruppen unterschiedlicher Settings miteinander vergleichen zu können.

2. Unspezifische Vergleiche von Online- versus Präsenzlernen sind nicht sinnvoll anwendbar bzw. nicht zu verallgemeinern.

Diese Hypothese ist zutreffend.

Die Ergebnisse zeigen, dass keine der in dieser Arbeit untersuchten Studien sichere Aussagen über den Lernerfolg in dem einen oder anderen Lernsetting treffen konnte. So wurden unterschiedliche Bedingungen in den meisten Fällen mit dem gleichen Test überprüft. Zusätzlich wurden Interferenzen zwischen Lernzielen und Prüfungsdesign nicht durchgängig beachtet.

3. Die Methode der Metaanalyse ist nur bedingt geeignet, um Rückschlüsse auf den Lernerfolg zuzulassen.

In Metaanalysen werden Ergebnisse verschiedener Primärstudien zur gleichen Thematik zusammengefasst und unter einem gemeinsamen Nenner subsumiert. Problematisch ist hierbei die Vergleichbarkeit der einzelnen Studien, da oftmals unterschiedliche Auswertungsmethoden, Eigenschaften der Stichprobe und Operationalisierungen zu Grunde liegen.

Wenn jedoch unspezifische Vergleiche von Online- versus Präsenzlernen nach der Ergebnislage nicht ohne weiteres sinnvoll anwendbar sind, kann auch die Methode der Metaanalyse nur bedingt genutzt werden, um verallgemeinerbare Rückschlüsse auf den Lernerfolg zuzulassen. Die verschiedenen Sichtweisen in den einzelnen Primärstudien machen generalisierbare Zusammenfassungen fast unmöglich. Solange nicht auf die unterschiedlichen Dimensionen des Konstrukts Lernerfolg Bezug genommen wird, bleibt der Lernbegriff zu allgemein.

Insgesamt wird deutlich, dass der Aspekt der Lernerfolgsmessung in allen untersuchten Studien detaillierter hätte operationalisiert werden müssen. Die Aufmerksamkeit der Evaluation wurde in den meisten Fällen eher auf die Methodik und auf formale Aspekte gerichtet. Dies wird vor allem dadurch erkennbar, dass die unterschiedlichen Dimensionen des Konstrukts Lernerfolg nicht berücksichtigt worden sind. Viele Studien bezogen sich in der Formulierung ihrer Lernziele nur auf die niedrigen Stufen der kognitiven Prozesse. Die Messung von Kompetenzen war dabei jedoch nicht möglich.

9 Fazit

In dieser Arbeit wurde versucht, die Problematik der Vergleichbarkeit von Lernerfolg durch verschiedene Instruktionsmethoden zu betrachten. Von Beginn an bestand dabei die Überlegung, dass ein Vergleich von Online- versus Präsenzlernen nicht uneingeschränkt möglich sein kann, da beide Lernsettings unterschiedliche Methodiken, Ziele und Lernformen voraussetzen.

Zur theoretischen Ausarbeitung dieser noch eher unsystematischen Vorstellung wurde in einem ersten Schritt die Lernqualität beleuchtet. Wie kann eine hohe Qualität beim Lernen erreicht werden? Es wurde schnell deutlich, dass hier mehrere Aspekte eine Rolle spielen. So kann sich die Lernqualität sowohl auf das Lernangebot als auch auf das Lernergebnis oder den Lernprozess beziehen. Die Qualität der Behaltensleistung ist also ebenso ein Aspekt von Lernqualität, wie beispielsweise die Qualität der Lernumgebung. Selbst ein Konsens über die Definition von Lernqualität konnte in der Diskussion bisher nicht erzielt werden. Lernqualität ist also – wie Qualität generell – als ein Konstrukt zu sehen, das in der Realität nicht in unmittelbar greifbarer Form existiert. Wird beispielsweise eine Software als Produkt mit hoher Lernqualität umworben, so ist daraus nicht direkt ersichtlich, ob damit die Qualität des Lernprozesses gemeint ist, oder ob es nicht doch darum geht, möglichst vorhersagbare und leicht messbare Lernergebnisse zu erzielen. Will man das Konstrukt Lernqualität überprüfen oder evaluieren, müssen sowohl Lernerfolg, Lernprozess und Lernleistung an Hand bestimmter Kriterien bewertet und in messbare Indikatoren zerlegt werden. Der Lernerfolg dient bei der Operationalisierung der Kriterien in vielen sozialwissenschaftlichen Studien jedoch als einziger Indikator: Je größer der Lernerfolg, desto höher die Qualität des Lernangebotes.

Auf diese Erkenntnis folgten schnell die nächsten Fragen: Was macht Lernen erfolgreich? Wie definiert man Lernerfolg? Wie kann er gemessen werden? Vermutlich unumstritten ist, dass sich ein hoher Lernerfolg durch intensives Trainieren von Kompetenzen einstellen kann. Betrachtet man jedoch das Lernergebnis, muss festgehalten werden, dass auch eine schlichte Behaltensleistung zu guten Lernergebnissen und qua Definition damit auch zu hoher Lernqualität führen kann. Bei einer näheren Betrachtung der verschiedenen Aspekte und Dimensionen, die Lernerfolg ausmachen, wurde deutlich, dass Lernerfolg nicht unmittelbar messbar ist – bewertet wird in den meisten Fällen nur das Zeigen einer erwarteten Leistung. Ob diese Fähigkeit auch beherrscht wird, kann daraus nicht abgeleitet werden. Wenn eine Medizinstudentin zum Beispiel in einem schriftlichen Test zeigt, dass sie weiß, wie man einen Blinddarm entfernt, bedeutet dies noch längst nicht, dass sie dies auch in der Praxis meistern kann! Es muss also zwischen Lernaktivitäten und Lernprozessen unterschieden werden. Darüber hinaus ist der Lernerfolg von Kontexten und Komponenten des jeweiligen Lernsettings abhängig.

Aus diesen Überlegungen wurde klar, dass auch der Lernerfolg als ein Konstrukt anzusehen ist, das aus verschiedenen Dimensionen besteht.

Um beispielhaft einige dieser Dimensionen aufzuzeigen, wurde in einem nächsten Schritt das Modell der lernrelevanten Kontexte und Komponenten nach Sigmar-Olaf Tergan (2004) herangezogen, in welchem zwischen Lernenden und Lernszenarien unterschieden wird. Darüber hinaus werden kognitive und prozedurale Lernprozesse differenziert und es werden die Kontexte und Komponenten der Lernszenarien herausgestellt. Dieses Modell zeigte vor allem, dass eine vollständige Auflistung aller Faktoren erfolgreichen Lernens nicht realisierbar ist. Der Lernerfolg wird demnach also nicht durch einzelne Komponenten ermöglicht, sondern nur durch das komplexe Gesamtgefüge der Konstruktdimensionen. Aussagen, die auf Grund des Lernerfolgs über die Lernqualität getroffen werden, sind demzufolge nur bedingt brauchbar, da ein Konstrukt – die Lernqualität – durch ein anderes Konstrukt – den Lernerfolg – operationalisiert wird.

Festgehalten werden konnte jedoch der Anspruch, dass der Lernerfolg den Anforderungen des Lernangebotes entsprechen muss. Dies impliziert, dass bei der Bewertung der Lernleistung auch immer eine Vorstellung über die Lernziele erforderlich ist. Die unterschiedlichen Dimensionen des Konstrukts Lernerfolg dürfen also – auch in Bezug auf die Lernziele – nicht durcheinander gebracht werden. Genau diese Problematik taucht jedoch in vielen Studien auf: Um den Erfolg von E-Learning im Vergleich zu traditionellen Lernformen zu erfassen, werden oftmals Vergleichsuntersuchungen durchgeführt. Unter Berücksichtigung der eben aufgezeigten Schwierigkeiten kann davon ausgegangen werden, dass auch Evaluationen, die eine Maßnahme in Bezug auf ihren Lernerfolg untersuchen, nicht problemlos durchzuführen sind. In der Regel haben wir es dabei mit zwei verschiedenen Gefügen von Lernerfolgsdimensionen zu tun, deren Faktoren unterschiedlich in Beziehung zu setzen sind. Dazu kommen generelle Probleme von Vergleichsuntersuchungen, wie beispielsweise, dass der Lernerfolg in den meisten Fällen durch das Testverfahren definiert wird. In der Regel wird für beide Testgruppen das gleiche Prüfverfahren festgelegt, damit eine hohe Vergleichbarkeit der Ergebnisse gegeben ist. Dadurch besteht jedoch die Gefahr, die spezifischen Eigenschaften eines Evaluanden zu vernachlässigen und diese nicht ausreichend zu erfassen.

Ein solches Verfahren entspräche dem folgenden Vorgehen in der Analogie von Peter Baumgartner (1997:135), die ich an dieser Stelle erneut aufgreifen möchte⁹⁹: Um zu überprüfen, in welchem Fahrzeug das subjektive Fahrerlebnis das bessere sei, wird ein Vergleich zwischen einem PKW und einem LKW angestellt. Das Konstrukt „Fahrerlebnis“ wird dabei in verschiedene Dimensionen gegliedert, beispielsweise in „Spritverbrauch“, „Geschwindigkeit“, „Komfort“, „Beförderung von Lasten“ usw. Der Vergleich der beiden Fahrzeuge bezieht sich dabei jedoch immer nur auf eine dieser Dimensionen, beispielsweise den Spritverbrauch (da dieser als grundlegend betrachtet wird). Dabei schneidet der PKW natürlich deutlich besser ab, da seine Verbrauchswerte signifikant unter denen des LKWs liegen. Alle anderen Dimensionen werden jedoch vernachlässigt: Dass der LKW beispielsweise mehr Lasten befördern kann,

⁹⁹ vgl. Seite 50

wird an dieser Stelle nicht als relevant betrachtet. Es bleibt jedoch die Frage, warum der LKW dann überhaupt Teil des Vergleichs ist. Wenn es nur um den Spritverbrauch geht, warum sollte man überhaupt einen LKW auswählen? Letztendlich wird dennoch gefolgert, das Fahrerlebnis in einem PKW sei besser als in einem LKW.

In Bezug auf die Forschungsfrage dieser Arbeit entspräche der Spritverbrauch beispielsweise der Menge des Behaltens von Faktenwissen. Zwei unterschiedliche Lernsettings (ein Online-Kurs in einer interaktiven Lernumgebung und eine klassische Vorlesung) werden miteinander verglichen, doch statt die verschiedenen Lernziele zu beachten, wird nur eine Dimension berücksichtigt. So wird die Behaltensleistung mit dem Lernerfolg gleichgesetzt und es wird resümiert, dass beispielsweise beim Auswendiglernen mittels interaktiver Angebote keine Vorteile gegenüber Drill-and-Practice-Verfahren auszumachen wären. Es werden also nur die Dimensionen verglichen, die ein Evaluand von vornherein nicht besser erfüllen kann als der andere. Alle übrigen Dimensionen, in denen dieser Evaluand bevorteilt wäre (z.B. Interaktivität, Kommunikationsmöglichkeiten usw.), werden vernachlässigt. Betrachtet man dabei zusätzlich die Dimension der kognitiven Prozesse, so fällt auf, dass jeweils nur der niedrigste Prozess der Hierarchie (nämlich „erinnern“) verglichen wird. Höhere Prozesse, die im Online-Kurs möglich wären, werden gar nicht erst in die Untersuchung einbezogen. Wenn es nur um die Behaltensleistung geht, warum sollte man dann überhaupt einen interaktiven Kurs belegen?

Ein Testverfahren, das alle Testgruppen mit der gleichen Methode prüft, kann zusammengefasst also nicht alle Funktionen und Eigenschaften der einzelnen Evaluanden berücksichtigen. Das bedeutet jedoch, dass sich die Interpretation der Ergebnisse nicht auf diese Aspekte beziehen darf. Die Schlussfolgerung aus diesen Überlegungen lautete also, dass das Evaluationsdesign der Komplexität der Wechselwirkungen entsprechen muss.

Wesentlich für eine realistische Feststellung des Lernerfolgs ist daher die differenzierte Festlegung von Lernzielen, auf deren Basis das Forschungsdesign aufgezogen werden muss. Entspricht das, was in einem Lernsetting intendiert war, auch tatsächlich dem, was letztlich überprüft wurde? An dieser Stelle bot sich die Taxonomie der Lernziele nach Lorin W. Anderson und David R. Krathwohl (2001) an. Sie schien geeignet zu sein, um Lernziele und damit auch den intendierten Lernerfolg zu vergleichen. Von besonderem Interesse für diese Untersuchung waren dabei die kognitiven Lernzielkategorien „Erinnern“, „Verstehen“, „Anwenden“, „Analysieren“, „Bewerten“ und „(Er)Schaffen“, die die kognitiven Abläufe des Lernprozesses beschreiben. Jede Dimension kann dabei in Unterkategorien aufgeschlüsselt werden. So gehören beispielsweise zu „analysieren“ die Subtypen „differenzieren“, „organisieren“ und „kennzeichnen“. Für die weitere Argumentation war dies ein wichtiger Aspekt, da hiermit bereits der erste Schritt zur Operationalisierung der Prozessdimensionen erfolgte.

Im weiteren Verlauf wurden den einzelnen Lernzielen in den Prozessdimensionen Aufgabentypen zugeordnet. Dies diente dazu, deutlich zu machen, dass unterschiedliche Lernziele auch unterschiedliche Arten von Aufgaben erfordern. So kann die Prozessdimension „erinnern“, in der es um die Abfrage von Faktenwissen geht, sehr gut mit einem Multiple-Choice-Test erfasst werden. Hingehen würde man die Fähigkeit des Anwendens und Analysierens kaum mit einem solchen Test messen können. Hier bedarf es beispielsweise eines schriftlichen Aufsatzes.

Die Beschreibung der Aufgabentypen wurde der Ontologie von Norbert Meder & Andrea Frick (2006:75) entnommen und auf die Taxonomie von Lorin W. Anderson und David R. Krathwohl (2001) übertragen. Auf diese Weise wurden einzelnen Lernzielen entsprechende Prüfungsformen zugeordnet, durch welche die Erreichung der Lernziele bewertet werden können. Lernziele und Prüfungsformen müssen sich für eine realistische Bewertung demnach auf der gleichen Ebene der kognitiven Prozessdimension befinden. Die Prüfung muss also äquivalent zum Lernziel sein.

An dieser Stelle wurde die Problematik von Untersuchungen, die Online- mit Präsenzlernen vergleichen, erneut aufgegriffen. Wie bereits festgestellt wurde, ist ein Vergleich deshalb schwierig, da die Leistungsüberprüfung den unterschiedlichen Bedingungen nicht durchgängig gerecht wird. Auch unterscheiden sich die Lernziele der Gruppen oftmals zu stark voneinander. Ein sinnvoller Vergleich von Untersuchungs- und Kontrollgruppe kann also nur möglich sein, wenn Faktoren miteinander verglichen werden, die die selbe Ausprägung von Lernerfolg zu Grunde legen, d.h. deren Lernziele in einer Ebene liegen.

Um zu prüfen, ob diese Überlegungen in der Forschung berücksichtigt werden, wurde eine Meta-Evaluation durchgeführt. Ziel war es, bestehende Vergleichsstudien in Form eines Reviews im Hinblick auf diese Problematik zu analysieren und zu bewerten. Darüber hinaus sollten die Dimensionen des Konstrukts Lernerfolg identifiziert werden, bei denen ein Vergleich möglich ist. Dabei ging es nicht um die Repräsentativität der Ergebnisse, sondern um einen Beleg für indifferente Sichtweisen von Lernerfolg und dessen Operationalisierung. Die Meta-Evaluation sollte keinesfalls dazu dienen, die Qualitäten und Kompetenzen der beteiligten Personen in Frage zu stellen oder einzelne Studien als schlechte Beispiele darzustellen. Vielmehr sollte die Meta-Evaluation einen sehr spezifischen Charakter haben, um die Primärstudien in Bezug auf diese speziellen Kriterien miteinander zu vergleichen. Ein Vergleich von Einzelbefunden war jedoch nicht intendiert.

Nach der Auswahl und Festlegung relevanter Kriterien wurden 11 Primärstudien analysiert und bewertet. Die Ergebnisse zeigten, dass in keiner dieser Studien sichere Aussagen über den Lernerfolg in dem einen oder anderen Lernsetting getroffen werden konnten. Insgesamt wird deutlich, dass der Aspekt der Lernerfolgsmessung in allen untersuchten Studien detaillierter hätte operationalisiert werden müssen. Dies begründet sich vor allem darin, dass die unterschiedlichen Dimensionen des Konstrukts Lernerfolg nicht berücksichtigt worden waren. Viele

Studien bezogen sich in der Formulierung ihrer Lernziele nur auf die niedrigen Stufen der kognitiven Prozesse. Die Messung von Kompetenzen war dabei jedoch nicht möglich. Überdies wurde die Aufmerksamkeit der Evaluation in den meisten Fällen eher auf die Methodik und auf formale Aspekte gerichtet. Es wurde dennoch deutlich, welche Voraussetzungen an zukünftige Evaluationen zu richten sind, die weitere Forschung in diesem Bereich betreiben.

9.1 Schlussfolgerungen

Insgesamt kann gefolgert werden, dass die unterschiedlichen Dimensionen des Konstrukts Lernerfolg i.d.R. nicht berücksichtigt worden sind. Gerade die Einordnung in die Dimension der kognitiven Prozesse hat gezeigt, dass es nicht DEN Lernerfolg gibt, sondern dass das Konstrukt und die Verflechtung der einzelnen Dimensionen eine große Rolle spielt.

Dies wurde insbesondere in den Prüfungen deutlich. In den untersuchten Primärstudien wurde der Lernerfolg in den meisten Fällen mittels der gleichen Tests für beide Untersuchungsgruppen gemessen. Dies hatte zur Folge, dass die getesteten Lernziele zwar auf derselben Ebene der kognitiven Prozesse lagen, andererseits konnte dadurch, dass das Prüfungsdesign nicht variierte, auch nur ein kleiner Teil aller relevanten Eigenschaften der einzelnen Evaluanden erfasst werden. Da der Lernerfolg fast durchgängig als Behaltensleistung definiert wurde, fanden andere Aspekte dieses Konstrukts – wie Lernprozess oder Kompetenz – meist keine Beachtung.

Die meisten Online-Lernsettings waren ferner nur an Zielen der niedrigen Prozessdimensionen ausgerichtet, wie „erinnern“ oder „verstehen“, während beispielsweise Gruppenaktivitäten im Klassenraum mit höheren Prozessen erfasst wurden. Damit zeigt sich erneut die Schwierigkeit, die einzelnen Evaluanden in ihrer Gesamtheit zu erfassen. Denn selbstverständlich könnten auch E-Learning-Angebote in den oberen Prozessdimensionen angesiedelt werden. So entspräche beispielsweise die gemeinsame Arbeit in Wikis dem weitaus höheren Prozess „bewerten“, wenn die Arbeit anderer kritisiert und eingeschätzt werden soll.

Zusammengefasst kann angenommen werden, dass die Aufmerksamkeit der Evaluation in den meisten Fällen eher auf die Methodik sowie auf formale Aspekte gerichtet war, jedoch die Frage nach den Lernzielen und der bestmöglichen Operationalisierung von Lernerfolg vernachlässigt wurde. Unter anderem zeigt sich dies dadurch, dass die Studien, die in den formalen Kriterien positiv bewertet wurden, oftmals die für die Forschungsfrage relevanten Kriterien weniger gut erfüllen konnten.

Auch in Bezug auf die Formulierung von konkreten Lernzielen wurden in den Primärstudien unterschiedliche Vorgehensweisen aufgezeigt. Offenkundig ist, dass nur in sehr wenigen Fällen überhaupt Lernziele benannt worden waren. M.E. waren sich viele EvaluatorInnen der

Relevanz der Lernziele nicht bewusst. Es ist auffällig, dass oftmals zwar die gleichen Lernziele für beide Gruppen intendiert waren, für die Prüfung allerdings ganz andere Aspekte gezeigt und erbracht werden mussten. Die größte Schwierigkeit liegt m.E. also in der Entsprechung von intendierten Lernzielen und Prüfungsform, welche oft nicht erreicht werden konnte. Wie bereits erwähnt, muss die Prüfung äquivalent zum Lernziel sein. Wenn dies schon innerhalb einer Untersuchungsgruppe nicht gewährleistet ist, ist es umso weniger sinnvoll, auch noch einen Vergleich mit einer anderen Untersuchungsgruppe anzustellen. Daraus muss gefolgert werden, dass unspezifische Vergleiche von Online- versus Präsenzlernen nicht sinnvoll anwendbar bzw. nicht generalisierbar sind und keine verallgemeinerbaren Rückschlüsse auf den Lernerfolg zulassen. Es leuchtet ein, dass bei der Evaluation einer Maßnahme auch eine gute Vergleichbarkeit der Gruppen gegeben sein muss. Dabei wird jedoch oft nur die empirische Perspektive berücksichtigt und nicht die didaktische.

Aus den dargestellten Gründen können keine verlässlichen Aussagen über die Lernqualität und den Lernerfolg in den untersuchten Settings getroffen werden. Die Frage, welches Setting besser für das Lernen geeignet ist, lässt sich jedoch mit einer Gegenfrage beantworten: „Besser wofür?“

9.2 Empfehlungen

Aus der bisherigen Argumentation ist bereits deutlich geworden, dass die Messung von Lernerfolg nicht problemlos möglich ist, aber auch zu relativistischen Positionen wenig sinnvoll sind. So kann an dieser Stelle nicht der Schluss gezogen werden, dass überhaupt nichts mehr gemessen werden könne. Auch die an einer früheren Stelle gezogene Folgerung, dass sich EvaluatorenInnen der Wechselwirkungen der untersuchten Konstrukte bewusst sein müssen (vgl. Annabell Preussler & Peter Baumgartner 2006:81), ist nicht mehr weitreichend genug.

Wesentlich ist, dass das Untersuchungsdesign den einzelnen Konstruktdimensionen entsprechen muss. Darüber hinaus müssen diese Dimensionen zusätzlich in ihren Wechselwirkungen analysiert werden. Dies muss jedoch unter Berücksichtigung der unterschiedlichen Lernsettings geschehen.

Es wäre also zu kurzfristig, zu behaupten, dass sich Lernerfolg nicht operationalisieren lasse. Der Punkt ist, dass der Lernerfolg in der Praxis durchaus operationalisiert wird, nur nicht ausreichend detailliert.

Die Schlussfolgerung muss also lauten: Um sinnvolle Vergleiche zwischen Präsenz- und Onlinelernen anstellen zu können, müssen die Untersuchungen bei der Operationalisierung bestimmte Kriterien erfüllen. Eine Lösung des Problems liegt in der Angemessenheit von Lernziel, Lernerfolgsprüfung und Lernsetting. Um die Evaluanden miteinander vergleichen zu können, müssen in beiden Untersuchungsgruppen die gleichen Lernziele formuliert, also die gleichen Ausprägungen von Lernerfolg zu Grunde gelegt werden. Dabei muss das Prüfungsdesign den

intendierten Lernzielen entsprechen. Zusätzlich müssen auch die Lernsettings auf dieser Ebene vergleichbar sein.

Konkret bedeutet das: Für einen sinnvollen Vergleich des Lernerfolgs beim Online- bzw. Präsenzlernen sind also drei Aspekte nötig:

1. Die Lernziele der Vergleichsgruppen müssen sich in einer Ebene der kognitiven Prozessdimension befinden.
2. Die Prüfungen müssen adäquat zu den Lernzielen sein.
3. Das Prüfungsdesign der Untersuchungsgruppen muss den unterschiedlichen Lernsettings entsprechen.

Nur wenn alle drei Faktoren erfüllt sind, ist ein Vergleich überhaupt möglich. In Tabelle 29 ist eine optimale Vergleichssituation dargestellt.

DIE KOGNITIVE PROZESS-DIMENSION					
1. Erinnern	2. Verstehen	3. Anwenden	4. Analysieren	5. Bewerten	6. (Er)schaffen
	Lernziele Online Lernziele Präsenz Tests				

Tabelle 29: Beispiel zur Einordnung einer optimalen Studie in die Dimension der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)

Wichtig ist jedoch, dass keine Aussagen über die einzelnen Dimensionen getroffen werden, sondern nur über das Gesamtgefüge aus Lernziel, Lernzielprüfung und Lernsetting bzw. Lernsituation. Die alten Worte Aristoteles'¹⁰⁰ „Das Ganze ist mehr als die Summe seiner Teile“ sind also nach wie vor gültig, da wir es mit Konstrukten zu tun haben, die sich durch verschiedene, aber voneinander abhängige Dimensionen auszeichnen. Jedoch ist das Gesamtgefüge mehr als sich durch einzelne Dimensionen erfassen lässt. Nicht dessen Komplexität stellt ein Hindernis dar, sondern vor allem die Verflechtung der Faktoren. Eine empirische Erfassung des komplexen Konstrukts „Lernerfolg“ ist also nicht möglich, wenn nur einzelne Dimensionen herausgelöst betrachtet werden.

Eine weitere Folgerung daraus ist, dass durchaus legitim ist, Lernerfolg als Behaltensleistung zu definieren. Allerdings sind dabei zwei Überlegungen relevant: Erstens wird auf diese Weise nur eine Dimension von Lernerfolg dargestellt, nämlich „erinnern“. Es wird jedoch vernachlässigt, dass es noch andere, wichtigere Dimensionen gibt, die aber nicht untersucht werden. Zweitens muss beachtet werden, dass sich auch die Lernziele und die Tests auf diese Ebene beziehen. Ein Vergleich

¹⁰⁰ 384 – 322 v.Chr.

mit einer anderen Untersuchungsgruppe ist nur dann möglich, wenn diese dieselben Eigenschaften aufweist.

In einem anderen Fall können Lernziele auf der Ebene des Bewertens formuliert werden. Solange sich auch die Tests auf diese Ebene beziehen, kann ein Vergleich mit einer anderen Untersuchungsgruppe angestellt werden, die diese Voraussetzungen ebenfalls erfüllt. Es geht also um die Betrachtung von zwei ganzheitlichen Settings und nicht um das herausgelöste Analysieren einzelner Dimensionen.

Für die Weiterentwicklung dieses Forschungsgebietes halte ich diese Aussicht für relevant. Zukünftig wären Untersuchungen sinnvoll, in denen die hier argumentativ aufgezeigten Bedingungen berücksichtigt werden. Die auf diese Weise erzielten Ergebnisse wären – da detaillierter operationalisiert – wesentlich wegweisender für die Frage, welches Lernsetting für welche Lernziele (!) nun tatsächlich das bessere sei.

10 Literatur

- Adomßent, Maik (2001). *Realisierungspotentiale kooperativen Lernens mit Hilfe virtueller Kommunikation*. In Lück, H. (Hrsg.) *Gruppendynamik und Organisationsberatung*. 32. Jahrg., Heft 2, S.161-175.
- American Evaluation Association (AEA) (1994). *Guiding principles for evaluators*. In: American Evaluation Association: *Evaluation Practice News*, June 1994.
- Anderson, Lorin W. & Krathwohl, David R. (2001). *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*. Addison Wesley Longman.
- Anderson, Scarvia & Ball, Samuel (1978). *The prozession and practice of program evaluation*. San Francisco. Jossey-Bass.
- Aufenanger, Stefan (2001). *Aufgaben der Erziehungswissenschaft in der Wissensgesellschaft*. In: Herzig, Bardo (Hrsg.). *Medien machen Schule. Grundlagen, Konzepte und Erfahrungen zur Medienbildung*. Bad Heilbrunn/Obb: Klinkhardt. S. 255-266.
- Bartos, Thomas (2004). *Selbstgesteuertes und kooperatives Lernen mit Neuen Medien - Eine longitudinale Evaluation einer webbasierten Lernumgebung zur deskriptiven Statistik* [WWW-Dokument unter <http://deposit.fernuni-hagen.de/volltexte/2006/41/>].
- Baumgartner, Peter (1997). *Evaluation vernetzten Lernens: 4 Thesen*. In: Wagner, Erwin & Kindt, Michael (Hrsg.). *Virtueller Campus. Forschung und Entwicklung für neues Lehren und Lernen*. Münster: Waxmann. S. 131-146.
- Baumgartner, Peter (1999). *10 Todsünden in der Evaluation interaktiver Lehr – und Lernmedien*. In: Lehmann, Klaus (Hrsg.) *Studieren 2000 - Alte Inhalte in neuen Medien?* Münster: Waxmann. S. 199-220.
- Baumgartner, Peter (1999a). *Evaluation mediengestützten Lernens. Theorie - Logik - Modelle*. In: Kindt, Michael (Hrsg.). *Projektevaluation in der Lehre - Multimedia an Hochschulen zeigt Profile*. Medien in der Wissenschaft, Bd. 7. Münster: Waxmann. S. 61-97.

- Baumgartner, Peter (2006). *Unterrichtsmethoden als Handlungsmuster - Vorarbeiten zu einer didaktischen Taxonomie für E-Learning*. [WWW-Dokument unter <http://www.peter.baumgartner.name/article-de/handlungsmuster-taxonomiepdf.pdf> , letzter Aufruf: 31.07.2007].
- Baumgartner, Peter & Frank, Stefan (2000). *Der Mediendidaktische Hochschulpreis (MEDIDA-PRIX) – Idee und Realisierung*. In: *Campus 2000: Lernen in neuen Organisationsformen*. Medien in der Wissenschaft, Bd. 10. Münster: Waxmann. S. 63-82.
- Baumgartner, Peter & Preussler, Annabell (2005). *E-Learning-Qualität: Evaluation und Assessment*. Studienbrief für Modul 8 des Weiterbildungsstudiengangs „Technologiebasiertes Lehren und Lernen“ Virtueller Campus Rheinland-Pfalz.
- Beelmann, Andreas & Bliesener, Thomas (1994). *Aktuelle Probleme und Strategien der Metaanalyse*. Psychologische Rundschau, 45, S. 211-233.
- Beethoven-Haus Bonn (o.J.). *Digitales Archiv*. Friedrich Oelenhainz (1745-1804), Franz Joseph Maximilian Fürst von Lobkowitz (1772-1816) - Fotografie eines wohl von Friedrich Oelenhainz stammenden Ölgemäldes. Beethoven-Haus Bonn, Ley, Band I, Nr. 173 [http://www.beethoven-haus-bonn.de/sixcms/detail.php?id=2452&template=dokseite_digitales_archiv_de&_eid=2235&_ug=Lieder%20und%20Ges%203%20a4nge&_werkid=99&_dokid=i3113&_opus=op.%2098&_mid=Werke%20Ludwig%20van%20Beethovens&_seite=1 letzter Aufruf 04.12.2006].
- Bildungskommission NRW (1995). *Zukunft der Bildung – Schule der Zukunft*. Neuwied: Luchterhand.
- Bloom, Benjamin S. (1956). *Taxonomy of Educational Objectives. Handbook I: Cognitive Domain*. New York.
- Bloom, Benjamin S., (Hg.) (1972). *Taxonomie von Lernzielen im kognitiven Bereich*. 4. Auflage. Weinheim und Basel, Beltz.
- Blumschein, Patrick (2004). *Eine Metaanalyse zur Effektivität multimedialen Lernens am Beispiel der Anchored Instruction*. Dissertation in der Wirtschafts- und verhaltenswissenschaftlichen

Fakultät. Freiburg, Albert-Ludwigs Universität Freiburg im Breisgau.

- Boblin, Sheryl L. & Barbara A. Gibson (1986). *Computer managed instruction at the University of Alberta Hospitals School of Nursing*. Unpublished master's project report, University of Alberta. Edmonton.
- Bonsen, Martin & Büchter, Andeas (2005): *Sozialwissenschaftliche Forschungsmethoden für Schulevaluation*. Studienbrief für den postgradualen Fernstudiengang "Schulmanagement" an der Technischen Universität. Kaiserslautern: Zentrum für Fernstudien und Universitäre Weiterbildung 2005.
- Bortz, Jürgen & Döring, Nicola (1995). *Forschungsmethoden und Evaluation für Sozialwissenschaftler*. 2. vollständig überarbeitete und aktualisierte Auflage. Berlin: Springer.
- Bower, Gordon H. & Hilgard, Ernest R.(1983). *Theorien des Lernens*. Stuttgart: Klett-Cotta.
- Bührlen, Bernhard (1999). *Metaevaluation und Metaanalyse*. Script einer Vorlesung im Sommersemester 1999 an der Universität Freiburg. [WWW Dokument unter: <http://home.debitel.net/user/buehrle/downloads/Metascr.pdf> , letzter Aufruf: März 2005].
- Bührlen-Armstrong, Bernhard (1998). *Evaluationsforschung und Qualitätssicherung in der Personalkommunikativen AIDS-Prävention. Eine Meta-Evaluation*. Inaugural-Dissertation zur Erlangung der Doktorwürde der Philosophischen Fakultäten der Albert-Ludwigs-Universität zu Freiburg i. Br. WS 1997/1998.
- Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung, D., W. u. K. Bundesministerium für Bildung, Österreich, et al., Eds. (2002). *Lernen in der Wissensgesellschaft*. Innsbruck: StudienVerlag.
- Carey Jane M. (2001). *Effective Student Outcomes: A Comparison of Online and Face-to-Face Delivery Modes*. In: DEOSNEWS Vol.11. [WWW-Dokument unter http://www.ed.psu.edu/acsde/deos/deosnews/deosnews11_9.asp, letzter Aufruf: 11.07.2007].

- Christmann, Edwin P.; Badgett, John L. & Lucking, Robert (1997). *Progressive Comparison of the Effects of Computer-Assisted Instruction on the Academic Achievement of Secondary Students*. Journal of Research on Computing in Education, 29(4), S. 325-337.
- Clark, Richard E. (1983). *Reconsidering Research on Learning from Media*. Review of Educational Research, 53(4), S. 445-459.
- Clark, Richard E. (1994a). *Media will Never Influence Learning*. Educational Technology Research and Development, 42(2), S. 21-29.
- Clark, Richard E. (1994b). *Media and Method: The Media Influence Debate: Read the Fine Print, but don't Lose the Sight of the Big Picture*. Educational Technology, Research & Development, 42(2), S. 7-10.
- Clement, Ute & Martens, Bernd (2000). *Effizienter Lernen durch Multimedia? Probleme der empirischen Feststellung von Ursachen des Lernerfolgs*. Zeitschrift für Pädagogik, Vol. 46, No. 1, Januar/Februar, S. 97–112.
- Cook, Thomas D. & Gruder, Charles L. (1978). *Metaevaluation research*. Evaluation Quarterly, 2, S. 5-51.
- Corno, Lyn & Snow, Richard E. (1986). *Adapting teaching to individual differences among learners*. In: Wittrock, M. C. (Ed.): *Handbook of research on teaching*. New York: Macmillan. S. 605-629.
- Davies, Randall S. & Mendenhall, Robert (1998). *Evaluation comparison of online and classroom instruction for HEPE 129—Fitness and lifestyle management course*. Department of Instructional Psychology and Technology, Brigham Young University, Utah, (ERIC Document Reproduction Service ED 427 752).
- Degele, Nina (2000). *Informiertes Wissen. Eine Wissenssoziologie der computerisierten Gesellschaft*. Frankfurt am Main: Campus.
- Deutsche Gesellschaft für Evaluation e.V. (2001): *Standards für Evaluation*. Geschäftsstelle DeGEval. Köln.
- Deutsche Gesellschaft für Evaluation e.V. (2002): *Angefragte Gutachten / Meta-Evaluationen auf Basis der DeGEval-Standards*. [WWW-

Dokument unter <http://www.degeval.de/calimero/tools/proxy.php?id=63> , letzter Aufruf 01.05.2007].

de Witt, Claudia & Czerwionka, Thomas (2007). *Studienbuch Mediendidaktik*. Studentexte der Erwachsenenbildung. (DIE). Bielefeld: Bertelsmann Verlag.

Döring, Nicola (1997). *Lernen mit dem Internet*. In L. J. Issing & P. Klimsa (Hrsg.). *Information und Lernen mit Multimedia*. 2. Aufl. Weinheim: Beltz. S. 305-336.

Ehlers, Ulf-Daniel (2002). *Qualität beim eLearning. Der Lernende als Grundkategorie der Qualitätssicherung*. In: DGfE (Hrsg.). Medienpaed.com – Onlinezeitschrift für Theorie und Praxis der Medienbildung. Ausgabe 01/2002. [WWW-Dokument, <http://www.medienpaed.com/02-1/ehlers1.pdf> , letzter Aufruf 07.06.2005].

Ehlers, Ulf-Daniel (2004): *Qualität im E-Learning aus Lernalternsicht*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Eisend, Martin (2004). *Metaanalyse. Einführung und kritische Diskussion*. Diskussionsbeiträge des Fachbereichs Wirtschaftswissenschaft der Freien Universität Berlin. Betriebswirtschaftliche Reihe, 2004/8 [WWW-Dokument unter http://www.wiwiiss.fu-berlin.de/files/K6UAD7B/Metaanalyse_DP.pdf, letzter Aufruf 24.04.2007].

Eriksson, Thommy; Goller, Alois; Muchin, Sergei (2001). *A comparison of online communication in distance education and in conventional education*. 31st Annual Frontiers in Education Conference, Impact on Engineering and Science Education. Conference Proceedings IEEE Part vol .1, 2001, T2F-20-5 vol. 1. Piscataway, NJ, USA.

Erpenbeck, John und Rosenstiel, Lutz von (Hrsg.) (2003). *Handbuch Kompetenzmessung. Erkennen, Verstehen und Bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis*. Stuttgart, Schäffer-Poeschel.

Euler, Dieter (1994). *(Multi)mediales Lernen - Theoretische Fundierungen und Forschungsstand*. Unterrichtswissenschaft, 22. S. 291-311.

Evaluation Research Society Standards Committee (1982). *Evaluation Research Society standards for program evaluation*. In: Peter Rossi

- (Hrsg.). *Standards for evaluation practice*. New Directions for Program Evaluation No. 15. San Francisco.
- Eysenck, Hans J. (1978). *An exercise in mega-silliness*. American Psychologist, 33. S.517.
- Farin, Erik (1994). *Forschungsperspektive und Methodik der Metaanalyse*. Forschungsberichte des Psychologischen Instituts der Albert-Ludwigs-Universität Freiburg i. Br., Nr. 113. Freiburg i. Br.: Abteilung für Rehabilitationspsychologie der Univ. Freiburg i. Br.
- Fisch, Martin & Gscheidle, Christoph (2006). *Onliner 2006: Zwischen Breitband und Web 2.0 – Ausstattung und Nutzungsinnovationen. Ergebnisse der ARD/ZDF-Online-Studien 1997-2006*. In: MediaPerspektiven 8/2006. S. 431-440.
- Fletcher-Flinn, Claire M. & Gravatt, Breon (1995). *The Efficacy of Computer Assisted Instruction (CAI): A Meta-Analysis*. Educational Computing Research, 12(3). S. 219-242.
- Freudenreich, Martin & Schulte, Carsten (2002). *Von der Evaluation von Lernsoftware zur Gestaltung von Unterricht*. MedienPädagogik. Online-Zeitschrift für Theorie und Praxis der Medienbildung, 18.3.2002 [WWW-Dokument unter http://ddi.uni-paderborn.de/didaktik/Veroeffentlichungen/freudenreich_schulte.pdf , letzter Aufruf 23.02.2007].
- Fricke, Reiner & Treinies, Gerhard (1985). *Einführung in die Metaanalyse*. Bern: Huber.
- Fricke, Reiner (1991). *Zur Effektivität computer- und videounterstützter Lernprogramme*. Braunschweig: Arbeiten aus dem Seminar für Pädagogik der Technischen Universität.
- Fricke, Reiner (1999). *Anmoderation zu Evaluation*. In: Lehmann, Klaus (1999) *Studieren 2000 - Alte Inhalte in neuen Medien?* Münster: Waxmann. S. 195-197.
- Fricke, Reiner (2000). *Qualitätsbeurteilung durch Kriterienkataloge. Auf der Suche nach validen Vorhersagemodellen*. In: Schenkel, P., Tergan, S.-O. & Lottmann, A. (Hrsg.): *Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand*. Reihe multimediales Lernen in der Berufsbildung. Nürnberg: BW-Verlag.

- Glass, Gene V. & McGraw, Berry et al. (1981). *Meta-analysis in social research*. Beverly Hills: Sage Publications.
- Glass, Gene V. (1976). *Primary, secondary and meta-analysis of research*. Educational Researcher, 5. S.3-8.
- Green, Ruth & Gentemann, Karen (2001). *Technology in the curriculum: An assessment of the impact of on-line courses*. [WWW-Dokument unter <http://assessment.gmu.edu/Results/Other/2000/Eng302/Eng302Report.pdf> , letzter Aufruf 22.09.2007].
- Hartzoulakis, Vassilis (2002). *Investigating Computer Mediated Instruction in a Greek State Lykeio*. Unpublished Master's Thesis. Hellenic Open University. [WWW-Dokument unter <http://users.att.sch.gr/hartzoulakis/Dissertation.pdf> , letzter Aufruf 27.01.2007].
- Harvey, Lee & Green, Diana (2000). *Qualität definieren. Fünf unterschiedliche Ansätze*. Zeitschrift für Pädagogik, 41. Beiheft. S 17-40.
- Hasebrook, Joachim (1995). *Lernen mit Multimedia*. Zeitschrift für Pädagogische Psychologie 9(2). S. 95-103.
- Hauf, Oliver (1996). *Die Informationsgesellschaft: Anatomie einer Lebenslüge*. Demokratie, Ökologie, Föderalismus, Bd. 11. Frankfurt am Main: P. Lang.
- Heyer, Susanne (2006). *Didaktische Szenarien und deren Verhältnis zu Lernmaterialien*. Forschungsberichte des Fachbereichs Elektrotechnik & Informationstechnik. 1/2006. [WWW-Dokument unter http://www.fernuni-hagen.de/etit/forschung/Forschungsbericht_1_2006.pdf , letzter Aufruf 07.02.2007].
- Hunneshagen, Heike; Schulz-Zander, Renate & Weinreich, Frank (2000). *Schulen am Netz. Veränderung von Lehr- und Lernprozessen durch den Einsatz Neuer Medien*. In: Rolff, Hans-Günter; Bos, Wilfried; Klemm, Klaus; Pfeiffer, Hermann & Schulz-Zander, Renate (Hrsg.): *Jahrbuch der Schulentwicklung*. Band 11. Weinheim: Juventa. S. 155-180.
- Jacobs, Bernhard (1998). *Einführung in die Versuchsplanung*. [WWW-Dokument unter <http://www.phil.uni-sb.de/~jakobs/seminar/vpl/> , letzter Aufruf 21.03. 2005].

- Jelitto, Marc (2006). *Der Prototyp eines Lernziel-Editors basierend auf der Taxonomie von Anderson & Krathwohl*. [WWW-Dokument unter http://blog.cc.fernuni-hagen.de/uploads/cc_Lernziel_Editor_Prototyp_2006_12_19.pdf , letzter Aufruf 16.02.2007].
- Jenks, Michael S. & Springer, John M. (2002). *A View of the Research on the Efficacy of CAI*. Idaho State University, College of Education: Electronic Journal for the Integration of Technology in Education. Vol. 1, no. 2. S. 43-58.
- Joint Committee on Standards for Educational Evaluation (1981). *Standards for evaluations of educational programs, projects and materials*. New York: New York, N.Y.: McGraw-Hill Book Co.
- Jones, Edward R. (1999). *A comparison of an all web-based class to a traditional class*. Paper presented at the Society for Information Technology & Teacher Education, San Antonio. [WWW-Dokument unter <http://falcon.tamucc.edu/~ejones/papers/site99.pdf> , letzter Aufruf: 01.08.2007].
- Keller, Franziska (2004). *Individualisierung von Assessments*. Diplomarbeit im Fach Informatik. Angefertigt am Institut für Informatik der Universität Zürich. [WWW-Dokument unter http://www.ifi.unizh.ch/archive/mastertheses/DA_Arbeiten_2004/Keller_Franziska.pdf , letzter Aufruf: 07.02.2007].
- Kerres, Michael (2001). *Multimediale und telemediale Lernumgebungen. Konzeption und Entwicklung*. 2. Auflage, München: Oldenbourg Verlag.
- Kerres, Michael; de Witt, Claudia & Stratmann, Jörg (2002). *E-Learning. Didaktische Konzepte für erfolgreiches Lernen*. In: Schwuchow, Karlheinz & Guttman, Joachim (Hrsg.): *Jahrbuch Personalentwicklung & Weiterbildung*. Neuwied: Luchterhand [WWW-Dokument unter <http://mediendidaktik.uni-duisburg-essen.de/system/files/jahrb-pe-wb-b.pdf> , letzter Aufruf: 18.09.2007].
- Köbberling, Almut (1971). *Effektiveres Lehren durch Programmierten Unterricht?* Weinheim: Beltz.
- Kot, Pauline, Skillen, Lynn D., & Wales, Marilyn (1986). *Learning effectiveness and efficiency using computer managed learning and lecture for health assessment theory*. Final report. Edmonton, AB: University of Alberta.

- Kozma, Robert B. (1994a). „*Will media influence learning? Reframing the debate.*“ Educational technology research and development, 42, 2/1994. S. 7-19.
- Kozma, Robert B. (1994b). *The Influence of Media on Learning: The Debate Continues*. In: School Library Media Research (SLMQ) Volume 22, Number 4.
[WWW-Dokument unter <http://www.ala.org/ala/aasl/aaslpubsandjournals/slmrb/editorschoiceb/infopower/selectkozmahtml.htm> , letzter Aufruf 23.02.2007].
- Krathwohl, David R. (2002). *A Revision of Bloom's Taxonomy: An Overview*. Theory into practice, Volume 41, Number 4.
- Krathwohl, David R.; Bloom, Benjamin S. & Masia, Betram B. (1978). *Taxonomie von Lernzielen im affektiven Bereich*. Weinheim: Beltz.
- Krempf, Stefan (2005). Wissen von unten. Süddeutsche Zeitung, 04.02.2005. [WWW-Dokument unter <http://www.sueddeutsche.de/trt3m1/computer/artikel/234/47187/> , letzter Aufruf 25.01.2007].
- Kübler, Hans-Dieter (1999). *Medienkompetenz – Dimension eines Schlagwortes*. In: Schell, Fred; Stolzenburg, Elke & Theunert, Helga (Hrsg.) (1999). *Medienkompetenz. Grundlagen und pädagogisches Handeln*. Reihe Medienpädagogik. München: KoPäd. S. 25-47.
- Kulik, Chen-Lin C. & Kulik, James A. (1991). *Effectiveness of Computer-based Instruction: An Updated Analysis*. Computers in Human Behaviour, 7. S. 75-94.
- Kulik, Chen-Lin C.; Kulik, James A. & Shwalb, Barbara (1986). *The effectiveness of computer-based adult education: a meta-analysis*. Journal of Educational Computing Research, 2. S. 235-252.
- Kulik, James A. (1994). *Meta-analytic studies of findings of computer-based instruction*. In Baker, E. L. and O'Neil, H. F. Jr. (Eds.), *Technology assessment in education and training*. Hillsdale, NJ: Lawrence Erlbaum. S. 9-33.
- Kulik, James A. and Kulik, Chen-Lin (1989). *Meta-analysis in education*. International Journal of Educational Research 13. S. 221-340.

- Mandl, Heinz & Reinmann-Rothmeier, Gabi (1997). *Medienpädagogik und -kompetenz: Was bedeutet das in einer Wissensgesellschaft und welche Lernkulturen brauchen wir dafür?* In: *Medienkompetenz im Informationszeitalter*. Schriftenreihe "Enquete-Kommission: Zukunft der Medien", Bd. 4. Bonn: ZV.
- Mandl, Heinz; Reinmann-Rothmeier, Gabi & Gräsel, Cornelia (1998). *Gutachten zur Vorbereitung des Programms „Systematische Einbeziehung von Medien, Informations- und Kommunikationstechnologien in Lehr-Lern-Prozesse“*. Materialien zur Bildungsplanung und zur Forschungsförderung, Heft 66. Bonn: BLK.
- Mandl, H. & Hron, A. (1985). *Förderung kognitiver Fähigkeiten und des Wissenserwerbs durch computerunterstütztes Lernen*. In U. Bosler, K. Frey, W. Hosseus, M. Kremer, P. Schermer & H. Wolgast (Hrsg.), *Mikroelektronik und Neue Medien im Bildungswesen*. Referate des Kieler Seminars vom 12. bis 16. November 1984. IPN-Arbeitsberichte. Kiel: IPN.
- Martens, Rob; Portier, Stanley J. & Valcke, Martin (1995). *Comparing study outcomes from different learning environments for statistics* (Vol. 65). Heerlen, Netherlands: Open University Center for educational Technological Innovation. ERIC Document Reproduction Service No. 383 745.
- Mayring, Philipp (1993). *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. 4., erweiterte Auflage. Weinheim: Beltz.
- Meder, Norbert & Frick, Andrea (2006). *Aufgaben beim Lernen im Internet*. In: Meder, Norbert (2006). *Web-Didaktik. Eine neue Didaktik webbasierten, vernetzten Lernens*. Bielefeld: Bertelsmann. S. 71-80.
- Middle States Commission on Higher Education (2001). *Characteristics of Excellence: Standards for Accreditation*. MS Draft for Discussion. Philadelphia, PA.
- Neisser, Ulric (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.
- Niegemann, Helmut M. (1995). *Computergestützte Instruktion in Schule, Aus- und Weiterbildung: theoretische Grundlagen, empirische Befunde und Probleme der Entwicklung von Lehrprogrammen*. Frankfurt am Main: P. Lang.

- Phipps, Ronald A. & Merisotis, Jamie (1999). *Quality on The Line: Benchmarks for Success in Internet-Based Distance Education*. Washington, D.C. Institute for Higher Education Policy.
- Postman, Neil (1985). *Wir amüsieren uns zu Tode. Urteilsbildung im Zeitalter der Unterhaltungsindustrie*. Frankfurt am Main: S. Fischer.
- Preussler, Annabell (2002). *Neue Medien und erziehungswissenschaftliche Innovationsprozesse in Schulen. Empirische Befunde der zweiten Koordinatorenbefragung „Schulen ans Netz“*. Diplomarbeit im Fachbereich Erziehungswissenschaften und Soziologie der Universität Dortmund.
- Preussler, Annabell (2004). *Über die Notwendigkeit des (geschlechter)gerechten Ausdrucks*. In: MaDonna. Zeitschrift der Gleichstellungsbeauftragten der Universität Dortmund.
- Preußler, Annabell (i. Vorb.). *Anlage und Durchführung einer empirischen Studie*. In: Baumgartner, Peter & Payr, Sabine (Hrsg.). *Studieren und Forschen mit dem Internet. Ein Handbuch zur Methodik und Technik des wissenschaftlichen Arbeitens*.
- Preussler, Annabell & Baumgartner, Peter (2006). *Qualitätssicherung in mediengestützten Lernprozessen – Zur Messproblematik von theoretischen Konstrukten*. In: Sindler, Alexandra (Hrsg.): *Qualitätssicherung im eLearning*. Reihe Medien in der Wissenschaft. Münster: Waxmann. S. 73-85.
- Pucel, David J. & Stertz, Thomas F. (2005). *Effectiveness of and Student Satisfaction with Web-Based Compared to Traditional In-service Teacher Education Courses*. Journal of Industrial Teacher Education, Volume 42, No. 1, S. 7-23.
- Pütz, Sebastian (2006). *E-Learning 2.0 - Buzzword oder ernstzunehmende Entwicklung?* [WWW-Dokument unter http://www.fortbildung-bw.de/wb/09_bildungsanbieter/05_e-learning/zwei_null.php?lvl=2455 , letzter Aufruf 16.02.2007].
- Richter, Sigrun (2002). *Die "kognitive Wende" in der Deutschdidaktik*. In: Hug, Michael & Richter, Sigrun (Hrsg.). *Ergebnisse soziologischer und psychologischer Forschung: Impulse für den Deutschunterricht*. Baltmannsweiler: Schneider-Verlag Hohengehren. S. 1-10.

- Rolff, Hans-Günter (2001). *Schulentwicklung konkret: Steuergruppe, Bestandsaufnahme, Evaluation*. Seelze: Kallmeyer.
- Rosenthal, Robert (1979). *The „file drawer problem“ and tolerance for null results*. Psychological Bulletin, 86. S.638-641.
- Russell, Thomas L. (1999). *The no significant difference phenomenon: as reported in 355 research reports, summaries, and papers: a comparative research annotated bibliography on technology for distance education*. [Raleigh, N.C.]: North Carolina State University.
- Rossi, Peter H. & Freeman, Howard E. (1989). *Evaluation. A Systematic Approach*. Newbury Park: Sage.
- Rossi, Peter H. (Ed.)(1982). *Standards for Evaluation Practice. New directions for program evaluation*. San Francisco: Jossey-Bass.
- Rossi, Peter H.; Freeman, Howard E. et al. (1988). *Programm-Evaluation. Einführung in die Methoden angewandter Sozialforschung*. Stuttgart: Enke.
- Rustenbach, Stephan Jeff (2003). *Metaanalyse. Eine anwendungsorientierte Einführung*. Bern: Huber.
- Ryan, William J. (2002). *Online and in the classroom: The numbers and what they might mean*. Paper presented at the League for Innovation in the Community College Conference, Boston. ERIC Document Reproduction Service No. ED 467 851.
- Salomon, Gavriel (2000). *It's not just the Tool, but the Educational Rationale that Counts*. EdMedia Meeting. Montreal [WWW-Dokument unter <http://www.aace.org/conf/edmedia/00/salomonkeynote.htm> , letzter Aufruf 17.09.2007].
- Sanders, James R. (Hrsg.) ,& Beywl, Wolfgang (dtsch. Übersetzung) (2005). *Handbuch der Evaluationsstandards: die Standards des "Joint Committee on Standards for Educational Evaluation"*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Schacter, John & Fagnano, Cheryl (1999). *Does computer technology improve student learning and achievement? How, when, and under what conditions?* Journal of Educational Computing Research, 20(4). S. 329-343.

- Schobel, Kurt & von Holdt, Ulrike (2004). *Entwicklung und Erprobung eines integrierten Leistungspunktsystems in der Weiterbildung modularisierter Studienangebote am Beispiel der Ingenieurwissenschaften. AP2: Verifikation von Qualifikation in einem Leistungspunktsystem*. Abschlussbericht des BLK-Projekts. Förderkennzeichen M 1605.05. Universität Hannover.
- Schroedel-Verlag (2001). *Arbeitsblatt „Gleichheit oder Fairness? - Die Karikatur von Traxler“*. schroedel aktuell (03/2001), S. 4.
- Schulmeister, Rolf; Vollmers, Burkhard; Gücker, Robert; Nuyken, Klaus (2005). *Konzeption und Durchführung der Evaluation einer virtuellen Lernumgebung: Das Projekt Methodenlehre-Baukasten*. In: Bachmair, Ben; Diepold, Peter & de Witt, Claudia (Hrsg.). *Jahrbuch Medienpädagogik Bd. 5*. Wiesbaden: VS, Verlag für Sozialwissenschaften. S. 37-52.
- Schulz-Zander, Renate (1997). *Lernen in der Informationsgesellschaft*. In: Zeitschrift für Pädagogik 3/97. S. 8-12.
- Schulz-Zander, Renate (2000). *Lernen mit Neuen Medien. Medienkompetenz als Schlüsselqualifikation*. In: Apflauer, R./ Reiter, A. (Hrsg.). *Schule online. Das Handbuch zum Bildungsmedium Internet*. Wien: Public-Voice. 105-115.
- Schutte, Jerald G. (1997). *Virtual teaching in Higher Education: The new Intellectual Superhighway or Just Another Traffic Jam?* [WWW-Dokument unter <http://www.csun.edu/sociology/virexp.htm> , letzter Aufruf 19.08.2006].
- Schwandt, Thomas A. & Halpern, Edward S. (1988). *Linking auditing and metaevaluation: Enhancing quality in applied research*. Applied social research methods series. Vol. 11. Newbury Park: Sage.
- Schwittmann, Dieter (1987). *Schwerpunkte und Aufgaben fernstudienorientierter Forschung*. In: H. Dichanz (Hrsg.): *Fernunterricht - Fernstudium. Beiträge zu Alternativen der Weiterbildung*. Hagen, S.186-212.
- Scriven, Michael (1980). *The Logic of Evaluation*. Pt. Reyes, Calif: Edgepress.
- Scriven, Michael (1991): *Evaluation Thesaurus*. Newbury Park: Sage.

- Sivin-Kachala, Jay. & Bialo, Ellen R. (1996). *The Effectiveness of Technology in Schools: A Summary of Recent Research*. School Library Media Quarterly, 25(1). S. 51-57.
- Skinner, Burrhus F. (1938). *The behavior of organismus: An experimental analysis*. Englewood Cliffs, N.J.
- Skinner, Burrhus F. (1958). *Teaching machines*. Science, 128, S. 969-977.
- Smolle, Josef; Staber, Reinhard; Neges, Heide; Reibnegger, Gilbert;Kerl, Helmut (2005). *Computer-based Training in der Dermatoonkologie – erste Ergebnisse zum Vergleich elektronischer Lernprogramme mit Präsenzlehre*. Journal der Deutschen Dermatologischen Gesellschaft Volume 3 Issue 11. S.883-888.
- Snow, Richard E. & Swanson, Judy. (1992). *Instructional psychology: Aptitude, adaption, and assessment*. Annual Review of Psychology, 43. Jg. S. 583-626.
- Spinner, Helmut F. (1998). *Die Architektur der Informationsgesellschaft: Entwurf eines wissensorienten Gesamtkonzepts*. Bodenheim: Philo.
- Spörer, Nadine (2003). *Strategie und Lernerfolg. Validierung eines Interviews zum selbstgesteuerten Lernen*. Dissertation in der Humanwissenschaftlichen Fakultät der Universität Potsdam.
- Stehr, Nico (1994). *Arbeit, Eigentum und Wissen. Zur Theorie von Wissensgesellschaften*. Frankfurt am Main: Suhrkamp.
- Strittmatter, Peter & Niegemann, Helmut (2000). *Lehren und Lernen mit Medien. Eine Einführung*. Darmstadt: Wissenschaftliche Buchgesellschaft.
- Surry, Daniel W. & Ensminger, David (2001). *What's Wrong with Media Comparison Studies?* In: Educational Technology, 41(4). S.32-35.
- Tergan, Sigmar-Olaf (2003). *Realistische Qualitätsevaluation von E-Learning*. In Meister, Dorothee, Tergan, Sigmar-Olaf & Zentl, Peter (Hrsg.): *Evaluation von E-Learning - Zielrichtungen, methodologische Aspekte, Zukunftsperspektiven*. Münster: Waxmann. S.131-154.

- Tergan, Sigmar-Olaf (2003a). *Was macht Lernen erfolgreich? Die Sicht der Wissenschaft*. In U. Beck & W. Sommer (Hrsg.). LEARNTEC 2003. Europäischer Kongress und Fachmesse für Bildungs- und Informationstechnologie. Tagungsband, Band 2. Karlsruhe: Karlsruher Messe- und Kongress-GmbH. S. 423-425.
- Tergan, Sigmar-Olaf (2004). *Was macht Lernen erfolgreich? Die Sicht der Wissenschaft*. In: Tergan, Sigmar-Olaf & Schenkel, Peter (Hrsg.). *Was macht E-Learning erfolgreich? Grundlagen und Instrumente der Qualitätsbeurteilung*. Berlin: Springer. S. 15-28.
- Tergan, Sigmar-Olaf & Schenkel, Peter (2004). *Qualität von E-Learning. Eine Einführung*. In: Tergan, Sigmar-Olaf & Schenkel, Peter (Hrsg.). *Was macht E-Learning erfolgreich? Grundlagen und Instrumente der Qualitätsbeurteilung*. Berlin: Springer. S. 3-13.
- Tinnerman, Larry S. (2006). *A Comparative Study Between Traditional and Distance Education Instructional Environments Involving Two Graduate Level Learning Disabilities Classes*. International Journal of Instructional Technology and Distance Learning, Vol. 3. No. 4.
- Traxler, Hans (1999). Karikatur: „Zum Ziele einer gerechten Auslese lautet die Prüfungsaufgabe für Sie alle gleich: Klettern Sie auf den Baum!“. In (ders.): *Alles von mir*. Frankfurt a. Main.
- Tulodziecki, Gerhard (2001). *Neue Medien - veränderte Schule?* Vortrag im Rahmen des Studenttags der PH Ludwigsburg am 22. Februar 2001. [WWW-Dokument unter www.ph-ludwigsburg.de/fileadmin/subsites/1b-mpxx-t-01/user_files/Tulodziecki.pdf , letzter Aufruf: 12.04.2007.
- Universität Dortmund, Dortmunder Zentrum Behinderung und Studium (DoBuS) (Hrsg.) (2004). *Studium in Dortmund. Ein Leitfaden für behinderte Studierende*.
- Uusikylä, Petrei & Virtanen, Petri (2000). *Meta-Evaluation as a Tool for Learning. A Case Study of the European Structural Fund Evaluations in Finland*. Evaluation, Vol. 6, No. 1. S. 50-65.
- von Prümmer, Christine (1999). *Medienevaluation als empirisch geleitete Reflexion*. In: Lehmann, Klaus (1999) *Studieren 2000 - Alte Inhalte in neuen Medien?* Münster: Waxmann. S. 221-237.

- Waxman, Hersh C.; Lin, Meng-Fen & Michko, Georgette M. (2003). *A Meta-Analysis of the Effectiveness of Teaching and Learning With Technology on Student Outcomes*. Learning Point Associates [WWW-Dokument unter <http://www.ncrel.org/tech/effects2/waxman.pdf>, letzter Aufruf: 17.04.2007].
- Weber, Peter J. (2005). „Computer literacy“ im Vergleich zwischen europäischen Nationen – Zielgröße des Bildungswettbewerbs in der Wissensgesellschaft. *Tertium Comparationis*, 11 (1/2005).
- Weinreich, Frank & Schulz-Zander, Renate (2000). *Schulen am Netz – Ergebnisse der bundesweiten Evaluation. Ergebnisse einer Befragung der Computerkoordinatoren und –koordinatorinnen an Schulen*. Zeitschrift für Erziehungswissenschaft 4/2000. S. 577-593.
- Widmer, Thomas (1996). *Meta-Evaluation: Kriterien zur Bewertung von Evaluationen*. Bern: Haupt.
- Wisan, Gail; Nazma, Shirin; Pscherer, Charles P. Jr. (2001). *Comparing online and Face-to-Face Instruction at a Large. Virtual University: Data and Issues in the Measurement of Quality*. AIR 2001 Annual Forum Paper. Paper presented at the Annual Meeting of the Association for Institutional Research. ERIC Document Reproduction Service ED 448 722.
- Wittmann, Werner (1985). *Evaluationsforschung: Aufgaben, Probleme und Anwendungen*. Berlin, Springer.
- Wottawa, Heinrich & Thierau, Heike (1990). *Lehrbuch Evaluation*. Bern.
- Yaakub, Mohammed N. (1998). *Meta-Analysis of the Effectiveness of Computer-Assisted Instruction in Technical Education and Training*. Virginia Polytechnical Institute and State University, USA. [WWW-Dokument unter http://scholar.lib.vt.edu/theses/available/etd-6398-20243/unrestricted/2_ABSTRACT1.PDF, letzter Aufruf: 01.08.2007].
- Yatrakis, Pan G.; Simon, Helen K. (2003). *The Effect of Self-selection on Student Satisfaction and Performance in Online Classes*. *International Review of Research in Open and Distance Learning*, 3(2).

Zimmer, Gerhard M. & Psaralidis, Elena (2000). *„Der Lernerfolg bestimmt die Qualität einer Lernsoftware!“ – Evaluation von Lernerfolg als logische Rekonstruktion von Handlungen*. In: In: Schenkel, Peter, Tergan, Sigmar-Olaf & Lottmann, Alfred (Hrsg.). *Qualitätsbeurteilung multimedialer Lern- und Informationssysteme. Evaluationsmethoden auf dem Prüfstand*. Reihe multimediales Lernen in der Berufsbildung. Nürnberg: BW-Vlg. , S. 22-51.

Anhang

I	Beschreibung der Primärstudien	II
I.I	P 1: Edward R. Jones et al. (1999)	II
I.II	P 2: Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr. (2001).....	V
I.III	P 3: Rob Martens, Stanley J. Portier & Martin Valcke (1995)	VIII
I.IV	P 4: Josef Smolle, Reinhard Staber, Heide Neges, Gilbert Reibnegger & Helmut Kerl (2005).....	XII
I.V	P 5: Randall S. Davies & Robert Mendenhall (1998).....	XV
I.VI	P 6: Vassilis Hartzoulakis (2002).....	XIX
I.VII	P 7: William J. Ryan (2002)	XXIII
I.VIII	P 8: Thommy Eriksson, Alois Goller & Sergei Muchin (2001)....	XXVI
I.IX	P 9: Ruth Green & Karen Gentemann (2001)	XXIX
I.X	P 10: David J. Pucel & Thomas F. Stertz (2005).....	XXXII
I.XI	P11: Larry S. Tinnerman (2006)	XXXVI
II	Bewertung der Primärstudien nach den übrigen Kriterien	XL
II.I	Erläuterung der Bewertungskriterien	XL
II.II	Operationalisierung der Kriterien	XLII
II.III	Bewertung der Primärstudien	XLIV
II.III.I	P 1: Edward R. Jones et al. (1999)	XLIV
II.III.II	P 2: Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr. (2001).....	XLVIII
II.III.III	P 3: Rob Martens, Stanley J. Portier & Martin Valcke (1995)	LI
II.III.IV	P 4: Josef Smolle, Reinhard Staber, Heide Neges, Gilbert Reibnegger & Helmut Kerl (2005).....	LV
II.III.V	P 5: Randall S. Davies & Robert Mendenhall (1998).....	LIX
II.III.VI	P 6: Vassilis Hartzoulakis (2002).....	LXIII
II.III.VII	P 7: William J. Ryan (2002)	LXVIII
II.III.VIII	P 8: Thommy Eriksson, Alois Goller & Sergei Muchin (2001)....	LXXI
II.III.IX	P 9: Ruth Green & Karen Gentemann (2001)	LXXIV
II.III.X	P 10: David J. Pucel & Thomas F. Stertz (2005).....	LXXVIII
II.III.XI	P11: Larry S. Tinnerman (2006)	LXXXII
II.IV	Vergleich der Primärstudien im Hinblick auf die übrigen Kriterien.....	LXXXV
III	Übersicht über die Operationalisierung der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)	LXXXVIII

I Beschreibung der Primärstudien

Es sollen nun die bereits in der Arbeit untersuchten Primärstudien vorgestellt werden. Dies dient dazu, Zielsetzung und Vorgehen der einzelnen Evaluationen im Gesamtzusammenhang betrachten zu können, ohne dass der Kontext durch die einzelnen Bewertungskriterien durchgekreuzt wird. Redundanzen zu den Angaben in der Meta-Evaluation sind also wahrscheinlich.

I.1 P 1: Edward R. Jones et al. (1999)

Die vorliegende Evaluation hatte zum Ziel, Vor- und Nachteile einer Online-Klasse gegenüber einer traditionellen Face-to-Face-Klasse aufzuzeigen. Dabei wurde eine Lerngruppe in traditioneller Art und Weise in Präsenzkursen unterrichtet, die andere komplett über das Internet.

Fragestellungen und Zielsetzungen

Hintergrund der Studie ist eine vorliegende Arbeit aus dem Jahr 1997ⁱ, die zeigte, dass Studierende in einem virtuellen Klassenzimmerⁱⁱ signifikant höhere Lernergebnisse erzielten als Studierende in einer traditionellen Klasse. Laut dieser Studie von Jerald G. Schutte (1997) betrage dieser Unterschied sogar 20%. Nachdem es verschiedene methodische Kritik an dieser Studie gegeben hatte, griffen Edward R. Jones et al. (1999) das Thema erneut auf und beschlossen, die Ergebnisse von Jerald G. Schutte in einer eigenen Untersuchung zu replizieren. Die untersuchte Maßnahme war das virtuelle Klassenzimmer im Vergleich mit traditionellem Unterricht.

Die Untersuchung wurde an 89 Studierenden in zwei Gruppen eines Statistikkurses an der Texas A&M University Corpus Christi durchgeführt. Davon wurde die eine Gruppe in traditioneller Art und Weise unterrichtet, die andere ausschließlich virtuell.

Das Untersuchungsdesign entsprach im Wesentlichen der Vorgängerstudie von Jerald G. Schutte (1997). In beiden Fällen setzte sich die Stichprobe aus Studierenden eines einführenden Statistikkurses zusammen. Auch die Unterrichtsmethode war mit der Vorgängerstudie vergleichbar. Es wurden drei Absichten verfolgt:

1. Die Leistungsunterschiede zwischen Studierenden in einem webbasierten und einem traditionellen Kurs sollten vor dem Hintergrund evaluiert werden, dass die Studierenden selbst die Wahl gehabt hatten, welchen dieser beiden Kurse sie besuchen wollten.

ⁱ Schutte, J.G. (1997): Virtual teaching in Higher Education: The new Intellectual Superhighway or Just Another Traffic Jam? Retrieved November 29, 1998 from WWW: <http://www.csun.edu/sociology/virexp.htm>.

ⁱⁱ Im Original wird der Begriff „Virtual classroom“ verwendet.

2. Darüber hinaus sollte die Hypothese untersucht werden, dass der Hintergrund der Studierenden, die einen Webkurs vorziehen, keine Unterschiede zu dem Hintergrund der Studierenden aufweist, die einen traditionellen Kurs bevorzugen.
3. Es sollte identifiziert werden, inwiefern Vorteile der Technologie (falls vorhanden) in einer webbasierten Klasse den Lernerfolg in einem einführenden Statistikkurs erhöhen (vgl. a.a.O.:2).

Voraussetzungen

Voraussetzung für die Einschreibung in den Kurs war das Bestehen eines Vortestsⁱⁱⁱ in Algebra. Ebenfalls ist als eine Voraussetzung der Untersuchung anzusehen, dass die Gruppenzuweisung auf freiwilliger Basis erfolgte und nicht forciert wurde. Es wird davon ausgegangen, dass diejenigen, die sich in den virtuellen Kurs eingeschrieben hatten, vertrauter im Umgang mit Computer und Internet waren als die Studierenden im traditionellen Kurs.

Über das Kursmaterial wird berichtet, dass anspruchsvolle Lektionen als PowerPoint Vortrag mit synchronisiertem Audio angeboten worden seien.

Während die Studierenden der traditionellen Klasse den Kurs an der Universität besuchten, hatte die virtuelle Klasse mehr Flexibilität in der Zeiteinteilung, da die Studierenden von zu Hause (oder anderen Orten) aus arbeiten konnten.

Methodik

Die Ergebnisse des erwähnten Vortests dienten nicht nur der Zulassung zum Kurs, sondern auch als Pretest, der anschließend zum Vergleich herangezogen wurde. Zur Feststellung des Lernerfolgs dienten Abbruchquoten und Noten der Multiple-Choice-Tests. Diese wurden in beiden Gruppen zeitgleich in einem zuvor abgestimmten Zeitrahmen durchgeführt. Die Untersuchung lief über 15 Wochen an zwei Kohorten von Studierenden in einem virtuellen bzw. traditionellen Statistikkurs. Dabei bezeichnete der virtuelle Kurs die Experimentalgruppe, der traditionelle Kurs die Vergleichsgruppe.

Beide Kurse hatten dieselbe Laufzeit und wurden von derselben Lehrperson durchgeführt. Beide deckten den gleichen Inhalt ab, nutzten das gleiche Kursmaterial und Hausaufgaben. Noten wurden auf Grund derselben Multiple-Choice-Tests vergeben. Auch die Prüfungen wurden unter gleichen Konditionen durchgeführt (vgl. ebd.).

Datengrundlage

Die Untersuchung basierte auf folgenden Daten:

ⁱⁱⁱ Test A4/F der Mathematical Association of America.

1. „Student demographics - age, GPA^{iv}, completed hours, and pretest scores
2. Scores on the three, in-class, multiple-choice in-class examinations
3. Student withdrawal and passing rates“ (ebd.).

Diese Daten wurden bei der Anmeldung zum Kurs sowie nach Abschluss des Kurses erhoben. Weitere Datenerhebungen während des Semesters fanden in Form von Multiple-Choice-Tests statt.

Die demographischen Angaben aus Experimental- und Vergleichsgruppe wurden mit Hilfe des Wilcoxon-Rangsummentests^v miteinander verglichen. Anschließend wurden auch die Examensnoten auf der Grundlage von Kovarianzanalysen mit Messwiederholung miteinander verglichen. Diese Analyse wurde zum einen mit und zum anderen ohne die Kovarianten (GPA und Algebravortest) durchgeführt. Mit dem „exakten Test auf Unabhängigkeit“ nach Fisher^{vi} wurden die Abbruchquoten analysiert.

Ergebnisse

Die erzielten Testergebnisse fielen in der virtuellen Klasse signifikant besser aus als in der traditionellen. Es zeigte sich aber, dass sich die Werte im Verlauf des Semesters einander näherten. Den Unterschied erklären sich die Verfasser durch den höheren Durchschnittswert der virtuellen Klasse im Vortest. Werden diese Werte berücksichtigt, so beträgt der Unterschied zwischen den Gruppen nur noch 3% und ist nicht mehr signifikant. Wird zusätzlich noch der GPA berücksichtigt, zeigt sich am Ende des Kurses sogar eine Überlegenheit der traditionellen Klasse.

Darüber hinaus wurden auch die Abbruchquoten in der Untersuchung einkalkuliert. Diese waren zwar nicht signifikant, wiesen aber höhere Durchschnittswerte in der virtuellen Klasse auf.

Zusammenfassend kann festgehalten werden, dass die Ergebnisse der Vorgängerstudie (Schutte 1997) nicht reproduziert werden konnten.

^{iv} Als GPA (Grade Point Average) wird im US-amerikanischen Bildungssystem eine Art Notendurchschnitt bezeichnet. Dabei wird jeder einzelnen Note ein Punktwert zugeordnet. Aus diesen wird anschließend ein gewichtetes arithmetisches Mittel berechnet (vgl. Wikipedia-Artikel unter http://de.wikipedia.org/wiki/Grade_Point_Average , letzter Aufruf: 25.08.2006).

^v Nichtparametrischer Test, um Mediane zu vergleichen.

^{vi} Hierbei handelt es sich um einen Test zur Signifikanzprüfung.

I.II P 2: Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr. (2001)

In der vorliegenden Studie sollte die Zufriedenheit der Studierenden mit der akademischen Qualität von Online- und Face-to-Face-Kursen untersucht werden. Darüber sollte diese auf Zusammenhänge mit der Anzahl von Online-Kursen, in die die Studierenden eingeschrieben sind, überprüft werden.

Fragestellungen und Zielsetzungen

Vor dem Hintergrund eines Papiers der „Middle States Commission on Higher Education“ (2001), das die Analyse studentischer Zufriedenheit^{vii} als einen wichtigen Indikator für die Effektivität einer höheren Bildungseinrichtung beschreibt und in Bezug darauf, dass die Zahl der Online-Einschreibungen am University of Maryland University College (UMUC) seit dem Jahr 2000 ständig ansteigt, zielte die hier untersuchte Studie von Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr (2001) darauf ab, die Zufriedenheit von Studierenden mit Online- und Face-to-Face-Betreuung zu vergleichen.

Ziel war es, signifikante Unterschiede zwischen diesen beiden Gruppen identifizieren zu können. Es wurde die Hypothese aufgestellt, dass keine Unterschiede zwischen den Gruppen vorhanden seien.

Vorraussetzungen

Die Untersuchung wurde am University of Maryland University College (UMUC) durchgeführt. Die AutorInnen verweisen dabei auf das Papier „Quality On the Line: Benchmarks for Success in Internet-Based Distance Education“ (Phipps & Merisotis 1999), welches als Bezugsrahmen empfiehlt, die Effektivität und den Lehr-Lernprozess mittels einer Evaluation zu messen, welche sich verschiedener Methoden und spezifischer Standards bedient.

Als eine dieser Methoden wird die Messung der Effektivität und Qualität von Präsenz- und Online-Lernen neben anderen Verfahren auch durch den Alumni-Survey angesehen. „‘Analysis of student satisfaction survey results’ is listed as an assessment tool to help an institution perform outcome assessment and evaluate its overall effectiveness“ (Middle States Commission on Higher Education“ 2001:43-44, zit. nach Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr 2001:4).

^{vii} Im Original: “One tool for studying institutional effectiveness and quality is the university's alumni survey. In the Middle States Commission's Characteristics of Excellence, they list a number of such tools and documents to use in the analysis and documentation of an institution's excellence. ‘Analysis of student satisfaction survey results’ is listed as an assessment tool to help an institution perform outcome assessment and evaluate its overall effectiveness“ (zit. nach Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr 2001:43-44).

Methodik

Zur Messung der Zufriedenheit mit der akademischen Qualität von Online- und Face-to-Face-Kursen und der Qualität des Services wurden alle Bachelor-AbsolventInnen des Jahres 1999 befragt. Dazu diente die AbsolventInnen-Befragung (Alumni-Survey)^{viii}, die an die ProbandInnen verschickt wurde. Es handelt sich hierbei um eine Querschnittsmessung. Es werden drei Ebenen von Studierenden untersucht, die nach Anzahl der belegten Online-Kurse eingeteilt werden.

Datengrundlage

Untersuchungsgegenstand war die AbsolventInnen-Befragung (Alumni-Survey) aller Bachelor-AbsolventInnen des Jahres 1999 (n=1802), die mit Fragen zur Bewertung des Colleges im Hinblick auf bestimmte Qualitätskriterien^{ix} versehen war. Zusätzlich wurde die Zufriedenheit mit der akademischen Qualität von Online- und Face-to-Face-Kursen und der Qualität des Services erhoben. Der Fragebogen enthielt 14 Items zur der Exzellenz oder Qualität des Studiengangs (academic program) und 9 Items zur Messung der Qualität von Service und Support.

Zur Bewertung durch die Studierenden wurde eine 5-stufige Likert-Skala benutzt, deren niedrigster Wert „sehr unzufrieden“ und deren höchster Wert „sehr zufrieden“ entsprach. So stellten höhere Werte gleichsam eine höhere Zufriedenheit dar. Ob signifikante Unterschiede zwischen Online- und Face-to-Face-Kursen vorlagen, wurde mittels des T-Tests auf Signifikanz berechnet. Der Fragebogen wurde von 567 Studierenden beantwortet, was einer Rücklaufquote von 31% entsprach. Davon bezogen sich 330 Rückmeldungen sowohl auf Online- als auch auf Face-to-Face-Kurse, 232 Studierende bewerteten nur Face-to-Face-Kurse^x.

Die AutorInnen analysierten die Ergebnisse zusätzlich in Bezug auf die Anzahl der belegten Online-Kurse der Studierenden. Es werden drei Ebenen untersucht:

1. Studierende, die nur einen Onlinekurs belegt hatten,
2. Studierende, die zwei oder drei Onlinekurse belegt hatten,
3. Studierende, die vier oder mehr Onlinekurse belegt hatten.

^{viii} Dies entspricht der Empfehlung des bereits erwähnten Papiers der „Middle States Commission on Higher Education“ (2001).

^{ix} Z.B. Inhalt des Curriculums, Qualität der Ausbildung, Entwicklung von kritischem Denken, Karriererelevanz, Tests und Bewertungen.

^x 5 Studierende haben nicht auf die Qualitätsfragen geantwortet.

Ergebnisse

Es konnte festgestellt werden, dass die Zufriedenheit mit den Online-Kursen signifikant zunahm, je mehr Kurse belegt wurden.

Allgemein wurde deutlich, dass die Zufriedenheit mit Präsenzkursen grundsätzlich höher war als mit Online-Kursen. Bei 12 von 14 Indikatoren waren diese Unterschiede statistisch signifikant. Bei einem Vergleich der Mittelwerte wurde jedoch klar, dass die Unterschiede, obwohl signifikant, von nur geringer praktischer Bedeutung waren, da sie zu klein waren als dass eine sinnvolle, unterscheidbare Interpretation möglich gewesen wäre. Lediglich hinsichtlich der „Interaktion von StudentIn zu StudentIn“ beträgt die Abweichung -0.646, was bedeutet, dass der Präsenzkurs auch in Bezug auf die zu Grunde liegende Skala deutlich besser abschneidet. In Bezug auf die Kategorien „develop critical thinking“ und „rigor and scholarship“ wurden keine signifikanten Unterschiede gefunden.

Betrachtet man die Auflistung nach Anzahl der belegten Kurse, so wird deutlich, dass die Mittelwerte in Bezug auf die Zufriedenheit mit Online-Kursen stiegen, je mehr dieser Kurse belegt wurden. Allerdings nahm dabei der Grad der Signifikanz ab: Unterschieden sich bei der Belegung von nur einem Kurs noch 12 von 14 Mittelwerten der Indikatoren signifikant, so konnten nur noch 5 signifikante Unterschiede bei der Belegung von 3 oder mehr Kursen gefunden werden.

Dabei ist ein Blick auf die einzelnen Indikatoren interessant: Beispielsweise bewerteten Studierende, die 3 oder mehr Online-Kurse besucht hatten, Präsenzkurse signifikant geringer als Studierende, die nur einen oder zwei Online-Kurse besucht hatten.

I.III P 3: Rob Martens, Stanley J. Portier & Martin Valcke (1995)

In der vorliegenden Studie wurde die Effizienz von Lernen in einer interaktiven Lernumgebung im Vergleich zu Lernen mit Printmaterialien und Face-to-Face-Kursen untersucht. Evaluiert wurden Prototypen von interaktiven Lernumgebungen. Dabei dienten die Studierenden, die den Stoff durch eine Präsenzvorlesung vermittelt bekommen haben, als Kontrollgruppe. Das Thema aller Kurse war „Inferenzstatistik“ im ersten Studienjahr im sozialwissenschaftlichen Studiengang „Psychology and Educational Sciences“ an der OpenUniversity der Niederlande.

Eine Schlüsselrolle in der Untersuchung spielten die sogenannten „embedded support devices (ESD)“, eine Art didaktische Zugabe zum Basiskursmaterial mit Lösungshinweisen, Glossaren, Selbsttests, Beispielen usw. Die ESDs waren sowohl in den online vermittelten als auch in den Printkursen vorhanden, allerdings entschied in beiden Fällen der Autor bzw. die Autorin der jeweiligen Kurse über ihre Verwendung.

Fragestellungen und Zielsetzungen

Die Hauptfragestellung der Untersuchung zielte darauf ab, ob unabhängiges Lernen eine gute Alternative zu traditionellem Face-to-Face Lernen darstellte. Dies sollte vor allem vor dem Hintergrund einer kostengünstigen Ausbildung untersucht werden^{xi}.

In der Studie wirkten 502 Studierende mit, die einen sozialwissenschaftlichen Studiengang im ersten Semester belegten. Davon waren 342 ProbandInnen in der Kontrollgruppe, der Rest ist in 5 Untersuchungsgruppen aufgeteilt.

Von der Lernumgebung existierten verschiedene Versionen: in der ersten Version waren alle Möglichkeiten der Interaktivität geblockt, so dass Studierende die Umgebung als eine Art elektronisches Lernbuch ohne weitere Funktionen verwenden konnten. In zwei weiteren Versionen wurden diese angeboten, die Umgebungen unterschieden sich hier nur in der Art und Weise der angebotenen ESD. Diese waren entweder mehr oder weniger sichtbar.

Das Printmaterial entsprach dem Standard von Lehrbriefen. Es handelte sich hierbei um didaktische aufbereitete Kurse, die von den Studierenden eigenständig bearbeitet wurden. Zu Forschungszwecken wurden zwei verschiedene Versionen von Printmaterial eingesetzt – einmal mit und einmal ohne Hilfestellungen bzw. didaktischen Anweisungen. Als Kontrollgruppe diente der Untersuchung eine traditionelle Face-to-Face-Vorlesung der Inferenzstatistik.

^{xi} Dazu muss gesagt werden, dass die OpenUniversity ebenfalls eine FernUniversität ist, deren Lehrvermittlung traditionell über Printkurse erfolgt. Dies ist zumindest zur Zeit der Veröffentlichung der Fall, inzwischen hat sich die OU vor allem in Bezug auf die Entwicklung und Umsetzung interaktiver Lernangebote einen Namen gemacht.

Die ersten drei Forschungshypothesen zielten auf die Effektivität der Lernumgebung ab, die letzten drei auf die Interaktion zwischen Studierenden, den Gebrauch der Lernumgebung und dem Lernerfolg.

1. Studierende, die mit der interaktiven Lernumgebung arbeiten werden zumindest gleiche Ergebnisse (study outcomes) erzielen, wie diejenigen, die „face-to-face“ studieren.
2. Studierende, die mit den ESDs arbeiten, profitieren von Lernumgebungen, in welchen die ESDs weniger sichtbar sind. Andererseits profitieren auch Studierende, die keine ESDs benutzen, von Lernumgebungen, in denen ESDs sichtbar sind^{xii}.
3. Studierende, die in der interaktiven Lernumgebung studieren, erzielen höhere Lernerfolge als Studierende in einer nicht-interaktiven Lernumgebung.
4. Vorwissen macht sich im positiven Sinn bei der Benutzung der Funktionalitäten der interaktiven Lernumgebung bemerkbar und stellt eine signifikante Interaktionsvariable im Zusammenhang von Maßnahme und Lernerfolg dar.
5. Gute LeserInnen benutzen weniger ESDs, um die eigenen Aktivitäten zu unterstützen.
6. Es gibt eine positive Korrelation zwischen Motivation und ESD-Benutzung, ihre Interaktion beeinflusst den Lernerfolg.

Voraussetzungen

Da sich die Studierenden im ersten Studiensemester befanden, wurde das Vorwissen als sehr gering eingeschätzt, was in Bezug auf die Untersuchung hinsichtlich der Vergleichbarkeit vorteilhaft war. Es wurde dennoch ein Pretest absolviert, mit welchem das Vorwissen der Studierenden geprüft wurde, so dass dieses als eine Einflussvariable behandelt werden konnte.

Die Navigation innerhalb der Lernumgebung ist nachvollziehbar, ebenfalls die Benutzung der ESDs. Da eine Logfile-Analyse durchgeführt wurde, ist davon auszugehen, dass die Studierenden nicht zu Hause, sondern immer in von der Hochschule betreuten Computerräumen gearbeitet haben.

Methodik

Um verschiedene Prozesse und Variablen kontrollieren zu können, wurden im Vorfeld ein Leseverständnistest sowie ein Vorwissenstest durchgeführt, der als Pretest anzusehen war.

^{xii} “This interaction is expected because low-users may want to skip certain ESD. When ESD are nondiscernable this might hinder these learners because they have to read the ESD before they know what ESD it is. On the other hand high-users who always tend to read ESD might find the different lay-out of ESD obstructive”(Rob Martens, Stanley J. Portier & Martin Valcke 1995:8).

Darüber hinaus wurde auch der Einfluss der Lernmaterialien und der Lernumgebung selbst auf das Lernen untersucht. Als individuelle Variablen wurden Vorwissen und noch einmal Lesekompetenz in die Studie einbezogen. Darüber hinaus sollte der Gebrauch und der Einfluss der ESDs in Bezug auf das Vorwissen der Lernenden mittels Logfile-Analyse überprüft werden: Aus der Information, was genau benutzt worden war, sollte geschlossen werden, welche Hilfen jemand benötigte und welches Vorwissen vorhanden war.

Um Vergleiche von allen Versionarten und der Kontrollgruppe anstellen zu können, wurden also verschiedene Settings bzw. Experimentalgruppen miteinander verglichen:

1. Lernumgebung mit sichtbarem Hilfesystem (n=40)
2. Lernumgebung ohne sichtbares Hilfesystem^{xiii} (n=20)
3. Lernumgebung mit sichtbarem Hilfesystem (nicht-interaktiv)^{xiv} (n=20)
4. Printkurse mit Hilfesystem (n=40)
5. Printkurse ohne Hilfesystem (n=40)
6. Vorlesung (Kontrollgruppe) (n=342)

Die Zuweisung der ProbandInnen zu den einzelnen Lerngruppen erfolgte mittels randomisierter Verfahren. Testergebnisse von Studierenden mit höherem Ausbildungsniveau sowie Studierende, die den Kurs zum zweiten Mal besuchten, wurden von der Untersuchung ausgeschlossen.

Datengrundlage

Kognitive Fähigkeiten, Lesekompetenzen, Motivation und Vorwissen zählten in der vorliegenden Studie als Einflussfaktoren in Bezug auf das Lernen. Zur Feststellung des Lernerfolgs wurde neben den Abschlussnoten ein so genannter „Mastery Test“ durchgeführt, der aus einem dreiteiligen Fragebogen bestand:

- A) „includes questions about motivation, age, study habits, educational level and gender;
- B) asks students to judge the degree of accessibility of the learning materials studied;
- C) puts forward questions about the use and appreciation of each specific embedded support device“ (Rob Martens, Stanley J. Portier & Martin Valcke 1995:16).

Dieser Fragebogen beinhaltete vor allem Items zur den Rahmenbedingungen. Des Weiteren sollten die Studierenden den Grad des Zugangs zu den Studienmaterialien bewerten (für Studierende in der Kontrollgruppe). Als dritter Teil wurden Fragen angegeben, die sich auf

^{xiii} Die Arbeitshilfen sind in diesem Fall in den Content integriert und nicht als solche ausgewiesen.

^{xiv} Die Arbeitshilfen sind explizit als solche ausgewiesen.

die Nutzung jedes einzelnen ESDs bezogen sowie eine Skala zur Messung der Einstellung in Bezug auf das Lernen mit Computern (für Studierende in der Experimentalgruppe).

KursteilnehmerInnen in den Experimentalgruppen arbeiteten drei Lernabschnitte von jeweils zwei Stunden mit der spezifischen Lernumgebung. Nach jedem Lernabschnitt füllten sie den Fragebogen (Teil C) aus. Teilnehmende der Kontrollgruppen folgten dem regelmäßigen Vortragsformat. Drei Vorträge wurden gehalten, jeweils bezogen auf eins der gewählten Kapitel im Bereich der Inferenzstatistik.

In den weiteren Analysen sind die folgenden Techniken angewendet worden: Regressionsanalyse, ANOVA, MANOVA und Profilanalyse, was von den AutorInnen als geeignet beschrieben wurde, um Effekte zwischen unabhängigen Mehrebenenvariablen, Effekte der Maßnahme und abhängigen Variablen zu isolieren.

Ergebnisse

In den meisten getesteten Bereichen brachte die Studie keine signifikanten Ergebnisse zu Tage: keines der Forschungssettings führte zu signifikant besserem Lernerfolg der Studierenden. Ebenfalls keinen signifikanten Unterschied zeigte der Vergleich zwischen interaktiven und nicht-interaktiven Lernmaterialien, wobei die Experimentalgruppe, die mit der voll-interaktiven Lernumgebung gearbeitet hat, die besten Testergebnisse erzielte. In Bezug auf die Verfügbarkeit der ESDs kann festgehalten werden, dass auch hier keine signifikanten Ergebnisse erzielt wurden: keine der Lerngruppen erreichte signifikant höhere Testwerte. Allerdings zeigte sich hier ein interessanter Aspekt: Studierende, die oft mit ESDs arbeiteten, erreichten signifikant höhere Werte, wenn diese gerade nicht verfügbar waren. Umgekehrt hatten auch Studierende, die die ESDs wenig benutzten, ein höheres Testergebnis, wenn ESDs vorhanden waren.

In Bezug auf den Lernerfolg wurde ein signifikanter Zusammenhang zwischen dem Vorwissen der Studierenden und der Benutzung der ESDs festgestellt, ebenso ein (nicht verwunderlicher bzw. überraschender) Zusammenhang zwischen der Einstellung zum Lernen mit Computern und des Darreichungsmodus der Inhalte.

Werden die erwähnten Hypothesen betrachtet, so kann in Bezug auf die hier relevanten Hypothesen festgehalten werden, dass Hypothese 1 akzeptiert werden konnte bzw. die vermuteten Zusammenhänge nicht widerlegt wurden. Hypothesen 3 wurde dagegen abgelehnt. Auch Hypothese 6 musste den Ergebnissen der Studie nach verworfen werden. Die Messungen zeigten keinen Einfluss der Motivation auf verschiedene ESD-Typen. Allerdings ist ein leichter Einfluss von Einstellungen gegenüber Computern im Allgemeinen und Nutzung von ESDs erkennbar. Eine Post-Hoc-Analyse zeigt einen positiven signifikanten Zusammenhang zwischen Einstellung und Motivation. Des Weiteren wurden keine Effekte in Bezug auf den Einfluss der Motivation auf die Lernergebnisse gefunden. Schaut man sich für die interaktive Lernumgebung jedoch erneut die Variable „Einstellung gegenüber Computern“ an, so kann auch hier ein signifikanter Zusammenhang festgestellt werden.

I.IV P 4: Josef Smolle, Reinhard Staber, Heide Neges, Gilbert Reibnegger & Helmut Kerl (2005)

Bei der vorliegenden Publikation handelt es sich um einen Artikel, der in einer medizinischen Zeitschrift veröffentlicht wurde. Die AutorInnen verfolgen das Ziel, Online- mit Präsenzlernen zu vergleichen. Dazu diente ein Seminar im Studienmodul Dermatoonkologie, das als eines von 6 möglichen Modulen von den Studierenden belegt werden konnte, die damit die Untersuchungspersonen abbildeten. Das Seminar war auf 30 Stunden über einen Zeitraum von 5 Wochen ausgerichtet. Davon wurden 24 Stunden in Präsenzform und 6 Stunden virtuell abgehalten.

Fragestellungen und Zielsetzungen

Die Forschungsfrage lautete, ob der Erfolg beim Lernen mit Computerlehrprogrammen in Form von Computer Based Training geringer oder höher sei als der Lernerfolg, der mit „innovativen didaktischen Konzepten im Präsenzunterricht“ (Josef Smolle, Reinhard Staber, Heide Neges, Gilbert Reibnegger & Helmut Kerl 2005:883) erzielt würde.

Voraussetzungen

Im Präsenzkurs wurde Team-Techning als didaktisches Modell angewendet. Dabei wurden klinische Themen mit PowerPoint präsentiert, danach in Kleingruppen diskutiert und anschließend den anderen Studierenden vorgestellt. Weitere Informationsvermittlung fand durch Mini-Lectures statt. Bei den Lernunterlagen im Online-Kurs handelte es sich um Lernprogramme zu klinischen Fragestellungen, die Bildmaterial und eine differenzierte Rückmeldefunktion erhielten. Dieser Kursteil wurde in Form von Computer-Based-Training abgehalten. Für die Absolvierung der Aufgaben wurde im Vorfeld ein Termin festgesetzt. Zum Bestehen dieses Kursteils war eine Trefferquote von 90% erforderlich, was u.U. ein mehrfaches Durchlaufen einzelner Aufgaben erforderte.

Die Einbindung der Lernmaterialien im Online-Kurs erfolgte über den „Virtuellen Medizinischen Campus“ der Universität Graz. Im Semester der Untersuchung waren etwa 300 Studierende in den Studiengang eingeschrieben. Die Zulassung zu dem untersuchten Modul war jedoch auf 26 Studierende beschränkt.

Methodik

Zum Vergleich von Präsenz- und Online-Lernen wurde ein Modul im Studienmodul (Wahlpflichtfach) Dermatoonkologie untersucht, das beide Lernsettings beinhaltet. Die Untersuchung wurde im Wintersemester 2003/2004 mit 16 Studentinnen und 10 Studenten

durchgeführt, die den 2. Studienabschnitt des Diplomstudiums Humanmedizin im 3. Semester belegt hatten.

Das Modul schloss mit einen Multiple-Choice-Test ab, der zu gleichen Teilen Fragen zum Inhalt von Präsenz- bzw. Online-Unterricht enthielt. Da die Studierenden jeweils den Präsenz- und den Online-Teil des Kurses absolvieren mussten, bildeten sich dementsprechend die zwei Untersuchungsgruppen.

Datengrundlage

Um zu überprüfen, welches Lernsetting den höheren Lernerfolg erzielt, wurden die Prüfungsergebnisse untersucht, „die mit einem elektronischen Format im Vergleich zu einem Format der Präsenzlehre erzielbar sind“ (a.a.O.:884). Diese Leistungsüberprüfung fand in Form des Multiple-Choice-Tests statt, der am Ende des gesamten Seminars durchgeführt wurde. Bei der Bearbeitung der Fragen beim CBT wurde überdies jeder Mausklick registriert und so die Trefferquote bzw. der Anteil an falschen Klicks aufgezeichnet.

„Zur Ermittlung der Trefferquote wurde jeder Auswahlvorgang, d. h. jeder Mausklick, registriert, und der Anteil an „falschen“ Klicks im Verhältnis zur Gesamtzahl der Klicks aufgezeichnet. Wenn z. B. jemand jede Frage auf Anhieb richtig beantworten konnte, ergab sich eine Trefferquote von 100%. Wenn dagegen jemand z. B. bei jedem Frame einen Fehlversuch aufwies, bevor er die richtige Antwort fand, dann ergab sich eine Trefferquote von 50 %“ (a.a.O.:885).

Als Datenmaterial für die statistische Analyse dienten die durch die Multiple-Choice-Prüfung erzielten Ergebnisse sowie die Leistungszahl, ein allgemeines Maß für den bisherigen Studienerfolg. Darüber hinaus wurden die Kommentare der Studierenden untersucht, zu denen nach jeder Absolvierung eines Lernobjekts Gelegenheit gegeben wurde.

Am Ende des Moduls fand eine Befragung mittels formalisiertem Fragebogen statt.

Ergebnisse

In der vorliegenden Evaluation wurde festgestellt, dass das CBT bessere Ergebnisse hervorgebracht hat als der Präsenzunterricht – die Prüfungsergebnisse aus diesem Teil der Abschlussprüfung fielen signifikant höher aus. Dies wurde im nicht-parametrischen paarweisen Vergleich deutlich herausgestellt.

„Die am Computer abgehaltene Multiple-Choice-Prüfung im Online-Format erbrachte für die Fragen aus dem CBT über alle Studierenden $90,4 \pm 12,5$ % richtige Antworten. (42-100%), für die Fragen aus dem Team-Learning $80,6 \pm 13,1$ % richtige Antworten (55-100%). Die mittlere Differenz zwischen den richtigen Prozentsätzen für CBT und Team-Learning betrug $9,8 \pm 12,5$ % (-13-39%). Insgesamt waren bei 17 Studierenden die Leistungen in den CBT-Fragen besser als bei den Team-Learning-

Fragen, bei 6 Studierenden war es umgekehrt, und bei 3 Studierenden bestand kein Unterschied“ (a.a.O.:885).

Ebenfalls wurde die Trefferquote der eingesandten Rückmeldungen mit dem CBT mit den Prüfungsleistungen der einzelnen Studierenden auf Signifikanz getestet. Hier ist ein hoch signifikanter Zusammenhang festzustellen ($r=0,791$, $p<0,001$). Auch der Vergleich der Rückmeldungen aus dem Team-Lernen mit den Prüfungsleistungen weist einen signifikanten Zusammenhang auf ($r=0,40$; $p<0,05$).

I.V P 5: Randall S. Davies & Robert Mendenhall (1998)

Fragestellungen und Zielsetzungen

Bei der vorliegenden Primärstudie aus dem Jahr 1998 handelt es sich um einen Forschungsbericht zu einer Evaluation, die WWW-basiertes Online-Lernen mit Instruktion im Klassenraum verglich. Beide Darbietungsmodi bezogen sich auf den Kurs „HEPE 129 - Fitness and Lifestyle Management“, der im Studiengang „Health Education/ Physical Education“ an der Brigham Young University im US-Bundesstaat Utah durchgeführt wurde. Ziel der Untersuchung war, Daten über die Qualität des Unterrichts in die Diskussion um den Wert webbasierter Kurse einzubringen.

Die Evaluation zielte ab auf:

- Die Effektivität von Instruktion im Online- und im Klassenraummodus in Bezug auf die Vermittlung der Grundlagen des Kurses HEPE 129,
- den Grad, inwieweit die Instruktion die Haltung der Studierenden in Bezug auf Gesundheit und Fitness beeinflusst,
- Vor- und Nachteile der beiden Instruktionsmethoden,
- Signifikante Unterschiede der Testergebnisse und Auswertungen des studentischen Feedbacks (vgl. Randall S. Davies & Robert Mendenhall (1998:8, Übersetzung AP).

Darüber hinaus wurden verschiedene Forschungsfragen aufgestellt:

1. Erreichen Studierende, die mit unterschiedlichen Instruktionsmodi lernen, gleiche Ergebnisse wie die Studierenden, die auf dem Campus lernen?
2. Sind die Studierenden der Ansicht, dass sich ihre Ansichten und ihr Verhalten nach Absolvieren des Kurses geändert hat?
3. Benutzen die Studierenden die verfügbaren instruktionalen Ressourcen?
4. Präferieren die Studierenden eine Darreichungsform gegenüber der anderen?
5. Finden die Studierenden eine Darbietungsform nützlicher als die andere?

Voraussetzungen

Zum Zeitpunkt der Untersuchung (1998) waren etwa 1000 Studierende in den Kurs HEPE 129 eingeschrieben. Abhängig von der Kapazität wurde dieser Kurs in 5 bis 6 Sektionen unterteilt, deren Klassengröße etwa jeweils 200 Studierende betrug.

Während der ursprüngliche Kurs im Klassenraum auf dem Campus unterrichtet wurde, war der Online-Kurs, der vom „Independent Studies Departement“ entwickelt wurde, über das WWW verfügbar. Beide Kurse bezogen sich jedoch auf die gleichen Inhalte.

„Both courses cover the same basic objectives in thirteen independent modules or lessons; both require students to complete similar additional learning activities; yet the methods for providing the instruction are quite different“ (Randall S. Davies & Robert Mendenhall (1998:6).

Bei dem Kurs, der im Klassenraum unterrichtet wurde, handelte es sich um einen traditionellen Kurs, der in 18 Vorlesungsmodulen aufgeteilt war. Die Studierenden wurden in 13 dieser Module getestet und mussten darüber hinaus 3 von 5 Wahlpflichtmodulen (Marriage and Family Relations, Women's Health, Environmental Health, Maintaining a Healthy Weight, and Crisis Intervention) belegen, die jedoch nicht prüfungsrelevant waren.

Das Material des Online-Kurses wurde von verschiedenen Lehrenden des HEPE 129-Kurses erstellt und diente vor allem der Reflexion. Es gab 13 Lerneinheiten, die zur Absolvierung des Kurses belegt werden mussten. Das gesamte Kursmaterial war online verfügbar und bestand aus Videoclips, Animationen, Sound, Grafiken und weiterführenden Links. Weiterhin gab es ein „did you know“-Feature, das verwandte Fakten präsentierte. Darüber hinaus bekamen die Studierenden vier CD-ROMs, die Videoclips und Animationen enthielten. Durch ein Bulletin-Board bestand die Möglichkeit zum Kontakt mit anderen Lernenden - dies war jedoch freiwillig. Im Online-Kurs wurde nach jeder Lektion ein Test durchgeführt, auf den direkte Rückmeldungen gegeben wurden. Ebenfalls war Feedback in Bezug auf den Lernprozess und die Fitnessprogramme vorgesehen. Im Online-Kurs gab es keine Wahlpflichtmodule.

Um die Zielgruppe für die Untersuchung zu rekrutieren, wurden die Studierenden, die die Stunden online absolvieren sollten, um eine freiwillige Beteiligung gebeten. Es stellten sich letztlich 39 Studierende zur Verfügung, von denen 15 die Inhalte zu Hause bearbeiteten, 24 im Computerraum der Universität. Diese Freiwilligen wurden daraufhin zufällig auf drei Gruppen aufgeteilt: Zwei Gruppen studierten die Inhalte nur unter Zuhilfenahme des Online-Materials, eine Gruppe wurde im Klassenraum unterrichtet.

Während des Kurses waren drei Prüfungen fällig. Zwischen diesen „exams“ wurden die Gruppen rotiert, damit jede Gruppe mit jeder Lernform Erfahrung machen konnte.

Methodik

Im Präsenzkurs wurden die Prüfungen zu 6 Zeiträumen während des laufenden Kurses durchgeführt und bezogen sich jeweils auf 2 bzw. 3 Module. Jede Prüfung konnte bis zu 2x gegen eine Gebühr wiederholt werden. Die Endnote des Kurses wurde zu 75% durch diese Modulprüfungen abgedeckt. Weitere 15% machten die Fitness-Übungen (Fitness-Vertrag, Fitness-Programm und Fitness-Check) aus. Für den Kurs im Klassenraum gab es 2 Lehrbücher und Leseaufgaben, die sich auf die einzelnen Module bezogen. Ferner wurde nach jeder Vorlesung auf ausgewähltes weiterführendes Material verwiesen, das in der Bibliothek beschafft werden konnte. Der Unterricht wurde im Klassenraum zusätzlich durch Beobachtungen untersucht.

Die Ergebnisse der Lektionentests des Online-Kurses flossen zu 65% in die Endnote ein. Ferner machten die Fitness-Übungen (Fitness-Vertrag, Fitness-Programm und Fitness-Check) 15% der Endnote aus. Weitere 20% der Endnote wurden durch einen protokollierten Paper-Pencil-Test abgedeckt.

In allen drei Untersuchungsgruppen wurden nach jeder Prüfung Interviews durchgeführt, die die Haltungen und Meinungen der Studierenden abfragten.

Der Fragebogen umfasste Fragen zur Vorbereitung auf die Prüfungen, die als Indikator dafür dienen sollten, in welchem Umfang sich die Studierenden sich mit dem zur Verfügung gestellten Material beschäftigt hatten. Jede Untersuchungsgruppe wurde wiederum in 2 Untergruppen aufgeteilt. Darüber hinaus wurden die Studierenden aufgefordert, an einem Fitness-Check teilzunehmen, bei dem ein Bluttest und ein Prä-/Post-Fitness-Test zu Beginn und zum Ende des Semesters durchgeführt wurde. Während des gesamten Semesters fand ein verpflichtendes Fitness-Programm statt:

„Students participating in the classroom experience must sign up for an appointment to have the fitness appraisal completed. Fitness appraisals are completed outside of regular class time and are administered by trained staff. This appraisal requires that students give a blood sample for analysis. Students record the results of the fitness appraisal on a fitness questionnaire. The questionnaire is submitted and processed. Students receive the results approximately one week later. With the online experience, students complete the fitness appraisal on their own following instructions provided. They are not required to give a blood sample. The fitness results are recorded on a fitness questionnaire and processed immediately by the computer“ (a.a.O.:7).

Datengrundlage

Die zu analysierenden Daten wurden aus den Testergebnissen in den Untersuchungsgruppen gewonnen. Für jede Gruppe wurden statistische Berechnungen durchgeführt:

„Data was gathered from test results for both in-class and online students. Mean scores and standard deviations were calculated for

each group. A measure of statistical significance was used to calculate the effect size. In addition, survey results were tabulated; classroom observations and interview information was organized by topic in order to identify issues and trends" (ebd.).

Wie bereits erwähnt, wurden die Präsenzkurse darüber hinaus durch Beobachtungen untersucht. Ein Fragebogen wurde in allen Untersuchungsgruppen zusätzlich am Ende der Evaluation eingesetzt.

Ergebnisse

Die Ergebnisse der Evaluation werden ausführlich dokumentiert. So ist festzuhalten, dass ein Vergleich der Untersuchungsgruppen keine signifikanten Unterschiede hervorbringen konnte. Die Autoren schließen daraus, dass beide Lehr- /Lernmethoden gleichermaßen geeignet seien. Die Befragung der Studierenden zeigte jedoch, dass das Lernen im Klassenraum die bevorzugte Methode war.

Unter Bezugnahme auf die Forschungsfragen stellen sich die Ergebnisse wie folgt dar:

- Studierende, die mit unterschiedlichen Instruktionsmodi lernen, erreichen die gleichen Ergebnisse wie die Studierenden, die auf dem Campus lernen.

Zwei Drittel der Studierenden geben an, etwas gelernt zu haben, das sie in besonderem Maße motiviert, den eigenen Lebenswandel dauerhaft zu ändern.

„Most were optimistic that the changes they were making in their lifestyle would be permanent. This optimism was found in both groups but due to the limitations of the study's design, this optimistic attitude may speak more to the course in general and not to differences in instructional delivery methods (a.a.O.:16)“.

- Nicht alle Studierenden nutzen die Ressourcen. Die Beteiligung im Klassenraum war vergleichsweise gering und die Online-Studierenden beschränkten ihre Beschäftigung auf das Lesen des Materials.
- Die Studierenden präferieren das Lernen im Klassenraum.
- Ob die Studierenden eine Darbietungsform nützlicher finden als die andere, konnte nicht geklärt werden, da keine nennenswerten Unterschiede festgestellt wurden. Die Studierenden konnten in ihren Aussagen zu jeder Methode Nachteile angeben.

I.VI P 6: Vassilis Hartzoulakis (2002)

Die Hauptfragestellung der Untersuchung zielte darauf ab, ob sich die Effektivität vom Computer gesteuerten Unterricht (CALL) mit den Ergebnissen der traditionellen Form des Unterrichts vergleichen ließe. Die Untersuchung fand in der städtischen Schule (Lykeio) im Athener Stadtteil Peristeri statt.

Fragestellungen und Zielsetzungen

Die Studie untersuchte die Effektivität von computerunterstütztem Sprachenlernen im Vergleich zu traditionellem Lernen im Klassenraum. Die Forschungsfragen lauteten: Ist computervermittelte Instruktion mehr, weniger oder gleich effektiv, wie traditionelles Lehren? Ist computervermittelte Instruktion für Mädchen und Jungen gleichermaßen effektiv? Um diese Fragen zu klären, wurden die Lernergebnisse der verschiedenen Untersuchungsgruppen miteinander verglichen.

„The study compares the learning outcomes achieved through computer-assisted instruction with those obtained in a traditional classroom when teaching one of the forms - in this case, inversion to two groups of randomly selected students in the second and third grade of Lykeio“ (Vassilis Hartzoulakis 2002:7).

An der Untersuchung wirkten 24 SchülerInnen im Alter von 15-17 Jahren des zweiten und dritten Jahrgangs des Lykeion in Peristeri mit. Zu diesem Zeitpunkt hatten diese bereits mindestens 6 Jahre Englischunterricht absolviert.

Die Software, die zur Präsentation der Unterrichtsstunden eingesetzt wurde, wurde vom Autor selbst entwickelt. Es handelte sich um eine Serie von Webseiten in html, die im Internet zur Verfügung standen. Das Thema der Lernreihe war „Inversion“.

„The webpages presented the form ‘inversion’ in an interactive way, where the students would be able to explore -employing ‘point and click’ techniques- the various cases where inversion can be used“ (a.a.O.:28).

Auf den Webseiten wurden 6 Fälle von Inversion beispielhaft beschrieben und mit Macromedia Flash umgesetzt. Es folgen Beispiele und Übungen mit „Hot potatoe“-tests.

Im Präsenzunterricht wurden zwei verschiedene Stunden von derselben Lehrperson (dem Autor) zum Thema „inversion“ abgehalten. Die Beispielsätze und Übungen, die im Unterricht verwendet wurden, entsprachen denen, die auch am Computer präsentiert wurden. Das Material wurde an der Tafel im Stil einer Vorlesung präsentiert. Fragen von SchülerInnen wurden erst nach einer Zusammenfassung der gesamten Stunde beantwortet.

Voraussetzungen

Der Lehrplan des griechischen Ministeriums für Erziehung (Hellenic Ministry of Education) spricht sich für den Einsatz von computergestützten

Lehrmethoden im Englischunterricht aus, jedoch ohne anzugeben, auf welchem Weg dies geschehen soll.

„What is specified is a detailed list of forms, functions, and structures that are to be taught throughout the three year span of the Hellenic Secondary Education. The teachers are, in that way, guided towards covering the material set -the specified **forms** and functions - in the best way they see fit for their students“ (a.a.O.:5, Hervorhebung AP).

Da dem Begriff „form“ verschiedene Definitionen zu Grunde liegen, wird in der vorliegenden Arbeit zwischen „form“ und „formS“ unterschieden (ebd.).

„Doughty and Williams (1998:4) clarified the issue by differentiating the term form from the term formS. They adopted Long's (1988) suggestions who defined isolated linguistic features (e.g. grammar) as formS and the overall engagement in meaning as form. Doughty and Williams (1998) stressed that the two terms are not polar opposites; instead, they are interwoven in the sense that form entails isolation or extraction of linguistic features from context, whereas formS is limited to isolated features. This approach has been adopted by the author throughout this study“ (ebd.).

Zum Hintergrund ist zu sagen, sich der Autor vor allem auf das Lehren von „formS“ bezieht, was bedeutet, dass isolierte linguistische Merkmale und formale Sprachelemente aus dem Kontext extrahiert werden. Detailliertere Angaben zu „form“ und „formS“ sind bei der Beschreibung der Kontextanalyse zu finden.

Letztlich wird noch auf die besondere Situation am Lykeion Bezug genommen: Bevor die SchülerInnen das Lykeio besuchen, erhalten sie zuvor in der Primar- und Sekundarerbziehung insgesamt sechs Jahre Englischunterricht. Eine große Anzahl derer erhält weiteren Englischunterricht an privaten Institutionen. Dies führt zu einem Ungleichgewicht im Level der englischen Sprachkenntnisse der SchülerInnen am Lykeion. Daher musste vor der Ziehung der Stichprobe der Kenntnisstand im Einzelnen festgestellt werden.

Methodik

Die Stichprobe der Untersuchung stellten 103 Lernende in der zweiten und dritten Jahrgangsstufe der Schule dar. Vor der Ziehung der Stichprobe wurde ein Oxford Placement Test zur Feststellung des Wissensstandes im Fach Englisch durchgeführt, der von 83 Lernenden bearbeitet wurde. Ausgehend von den Ergebnissen dieses Tests wurden 47 Schülerinnen und Schüler einer „B“- und 36 Schülerinnen und Schüler einer „C“-Klasse zugeordnet:

Da jedoch die Punkteverteilung in beiden Klassen annähernd gleich war, wurde entschieden die Daten beider Klassen zu einer Gruppe zusammenzufassen.

„This distribution is almost equal for both classes and the average grade for C class is higher than the average grade for B class. Specifically, a range of 45 points and an average of 63.12% was

noted for B (second) class, while a range of 42 points and an average of 74.44% was observed for C (third) class" (a.a.O.:23).

Um die Homogenität der Gruppe sicherzustellen, wurde ferner der Gesamtdurchschnitt der Testergebnisse berechnet. Werte unter 50% und über 80% wurden als Extremwerte ausgeklammert, so dass letztendlich 54 repräsentative Lernende in die Untersuchungsgruppe einbezogen wurden. Diesen wurde ein erneuter Prätest zur Messung des Vorwissens vorgelegt, auf Grund dessen Ergebnis die 40 leistungsschwächsten Schülerinnen und Schüler ausgewählt wurden.

Da für das Experiment genau 24 Lernende benötigt wurden, wurde diese Anzahl zufällig aus der Gruppe der verbliebenen 40 Schülerinnen und Schüler ausgewählt. Dabei wurden 12 Lernende der Experimental- und 12 Lernende der Kontrollgruppe zugeteilt. Die erste Gruppe (the lab group) bekam ihren Platz im Computerraum, in dem jede/jeder einen eigenen PC zugeteilt bekam. Die zweite Gruppe (the traditional group) wurde in einem regulären Klassenraum unterrichtet.

Wie bereits erwähnt, wurden zwei Untersuchungsgruppen gebildet – die Experimentalgruppe erhielt das Unterrichtsmaterial online, die Kontrollgruppe während des Präsenzunterrichts. Zuvor wurde in einem Prätest der Leistungsstand der Lernenden erhoben.

Zwei Tage nach dem Unterricht wurde ein Posttest durchgeführt. Dies wird damit begründet, dass die Evaluation der Effektivität nicht durch Retentionsfaktoren beeinflusst werden sollte. Der Posttest entsprach im Wesentlichen dem Prätest, mit dem Unterschied, dass alle „unnötigen“ Items – also Items, die sich nicht direkt auf das Thema „inversion“ bezogen – entfernt wurden.

Datengrundlage

Um das Vorwissen der Studierenden abzufragen, wurde ein Prätest durchgeführt. Dieser wurde zunächst von zwei Experten in Bezug auf die Validität des Inhalts überprüft, woraufhin Anpassungen vorgenommen wurden. Vor dem Einsatz des Tests wurde dieser in einer anderen Kohorte desselben Lykeions getestet und auf Schwächen überprüft.

„Before giving the pretest to the students in the final shortlist, we piloted it with a different group of students in another Lykeio. The piloting did not reveal any weaknesses in the wording of the rubrics or the structure of the exercises and the test seemed clear to the students. The time they required to complete the test was measured and it coincided with our estimate (about 35 minutes on average)"(a.a.O.:26).

Nach Durchführung des Experiments wurden die Ergebnisse der Leistungsmessung (Punktzahlen) der Prä-, Post-, und Placement-Tests statistisch ausgewertet:

„A T-Test investigating overall differences between the pretest and the posttest findings.

An ANCOVA^{xv} investigating whether there were differences in the outcomes of the two teaching methods, and

An ANOVA^{xvi} investigating whether there were differences in the performance between males and females" (a.a.O.:32).

Im Vorfeld der Datenanalyse wurde eine Nullhypothese formuliert, um den T-Test durchführen zu können. Diese lautete:

„There were no significant differences between the pre- and the posttest scores" (a.a.O.:36).

Die Alternativhypothese lautete dementsprechend:

„The differences between the pre- and posttest scores were significant" (ebd.).

Auch für die Analyse der Kovarianz wurden ebenfalls Hypothesen aufgestellt:

„‘There is no difference between the two teaching methods’ (H0: $\mu_1 - \mu_2 = 0$). The alternative hypothesis is: ‘The two methods yield different results’ (H1: $\mu_1 \neq \mu_2$)" (a.a.o.:36).

Untersucht wurden im Wesentlichen fünf Variablen: die drei numerischen Variablen der eingesetzten Tests (Prä-, Post-, und Placement-Test) und zwei Dummy-Variablen^{xvii} für das Geschlecht und die Lehrmethode.

Ergebnisse

Die Studie kommt zu dem Ergebnis, dass schlechtere Lernergebnisse in der Computerklasse – der Experimentalgruppe – erzielt worden sind, deren Mittelwert 54.86 Punkte betrug. Die Kontrollgruppe im Klassenzimmer wies einen Mittelwert von 73.61 Punkten auf. Die Ergebnisse werden in der Arbeit sehr detailliert dargestellt.

Dabei wurde u.a. der Frage nachgegangen, ob die computervermittelte Instruktion für Jungen und Mädchen gleich effektiv ist. Dem kann auf Grund der Datenlage zugestimmt werden.

Statistisch signifikante Unterschiede konnten im Vergleich von Prä- und Posttest sowie in der Favorisierung einer Lehrmethode ausgemacht werden. Dabei wurden schlechtere Lernergebnisse in der Computerklasse – der Experimentalgruppe – erzielt, deren Mittelwert 54.86 Punkte betrug. Die Kontrollgruppe im Klassenzimmer wies einen Mittelwert von 73.61 Punkten auf.

Keine Differenzen wurden jedoch beim Vergleich von Mädchen und Jungen deutlich. Zwar schnitten die Jungen geringfügig besser ab, diese Unterschiede waren jedoch nicht statistisch signifikant.

^{xv} analysis of covariance, Anm. AP

^{xvi} analysis of variance, Anm. AP

^{xvii} Unter einer Dummy-Variablen wird eine binäre Variable verstanden, die die Ausprägungen 0 und 1 aufweist. Diese steht als Platzhalter (also als Dummy) für eine Ausprägung einer mehrstufigen Variablen.

I.VII P 7: William J. Ryan (2002)

Die vorliegende Publikation beschreibt eine Untersuchung, die am Lakeland Community College durchgeführt wurde. Diese hatte zum Ziel, die Auswirkungen webbasierter Angebote auf das Lernen im Vergleich mit äquivalenten Angeboten im Klassenraum zu vergleichen.

Die hier beschriebene Primärstudie wurde zwischen 1999 und 2000 durchgeführt. Sie war als Pilotstudie anzusehen, aus der sich anschließend ein eigenständiges Forschungsprojekt entwickelte. In den Jahren 2000 und 2001 wurden im Anschluss an die Pilotstudie alle angebotenen Fernlehrkurse untersucht, allerdings wurden dabei nur die Abgangsraten und die GPAs^{xviii} miteinander verglichen. Die Ergebnisse werden in der vorliegenden Publikation mit den Ergebnissen der Pilotstudie in Beziehung gesetzt. Da jedoch nur die erste Erhebungsphase ausführlich beschrieben wird, wird in der folgenden Analyse auch nur auf diesen Teil der Untersuchung eingegangen.

Fragestellungen und Zielsetzungen

In der Pilotevaluation wurde ein einzelner Kurs untersucht, der auf drei verschiedene Arten gelehrt wurde: als traditioneller Vorlesungskurs, als Telekurs und als webbasierter Kurs.

Da angenommen wurde, dass es in Bezug auf die erbrachte Punktzahl nach Abschluss des Kurses keine Unterschiede zwischen den Darbietungsmodi geben würde, wurden verschiedene Forschungsfragen geprüft:

- Gibt es signifikante Unterschiede in der durch die Abschlussarbeit gemessenen Leistung der Studierenden bezogen auf die Instruktionsmethoden?
- Wird das Lernergebnis von Geschlecht und Alter der Lernenden beeinflusst? Spielt der Darbietungsmodus eine Rolle?
- Beeinflusst die Zufriedenheit der Studierenden mit der Darbietungsmethode das Lernergebnis?
- Werden die Lernergebnisse je nach Darbietungsmethode von Reise- und Zeitoptionen beeinflusst?^{xix}

^{xviii} Als GPA (Grade Point Average) wird im US-amerikanischen Bildungssystem eine Art Notendurchschnitt bezeichnet. Dabei wird jeder einzelnen Note ein Punktwert zugeordnet. Aus diesen wird anschließend ein gewichtetes arithmetisches Mittel berechnet (vgl. http://de.wikipedia.org/wiki/Grade_Point_Average, letzter Aufruf: 25.08. 2006).

^{xix} Hintergrund ist die größere Unabhängigkeit von Ort und Zeit bei webbasierten Kursen oder Telekursen.

- Werden die Lernergebnisse von äußerlichen, mentalen, umfeldbedingten oder technischen Hindernissen bezüglich ihrer Leistung beeinflusst?

Zusammengefasst ging es also um die Effektivität von webbasierten Kursen und Telekursen. Dies ist insbesondere vor dem Hintergrund wichtig, dass US-Bildungsinstitutionen zum Zeitpunkt der Untersuchung gerade begannen, die Einführung dieser Kursarten zu diskutieren.

Voraussetzungen

Der Kontext wird sehr detailliert beschrieben. Insbesondere der Stand der Forschung in Bezug auf Fernlehre und E-Learning wird genau dargestellt. Darüber hinaus werden die Bezüge der vorliegenden Studie zur durchführenden Institution deutlich gemacht.

Methodik

Zur Gruppenaufteilung werden keine weiteren Angaben gemacht. Der Kurs „Math 155“ wurde in drei Darbietungsformen angeboten: als traditioneller Vorlesungskurs, als Telekurs und als webbasierter Kurs. Es ist anzunehmen, dass die Studierenden jeweils einer Kursart zugeordnet waren. Wie diese Zuordnung vorgenommen wurde, ist nicht bekannt.

Datengrundlage

Die Untersuchung wurde in einem Kurs zur Einführung in die Statistik während des Schulhalbjahres 1999-2000 am Lakeland Community College in Ohio, USA, durchgeführt. Beteiligt wurden dabei nur die Studierenden, die den Vortest und das Abschlussexamen absolviert hatten. Dabei handelte es sich um eine Kohorte von 78 Studierenden, von denen 31 unter und 47 über 22 Jahre alt waren.

„Of the students who completed the two tests 31, or 40%, were in the traditional^{xx} age bracket (under the age of 22), and 47, or 60%, were in the non-traditional age bracket (22 and older). This age factor was different than other distance learning demographic research that had indicated that "the independent study population has shifted towards younger students" (Wallace, 1996, p. 1)" (William J. Ryan 2002:7).

Die Studierenden waren zu 59% weiblich und zu 41% männlich.

Ergebnisse

Die Ergebnisse der Evaluation werden beschrieben. Dabei ist festzustellen, dass keine der erwähnten Forschungsfragen signifikante Ergebnisse lieferte. In der Leistung der Studierenden gab es keine Unterschiede in Bezug auf die Darbietungsmodi. Der Autor schließt

^{xx} „Traditionell“ bezieht sich in diesem Fall auf das Durchschnittsalter der Studierenden in Fernlehrkursen, nicht auf die Lehrmethoden.

daraus, dass es den Studierenden nichts ausmachte, mit alternativen Methoden unterrichtet zu werden.

„This research indicates that there is also no harm to the learner, as measured in final course grade, by institutions using alternative instructional delivery systems. The feedback obtained in the attitudinal survey suggests that alternative delivery systems such as telecourses and web-based instruction may work to the student's advantage in terms of time and access gained. Technology offers the student the ability to shift time and increase access to instructional materials compared to traditional lecture-based delivery“ (a.a.O.:8).

Es sind darüber hinaus keine signifikanten Unterschiede in Bezug auf Alter und Lernleistung festgestellt worden. Ebenfalls führte die Messung von Studienleistungen in Bezug auf die Einstellung zu webbasiertem Lernen zu keinen signifikanten Ergebnissen.

„There was no significant difference in the interaction of the delivery methods with the pre-test used as a covariate on the final course grade with traditional and non-traditional age students. The results showed no significant difference among the three instructional delivery methods on the overall final course grade (dependent variable) when comparing traditional age students to non-traditional age students. The results also indicated no significant statistical difference in student achievement with regard to gender. Therefore, it can be concluded that for this research project there is no significant statistical difference in student achievement, as measured by overall final course grade, using the three different instructional delivery systems when comparing traditional and non-traditional age students regardless of gender“ (a.a.O.:9).

Studierende in einem webbasierten Kurs oder einem Telekurs erbrachten also keine schlechteren Leistungen als in einem traditionellen Kurs. Auch Alter und Geschlecht übten keinen Einfluss auf die Abschlussnote aus. Die Zufriedenheit der Studierenden mit der Darbietungsmethode hatte ebenfalls keine Auswirkungen auf das Lernergebnis, ebenso wenig wie Zeit- und Ortsabhängigkeit. Es wurde überdies festgestellt, dass äußere oder mentale Hindernisse, ebenso wie technische Hindernisse keine Effekte hatten.

Der Autor resümiert, dass Institutionen auf Grund dieser Ergebnisse in ihren Bestrebungen unterstützt werden, alternative Lernmethoden einzusetzen und Studierende auf erfolgreiche Fernstudienenerfahrungen vorzubereiten.

I.VIII P 8: Thommy Eriksson, Alois Goller & Sergei Muchin (2001)

Die vorliegende Studie wurde 2001 an Chalmers Tekniska Högskola in Göteborg, Schweden, durchgeführt. In erster Linie ging es dabei um die Untersuchung der Online-Kommunikation, dennoch wurden auch Angaben zu den Lernergebnissen gemacht, weshalb diese Evaluation als Primärstudie berücksichtigt wird.

Fragestellungen und Zielsetzungen

Die Evaluation zielte auf die Untersuchung der Kommunikation in einer virtuellen Lernumgebung ab. Weiterhin wurde die Zufriedenheit der Studierenden mit dem Kurs untersucht und es wurden Vergleiche in Bezug auf die Lernergebnisse angestellt.

Es wurde ein Kurs evaluiert, der in zwei Versionen verfügbar war – als konventioneller Kurs mit Workshops und Vorlesungen sowie als reiner Online-Kurs mit Kursarbeiten, Übungen und Newsgroups.

Der Kurs wurde in Teilzeit angeboten und hatte das Thema „Planning and design of web pages“, weshalb eine gewisse Affinität zum Lernen mit dem Internet vorausgesetzt werden konnte.

Voraussetzungen

Der Kurs „Planning and design of web pages“ war an Studierende gerichtet, die an kleinen oder mittelständischen Unternehmen arbeiteten. Diese Zielgruppe macht etwa 60% der Studierenden an der Chalmers Tekniska Högskola aus.

Die Untersuchung wurde im Wintersemester 1999/2000 durchgeführt. Der evaluierte Kurs wurde während neun Wochen in Teilzeit angeboten. Ziel war eine holistische Annäherung zu Wissen und Fähigkeiten, die erfordert werden, um Webseiten erstellen zu können sowie Aspekte von Internet-Marketing, Projektmanagement, Webdesign, HTML und Programmierung.

1997 wurde dieser Kurs erstmalig eingesetzt und seitdem 7 Mal durchgeführt. Jedes Semester wurde er stufenweise durch neue Ansätze, neuen Inhalt oder neue Technologien weiterentwickelt.

Die Kurswebseite verfügte über ein „schwarzes Brett“, Zeitpläne, Glossare, ein netzbasiertes Lehrbuch, Quizfragen, Anweisungen und Tests, einen Überblick über deren Ergebnisse und eine Newsgroup.

Der Präsenzkurs bestand aus Vorträgen und praktischen Übungen. Eine Unterrichtseinheit hatte die Dauer von 3 Stunden.

Den Kursteilnehmenden beider Gruppen wurde jeweils am Mittwochnachmittag eine Aufgabe gestellt, die bis zum darauffolgenden Montag zu lösen war. Bevor diese abgegeben werden konnte, musste ein Multiple-Choice-Quiz absolviert werden. Beide Kurse partizipierten darüber hinaus an der gleichen Newsgroup. Die Debatten dort wurden in

der Regel von der Lehrperson initiiert. Es war den Studierenden jedoch vorgeschrieben, mindestens eine Antwort zu geben.

Methodik

Die Untersuchung wurde mit zwei verschiedenen Lerngruppen durchgeführt. Während eine Gruppe von Studierenden den evaluierten Kurs als traditionelles Seminar mit Workshop- und Vorlesungsanteilen sowie Web-Unterstützung absolvierte, arbeitete die andere Gruppe komplett online.

„To investigate this we made a comparison and evaluation of two versions of the same short course. The course "Planning and design of web sites" is given at Chalmers University of Technology, Chalmers Lindholmen University College, both as a conventional course (ordinary lectures and workshops at location with online support) and a distance course (fully online). All students use the same web site, which have functionality such as coursework, exercises, newsgroup, chat, etc.“ (a.a.O.:1).

Wie die Zusammensetzung der Gruppen genau vollzogen wurde, wird nicht näher erläutert.

Datengrundlage

In der Untersuchung wurden die Studierenden überwiegend zur ihrer Kommunikation befragt. Es wurde zunächst ein Review der relevanten Literatur durchgeführt. Kursevaluationen fanden in jedem Semester statt^{xxi}. Darüber hinaus wurden im Wintersemester 2000 5 Lehrpersonen und 7 Studierende interviewt. Auch Transkripte, das Webportal und die Ergebnisse von schriftlichen Übungen wurden analysiert. Die Kursevaluationen basierten dabei auf folgenden Fragestellungen:

- „A Are you satisfied with how much you have learned?
- B What's your general opinion about the course?
- C What's your opinion about teacher competence?
- D What's your opinion about web portal functionality?
- E What's your opinion about web portal content?
- F What's your opinion about communication with the teachers?
- G What's your opinion about the use of the newsgroup?
- H What's your opinion about peer-to-peer communication?
- I What's your opinion about teacher communication in the newsgroup?“ (a.a.O.:4).

Es werden keine Aussagen zu Validität und Reliabilität getroffen, obwohl davon auszugehen ist, dass diese Werte bei der Untersuchung nicht vernachlässigt worden sind. Auch über die Verfahren der Datengewinnung und -analyse ist nichts bekannt.

^{xxi} In die Untersuchung fließen jedoch nur Ergebnisse aus dem Wintersemester 2000 ein.

Ergebnisse

Neben den Ergebnissen, die zum Aspekt der Kommunikation vorgestellt werden^{xxii}, gibt es auch Resultate, die den Lernerfolg betreffen.

Demnach waren die Studierenden im Online-Kurs zufriedener mit den eigenen Lernergebnissen und der Funktionalität der Webseite. Daraus wird der Schluss gezogen, dass sich diese Studierenden besser an das Lernen mit einer Lernumgebung angepasst hätten:

Die Autoren äußern ihre Verwunderung darüber, dass diese Aussagen getroffen worden seien, obwohl der Online-Kurs mit mehr technischen Problemen konfrontiert gewesen sei als der Präsenzkurs:

„Could it be that the distance education students was more engaged in their own learning and had a more mature attitude towards their own learning and their own responsibility for it?“
(ebd).

Unterschiede zwischen den Lernergebnissen, die in den Übungen erhoben worden sind, wurden nicht beobachtet. Es gibt keine Hinweise darauf, dass die Online-Studierenden mehr Schwierigkeiten hatten, die Übungen zu verstehen oder abzuschließen.

Die wichtigste Schlussfolgerung der Studie ist die Erkenntnis, dass die gefundenen Unterschiede zwischen den Gruppen – wenn überhaupt vorhanden – nur sehr klein waren.

Die Ergebnisse wurden auf Signifikanz untersucht. Dabei wurden keine signifikanten Unterschiede zwischen den Gruppen festgestellt:

^{xxii} Der wichtigste Vorteil der computervermittelten Kommunikation sind der Zeitaspekt im Allgemeinen und die Asynchronität im Speziellen. Dies wird sowohl von Lehrenden, als auch von Lernenden kommuniziert. Die Präsenzgruppe war mit der Kommunikation zufriedener als die Online-Gruppe.

I.IX P 9: Ruth Green & Karen Gentemann (2001)

Die vorliegende Untersuchung wurde im Wintersemester 1999/2000 an der George Mason University im US-Bundesstaat Virginia durchgeführt. Dabei ging es den Autorinnen in erster Linie darum, die Unklarheit der unterschiedlichen Schlussfolgerungen, die in der Literatur in Bezug auf die Effektivität von Online vs. Präsenzlernen gezogen wurden, durch weitere Forschung aufzulösen.

Die Autorinnen legen eine Längsschnittstudie vor, in der die Unterschiede zwischen Studierenden in einem Online- und einem Face-to-Face-Kurs zu zwei Zeitpunkten erhoben wurden. Dabei wurden sowohl die Einstellungen zum Kurs als auch die Leistungen der Lernenden betrachtet.

Fragestellungen und Zielsetzungen

Aufbauend auf bereits vorhandenen Untersuchungsergebnissen in der Literatur möchten die Autorinnen die Effektivität von Online-Kursen durch ein Evaluationsdesign untersuchen, das nicht nur auf den reinen Vergleich der Leistung ausgerichtet ist.

„The study examines differences between the OLC students at the beginning and end of the semester (e.g., attitudes about OLC and FTF courses, perceived levels of technological expertise, use of technology) and differences in academic outcomes (e.g., course completion, grade) for students in the OLC vs. FTF (Ruth Green & Karen Gentemann 2001:3).

Die Untersuchung zielte darauf ab, die Leistungen und Einstellungen der Studierenden im Online-Kurs „advanced English“ zu zwei unterschiedlichen Zeitpunkten im laufenden Semester festzustellen und miteinander zu vergleichen. Als Kontrollgruppe diente der reguläre Kurs in Präsenzform. Beide hatten den gleichen Umfang (3 Kreditpunkte).

In Präsenzform bestand der Kurs in erster Linie darin, das Schreiben und Analysieren von Texten zu trainieren.

„English 302: Advanced Composition is a required 3-credit course with extensive pre-requisites (Completion of 45 credit hours, three credits of composition, and 6 credits of literature). The course provides intensive practice in writing and analyzing such expository forms as the essay, article, proposal, and technical and scientific report. Students must maintain a C average to fulfill degree requirements“ (ebd.).

Als Online-Angebot wurde der Kurs durch Kurs-Management-System mit einer Web-CT-Installation angeboten.

„The OLC site provided a course syllabus, academic rules and regulations, assignments and links to materials needed for each assignment. The students enrolled in the OLC were also connected via a class e-mail system (file exchange). Assignments varied but were typically scenario/case studies that required students to provide on-line, research-based responses using the university data bases and other materials identified on the website. For example, one assignment provided students a mock memo with an identified problem from a work supervisor. Students were asked to respond

via electronic memo with a research-based response and recommendations for the supervisor" (a.a.O.:4).

In zwei der drei Online-Klassen war ein Face-to-Face-Treffen zu Beginn und am Ende des Kurses vorgesehen, eine Klasse führte dies nur zu Beginn durch.

Voraussetzungen

Zunächst findet ein Review der relevanten Literatur statt. Hierbei wird insbesondere auf die Ergebnisse von Vergleichsstudienforschung sowie auf die didaktische Qualität von Online-Lernen Bezug genommen. Darüber hinaus werden die Rahmenbedingungen der Universität und die typische Zielgruppe erläutert.

Auf pädagogische Voraussetzungen der Evaluation oder Vorwissen der Studierenden wird nicht speziell eingegangen.

Methodik

Als Untersuchungsgruppe diente der Kurs „English 302: Advanced Composition“, der in drei als Online-Kurs mit begrenzten Präsenzeinheiten unterrichtet wurde:

„The three sections were OLC with limited face-to-face contact. One class met only at the beginning of the course and two classes met at the beginning and again at the end“ (a.a.O.:3)

An diesem Kurs nahmen 57 Studierende teil, von denen 47 während des gesamten Semesters im Kurs verblieben. 52 Studierende beteiligten sich zu Beginn des Semesters an der Untersuchung, 30 Studierende am Ende des Kurses.

Die Kontrollgruppe bildeten 12 Einheiten des gleichen Kurses, der als traditioneller Präsenzkurs abgehalten wurde:

„In contrast, there were 12 sections of English 302 that were traditional FTF courses that met at different times during the week for varying periods of time as shown below“ (ebd.).

Über die Anzahl der Studierenden im Präsenzkurs sowie über die Verfahren der Gruppenzusammensetzung werden keine Angaben gemacht.

Zur Messung der Effektivität wurden verschiedene Methoden miteinander kombiniert. Zum einen wurden schriftliche Befragungen zu Beginn und am Ende des Kurses durchgeführt^{xxiii}. Diese betrafen die Bereiche „educational strategies“, „experiences with technology“, „student academic and demographic information“ und schlossen darüber hinaus Interviews und Gruppenprotokolle ein. Hierbei ging es primär um den persönlichen Hintergrund und Erwartungen gegenüber dem Unterricht.

^{xxiii} Die Umfragen wurden mit Hilfe des Flashlight-Inventory erstellt. Hierbei handelt es sich um ein Evaluationswerkzeug, das mit externer Unterstützung entwickelt wurde.

In den Fragebögen waren verschiedene Skalen enthalten, die auf die Bereiche „Importance of academic/social interactions“, „Expertise with technology“, „Student expectations for instructor priorities“, „Use of Seven Principles of Good Practice“, „Relationships and contacts with other students“, „Course characteristics“, „Course procedures“, „Course feedback“, „Benefits of OLC“ und „Benefits of FTF“ abzielten.

Die Antworten wurden mit Hilfe einer statistischer Methoden ausgewertet. Signifikanztests wurden in Bezug auf die Abschlussnote, Bestehensquote, Geschlecht und ethnischer Herkunft gegenüber der Präsenzklasse durchgeführt.

„For the variables that compared the students in the OLC with the FTF, data were obtained using institutional databases. OLC and FTF groups were compared based on gender, age, cumulative and course GPA and course retention/attrition rates. [...] Average course grade and retention/attrition rates were compared for all OLC and FTF students enrolled in fall 1999 in the 15 sections of English 302 designated for business majors (OLC courses n=3, FTF n=12 courses)“ (a.a.O.:12f.).

Datengrundlage

Die Studierenden wurden mittels Fragebogen zur Einschätzung der Instruktionsmethode und Vergleich der Leistungen getestet. Dabei beinhaltete der Fragebogen diverse Skalen (Importance of academic/social interactions, Expertise with technology, Student expectations for instructor priorities, Use of Seven Principles of Good Practice, Relationships and contacts with other students, Course characteristics, Course procedures, Course feedback, Benefits of OLC, and Benefits of FTF), von denen im Forschungsbericht jeweils die Reliabilität (Cronbach's Alpha) angegeben waren.

Ergebnisse

Die Ergebnisse wurden auf Signifikanz untersucht. Dabei wurden in Bezug auf den Lernerfolg keine signifikanten Unterschiede zwischen den Gruppen festgestellt:

„There were no significant differences between students in the OLC compared to students in the FTF courses for course retention“ (a.a.O.:15).

Die Längsschnittuntersuchung innerhalb der Online-Kurse ergab, dass alle Studierenden über einen geeigneten Computer verfügten und ihnen darüber hinaus die erforderliche Zeit zum Bearbeiten der Kurse zur Verfügung stand. Ferner hatte sich der Großteil der Studierenden eingeschätzt, die erforderlichen technischen Fähigkeiten zu besitzen.

In Bezug auf Online-Lernen wurde deutlich, dass diesem eher realistische Erwartungen zugeschrieben wurden – es sei nicht über-, aber auch nicht unterzubewerten. Als Vorteile wurden hochwertiges Feedback und individuelle Betreuung genannt. Nachteilig empfanden die Studierenden den Zeitaspekt.

I.X P 10: David J. Pucel & Thomas F. Stertz (2005)

In dieser Studie wurde die Zufriedenheit der Studierenden mit zwei Online-Kursen und deren Leistung im Vergleich zu zwei gleichwertigen Präsenzkursen untersucht. Zielgruppe waren Lehramtsstudierende der University of Minnesota.

Fragestellungen und Zielsetzungen

Die vorliegende Arbeit zielt darauf ab, die Zufriedenheit der Studierenden mit zwei Online-Kursen und deren Leistung im Vergleich zu zwei gleichwertigen Präsenzkursen zu untersuchen. Dabei wurde das Ziel verfolgt, ein generalisierbares Modell zur Evaluation von Online-Kursen im Vergleich zu traditionellen Kursen zu entwickeln und gleichzeitig festzustellen, ob diese gleiche Leistungen der Studierenden hervorbringen würden.

Die Autoren geben darüber hinaus an, generell mehr Forschung in Bezug auf die Messung der Effektivität von Online-Kursen durchführen zu wollen. Dies könne durch die Lernleistung und die Zufriedenheit der Studierenden realisiert werden.

Voraussetzungen

Die Untersuchung wurde zwischen den Jahren 2001 und 2003 an insgesamt 84 Lehramtsstudierenden in (44 in Online-Kursen und 40 im Präsenzunterricht) durchgeführt, die zu dieser Zeit die Kurse „History and Philosophy of Career and Technical Education“ und „Instructional Methods for Business and Industry“ an der George Mason University belegt hatten. Diese Kurse wurden von zwei Doktoranden des Faches konzipiert und geleitet.

Die Autoren führten zunächst ein Review der relevanten Literatur durch, welche sich primär auf die Lehrerbildung sowie Web-Based-Instruction und Lernleistungen von Studierenden bezog. Ferner wird auf die Spezifika von Vergleichsstudienforschung eingegangen.

Darüber hinaus werden die Rahmenbedingungen Lehrerbildung an der George Mason University erläutert:

„Due to a continuing lack of qualified individuals who are both occupationally competent and possess teaching degrees, schools hire occupationally competent people and provide them with teacher education as they are teaching. The occupational teachers teach in fields such as communications technology, construction careers, creative design careers, early childhood careers, hospitality service careers, manufacturing careers, medical careers, and transportation careers. These new teachers are sparsely spread out throughout the state of Minnesota [...] Therefore, the Minnesota Department of Education asked the University of Minnesota to develop online- enhanced WBI course alternatives for some of its basic pedagogical courses [...]. It was hoped that offering the classes in a WBI format would reduce the amount of travel and in-classroom time and provide flexibility for students completing course requirements without compromising either student

satisfaction or performance”(David J. Pucel & Thomas F. Stertz (2005:11f.).

Methodik

Bei der vorliegenden Evaluation handelt es sich um eine summative Querschnittsuntersuchung, wenngleich die Messungen zu unterschiedlichen Zeitpunkten durchgeführt worden waren. Über die Zusammensetzung von Test- und Kontrollgruppen werden keine Angaben gemacht. Es wird lediglich darüber berichtet, dass zwei Online- und zwei Präsenzkurse miteinander verglichen wurden, von denen jeweils die Online- und Präsenzversion das selbe Kursmaterial und denselben Kursinhalt hatten:

1. „History and Philosophy of Career and Technical Education“ als Online-Kurs
2. „Instructional Methods for Business and Industry“ als Online-Kurs
3. „History and Philosophy of Career and Technical Education“ als Präsenzkurs
4. „Instructional Methods for Business and Industry“ als Präsenzkurs

Dabei waren Ziele und Inhalte der jeweiligen Online- und Präsenzversionen nahezu identisch, da die Inhalte der traditionellen Kurse übertragen wurden:

„To ensure that the instruction in the WBI-enhanced versions of the courses was as nearly equivalent as possible to the past traditional versions, two instructors with Ph.D.'s in education who had long histories of teaching the courses in a traditional manner, and who were interested in online instruction, re-formatted the courses for WBI. Specialists in the development of online instruction from the university's College of Education and Human Development provided guidance in using WebCT to assemble the material for online delivery“ (a.a.O.:12).

Der Präsenzkurs in Philosophie wurde bereits während des Wintersemesters 2002 untersucht, der Präsenzkurs in Methoden im Sommer 2001. Die Online-Kurse fanden im Jahr 2003 statt. Die Gründe für die Wahl dieses Evaluationsdesigns werden nicht begründet.

Auch die Online-Kurse hatten einen Präsenzanteil von insgesamt sechs Stunden, in denen beispielsweise die Präsentationen erstellt oder die Prüfungen absolviert wurden. Die technische Umsetzung wurde mit WebCT realisiert:

„Each WBI course contained a detailed syllabus, instructional modules designed to lead students through the learning activities and related resources, project guidelines and expectations, and exercises that required students to participate in cooperative learning activities. In addition, the course materials included Web links, video clips, audio clips, and PowerPoint presentations“ (a.a.O.:12).

Datengrundlage

Die Untersuchung basiert vorwiegend auf Test- oder Abschlussnoten, die im Laufe des Semesters von den Studierenden erbracht wurden.

„Grading information included student project scores, final test scores, and final grades gathered during each WBI course“ (ebd.).

In jedem Kurs wurden vier Erhebungen durchgeführt und zwar in Form von zwei Projekten, einem Abschlusstest und einer Gesamtnote. Darüber hinaus wurde die Zufriedenheit der Studierenden mit dem Kurs durch den universitären Fragebogen zur Kursevaluation erhoben. Als Vergleichsdaten wurden Ergebnisse früherer Jahrgänge in Präsenzkursen herangezogen. Alle Kursarten wurden mit dem gleichen Instrument bewertet, das im Wesentlichen aus sechs Dimensionen bestand:

- „1. Instructor’s overall ability?
2. Instructor’s knowledge of the subject matter?
3. How much would you say you learned from this course?
4. Overall quality of tests and handouts?
5. Helpfulness of feedback to you about your performance?
6. Degree to which evaluation procedures measured your knowledge and understanding?“ (a.a.O.:18)

Zusätzlich zu diesen standardisierten Verfahren wurden die Studierenden im Online-Kurs aufgefordert, ihre Zufriedenheit zu beschreiben, anzugeben, wieviel Zeit sie dafür verwendet hatten und die Schwierigkeit des Kurses im Vergleich mit ihren Erfahrungen im Präsenzkurs anzugeben.

Ergebnisse

In den Ergebnissen der Evaluation wird u.a. deutlich, dass der Zeitaufwand für die Bearbeitung des Online-Kurses von den Studierenden höher eingeschätzt wurde als für den Präsenzkurs. Im Gegensatz dazu fiel die Arbeit jedoch leichter. Insgesamt konnte eine überwiegende Zufriedenheit der Studierenden mit den Online-Kursen festgestellt werden. Die Leistung der Studierenden war im Kurs Philosophie in der Online-Version signifikant besser, im Methodenkurs in der Präsenzversion. Die Autoren geben jedoch an, dass die praktische Bedeutsamkeit dieser Ergebnisse relativ gering gewesen sei:

„Although some statistically significant differences were found on some of the student performance measures, the practical differences between the WBI and traditional versions of the courses were not large. The courses were graded using a criterion referenced system in which students were not graded against one another but rather against a set of standards. Given this system, it

is possible for all students to receive very high grades or very low grades depending upon their performance“(a.a.O.:21).

Insgesamt kann als Ergebnis der Studie festgehalten werden, dass in allen 4 Kursen gute Leistungen erbracht werden konnten.

Ferner wurde die Leistung der Studierenden untersucht. Diese fiel in der Online-Version des Philosophie-Kurses schlechter aus als in der Präsenzform. Hier ergaben sich signifikante Unterschiede zwischen den Kursversionen:

„In two cases the differences were statistically significant: 1) performance on the mission statement project, and 2) final grades“(a.a.O.:19).

Dem steht jedoch gegenüber, dass diese Ergebnisse für die Methodenkurse nicht bestätigt werden konnten. Hier zeigte sich eine signifikante Überlegenheit des Präsenzunterrichts gegenüber der Online-Version.

I.XI P11: Larry S. Tinnerman (2006)

In dieser Primärstudie wurden zwei Kurse miteinander verglichen, die sich mit Lernbehinderungen befassen. Beide wurden von derselben Lehrperson gegeben, allerdings wurde ein Kurs als Fernlehrrangebot online unterrichtet, der andere in einem traditionellen Klassenraum-Szenario. Neben dem Vergleich des Lernerfolgs und der Effektivität von Online-Lernen wurden Faktoren wie Kommunikation, Zeit, Kontinuität sowie technische Erfahrungen untersucht.

Fragestellungen und Zielsetzungen

Es sollten zwei Kurse miteinander verglichen werden, die von der gleichen Lehrperson sowohl in einem Präsenzkurs als auch online gelehrt wurden. Das Forschungsinteresse lag hierbei sowohl auf den Lernleistungen der Studierenden als auch auf ihren Meinungen und Haltungen bezüglich der Kursdidaktik und -organisation.

Beispielweise wurde zusätzlich erhoben:

- „The awkwardness of the communication effectively using a nonverbal typed method such as chat or discussion forum
- The compatibility of various software packages
- The difficulty of teacher modeling for the students
- The demand for the instructor to be both a content expert and a technical expert“ (Larry S. Tinnerman 2006:o.S.).

Die Untersuchung wurde an Hand von zwei Kursen zum Thema Lernbehinderungen durchgeführt. Dabei wurde die Instruktion entweder im Klassenraum, oder online gegeben.

Voraussetzungen

Der Kontext der Untersuchung wird durch Bezüge zum Stand der Forschung in Bezug auf E-Learning und Fernlehre dargestellt. Diese werden in Beziehung zum Wandel der akademischen Ausbildung gesetzt. Insbesondere wird auf die Relevanz der Forschung hinsichtlich Behinderung und Rehabilitation hingewiesen.

Die Rahmenbedingungen der Studie werden erläutert. In einer demographischen Erhebung wird auf die speziellen persönlichen Hintergründe der Studierenden eingegangen.

Methodik

An der Studie nahmen insgesamt 46 Studierende teil, die ihrer Einschreibung entsprechend in die Online- oder die traditionelle Gruppe eines Kurses zu Lernbehinderungen aufgeteilt wurden.

„Students were selected for this study using a census approach as constituted by their enrollment in the course by either traditional or distance instruction. It was this researcher’s intent to get as close to

total participation from all the students in both settings as possible“ (ebd.).

Bei der vorliegenden Evaluation handelt es sich um eine summative Längsschnittuntersuchung.

Die Zufriedenheit und Leistungen der Studierenden wurden sowohl in den Online- als auch in den Präsenzversionen der Kurse mittels derselben Kriterien evaluiert.

Der Lernerfolg wurde durch Vor- und Nachtests ermittelt. Dieser bestand aus 20 Multiple-Choice-Fragen aus der Kursdatenbank für die Kapitel 1 bis 6 des zu lernenden Stoffes. Beide Kurse wurden den gleichen Tests unterzogen.

„Students in the on-line program were adhering to the specified conditions of both the pre and post tests, i.e. that both tests were to be taken without books or notes“ (ebd.).

Datengrundlage

In der Studie wurden sowohl ein experimenteller als auch ein deskriptiver Ansatz verfolgt und integriert.

„Experimental - The independent variable was the method of instructional delivery, i.e. the traditional classroom instructional approach or the distance method of instruction. Enrollment in the class, the instructor and course content were the dependant variables. Since the class was offered by the same instructor to both groups at the same time, this provided an excellent opportunity to examine the similarities and differences between the groups in the areas of content delivery, student performance and student satisfaction“ (ebd.).

Bei der deskriptiven Vorgehensweise wurden vor allem Einstellungen und Meinungen der Studierenden untersucht. Hierzu wurden Likert-Skalen und offene Fragen benutzt.

Folgende Instrumente wurden eingesetzt:

- Vortest: 20 Multiple-Choice-Fragen aus der Kursdatenbank für die Kapitel 1 bis 6. Dieser Test wurde mit beiden Untersuchungsgruppen durchgeführt.
- Demographischer Fragebogen: basierend auf den Forschungsfragen.
- Student Perception Survey: ebenfalls basierend auf den Forschungsfragen.
- Nachtest: 20 Multiple-Choice-Fragen aus der Kursdatenbank für die Kapitel 1 bis 6.

Folgende Bereiche wurden u.a. im Perception Survey abgefragt:

- Zeitmanagement
 - „Perceived time spent reading class text
 - Perceived time spent reading supplemental assigned materials

- Perceived time spent writing assignments for class
- Perceived time spent in communication with instructor
- Perceived time spent in class research activities
- Perceived time spent in dealing with technology issues
- Perceived time spent in communication with other class members
- Perceived time spent each week in total class activities“ (ebd.)
- **Kommunikation**
 - „Accessibility of other students in your class
 - Effectiveness of student to student discussions to progress or interest in class
 - Frequency of contacting other classmates in completion of assignments
 - Other classmates having a broad scope of knowledge that helps make the course topics relevant
 - Contacting other classmates during the week in a social or causal context“ (ebd)
- **Hilfe und Feedback**
 - „Instructor accessibility
 - Prompt and helpful feedback from the instructor in a timely manner
 - Instructional methods learn and understandable
 - Instructor sympathetic to needs and situations of a graduate student
 - Instructor provides ample opportunity for questions and the voicing of concerns and opinions“ (ebd.)
- **Kurs-Effizienz und Relevanz**
 - „Factors beyond the control of the student, instructor or university that contribute to frustration and confusion
 - Relevance of material presented in regards to personal career goals
 - Generalizability of material to other professional and or personal life endeavors
 - Material found to be thought provoking and intellectually stimulating
 - Satisfaction as to how class is progressing
 - Perception of the difficulty of participation in the on-line version of this class
 - The course content is clear and understandable“ (ebd.)

Die Befragungen wurden anonym durchgeführt – den Studierenden wurden zur Vergleichbarkeit der Daten jedoch Identitätsnummern zugeteilt.

Ergebnisse

Es kann festgehalten werden, dass im traditionellen Unterricht mehr Interaktion stattgefunden hat. Die Studierenden (in der traditionellen Klasse!) erachteten dabei die freie Zeiteinteilung beim Lernen in der Online-Klasse als günstig. Allerdings wurde auch die Möglichkeit, Fragen zu stellen in der Präsenzkasse begrüßt. Es wurden Verbesserungsvorschläge seitens der Studierenden gemacht, die sich vor allem auf inhaltliche Aspekte (z.B. weniger Zeit für Wiederholungen anzusetzen) bezogen.

Die Online-Studierenden beurteilten ebenfalls die Möglichkeit der freien Zeiteinteilung als gut, sahen jedoch auch die Vorteile der Face-to-Face-Interaktion im Klassenraum. Die Arbeit empfanden sie als generell zeitintensiver und anstrengender in Bezug auf die Recherche nach relevanten Informationen. Die Verbesserungsvorschläge der Studierenden bezogen sich in erster Linie auf die Erreichbarkeit des Lehrenden, die Menge der Arbeit und die Organisation des Kurses.

Die Leistungen der Studierenden fielen in beiden Klassen gleich aus. Die Analyse der Testdaten wurde mittels eines T-Tests durchgeführt. Es wurden die Leistungen von 15 Online-Studierenden und 22 Studierenden im Präsenzkurs miteinander verglichen^{xxiv}.

Wenngleich in beiden Gruppen signifikante Anstiege der Lernleistung vor und nach der Unterrichtseinheit festgestellt werden konnten, ergab die Analyse keine signifikanten Unterschiede zwischen den Untersuchungsgruppen.

Die Ergebnisse wurden auf Signifikanz untersucht. Dabei wurden in Bezug auf den Lernerfolg keine signifikanten Unterschiede zwischen den Gruppen festgestellt:

„There was no significant difference between the performance on the pretest for the distance education group and the pretest for the traditional education group at least $p=.05$

There was no significant difference between the performance on the post-test for the distance education group and the post test for the traditional education group at least $p=.05$ “ (ebd.).

^{xxiv} Die restlichen Studierenden beteiligten sich nur an den Fragebogenerhebungen.

II Bewertung der Primärstudien nach den übrigen Kriterien

II.I Erläuterung der Bewertungskriterien

Im Folgenden wird insbesondere auf die Kriterien eingegangen, die in der Arbeit nicht näher untersucht wurden, da sie nicht ausreichend auf die Forschungsfrage fokussiert waren. Um zu prüfen, ob die Studien den Evaluationsstandards entsprechen, ist eine Prüfung jedoch auch nach diesen Kriterien sinnvoll.

Kriterien der Nützlichkeit

In der Erläuterung der Deutschen Gesellschaft für Evaluation (2001) wird angegeben, dass

„die Nützlichkeitsstandards [...] sicher stellen [sollen], dass die Evaluation sich an den geklärten Evaluationszwecken sowie am Informationsbedarf der vorgesehenen Nutzer und Nutzerinnen ausrichtet“ (ebd.).

Dazu gehören:

Identifizierung der Beteiligten und Betroffenen

„Die am Evaluationsgegenstand beteiligten oder von ihm betroffenen Personen bzw. Personengruppen sollen identifiziert werden, damit deren Interessen geklärt [...] werden können“ (ebd.).

Die Formulierung der Deutschen Gesellschaft für Evaluation geht an dieser Stelle weiter. Für die vorliegende Arbeit ist die Beschreibung des Kriteriums m.E. jedoch ausreichend. Thomas Widmer (1996:15) legt dar, welche Personengruppen im Einzelnen gemeint sind: Entscheidungsträger und Programmgestaltende, Programmfinanzierende, Evaluationsfinanzierende, Zielgruppe des Programms, Programmverantwortliche, EvaluatorInnen, Programm-Mitbewerbende, Evaluationsgemeinde (vgl. ebd.).

Klärung der Evaluationszwecke

„Es soll deutlich bestimmt sein, welche Zwecke mit der Evaluation verfolgt werden, so dass die Beteiligten und Betroffenen Position dazu beziehen können und das Evaluationsteam einen klaren Arbeitsauftrag verfolgen kann“ (Deutsche Gesellschaft für Evaluation 2001).

Klarheit der Veröffentlichung

Die Veröffentlichung der Primärevaluation

„[...] sollte das evaluierte Objekt und den Kontext ebenso beschreiben, wie die Absichten, die Verfahren und die Ergebnisse der Evaluation [...]“ (Thomas Widmer 1996:18).

Dies ist notwendig, damit LeserInnen verstehen können,

„was gemacht worden ist, warum dies getan wurde, welche Informationen gewonnen, welche Schlüsse gezogen und welche Empfehlungen gemacht wurden“ (ebd.).

Kriterien der Genauigkeit

„Die Genauigkeitsstandards sollen sicher stellen, dass eine Evaluation gültige Informationen und Ergebnisse zu dem jeweiligen Evaluationsgegenstand und den Evaluationsfragestellungen hervorbringt und vermittelt“ (Deutsche Gesellschaft für Evaluation 2001).

Beschreibung des Evaluationsgegenstandes

„Der Evaluationsgegenstand soll klar und genau beschrieben und dokumentiert werden, so dass er eindeutig identifiziert werden kann“ (ebd.).

Kontextanalyse

„Der Kontext des Evaluationsgegenstandes soll ausreichend detailliert untersucht und analysiert werden“ (ebd.).

Beschreibung von Zwecken und Vorgehen

„Gegenstand, Zwecke, Fragestellungen und Vorgehen der Evaluation, einschließlich der angewandten Methoden, sollen genau dokumentiert und beschrieben werden, so dass sie identifiziert und eingeschätzt werden können“ (ebd.).

Übertragbarkeit

„Die Übertragbarkeit [...] der Evaluationsbefunde ist deren Anwendbarkeit in der Programmpraxis“ (Bernhard Bührlen-Armstrong (1998:127). Dazu gehört der Bezug zum Programm, der gesellschaftlicher Kontext sowie die Nützlichkeit für das Programm (vgl. ebd.).

II.II Operationalisierung der Kriterien

Bei der Operationalisierung werden Indikatoren formuliert, die zur Klärung der Erfüllung eines Kriteriums beitragen. So dienen beispielsweise die Indikatoren „Die Veröffentlichung ist verständlich“, „Die Veröffentlichung folgt einer inhaltlichen Struktur“ und „Es bleiben keine offenen Fragen“ dazu, festzustellen, ob bzw. in welchem Ausmaß das Kriterium „Klarheit der Veröffentlichung“ erfüllt wird.

Jeder einzelne Indikator wird mit einer Skala von „trifft voll und ganz zu“ bis „trifft überhaupt nicht zu“ bewertet.

Nützlichkeit

Identifizierung der Beteiligten und Betroffenen

- Es wird deutlich, welche Personen an der Evaluation beteiligt waren.
- Es ist klar beschrieben, auf welchen Kurs/ welche Lerneinheit/ welche Lernenden sich die Untersuchung bezog.

Klärung der Evaluationszwecke

- Das Ziel der Untersuchung ist eindeutig angegeben.
- Der Zweck der Untersuchung wird begründet.

Klarheit der Veröffentlichung

- Die Veröffentlichung ist verständlich.
- Die Veröffentlichung folgt einer inhaltlichen Struktur.
- Es bleiben keine offenen Fragen.

Genauigkeit

Beschreibung des Evaluationsgegenstandes

- Der Evaluationsgegenstand wird präzise beschrieben.
- Es wird begründet, welche Kursarten getestet wurden.
- Es ist nachvollziehbar, welche Kursarten getestet wurden.
- Es wird begründet, welches Kursmaterial getestet wurde.
- Es ist nachvollziehbar, welches Kursmaterial getestet wurde.

Kontextanalyse

- Pädagogische Voraussetzungen der Evaluation werden beschrieben (z. B. Vorwissen der ProbandInnen).
- Es werden technische Voraussetzungen erläutert (z.B. Zugang zu Kursinhalten).
- Es werden Rahmenbedingungen (z.B. Kursgröße, Themen) beschrieben.

- Es wird deutlich, welchen theoretischen Überlegungen (didaktischen, pädagogischen) die Evaluation folgt.

Beschreibung von Zwecken und Vorgehen

- In der Primärstudie wird begründet, welches Evaluationsdesign zu Grunde gelegt wurde (Längs- oder Querschnitt, summative oder formative Evaluation).
- Die Methoden der Untersuchung werden begründet.

Übertragbarkeit

- Es wird angegeben, ob die Ergebnisse der Evaluation auf andere Bereiche übertragbar sind.
- Die Übertragbarkeit ist nachvollziehbar.
- Es werden Beispiele aus anderen Bereichen angegeben.

II.III Bewertung der Primärstudien

II.III.I P 1: Edward R. Jones et al. (1999)

II.III.I.I Nützlichkeit

Identifizierung der Beteiligten und Betroffenen

Die Untersuchung wurde an 89 Studierenden in zwei Gruppen eines Statistikkurses an der Texas A&M University Corpus Christi durchgeführt. Davon wurde die eine Gruppe in traditioneller Art und Weise unterrichtet, die andere ausschließlich virtuell.

In der Veröffentlichung werden Kurs und Lerneinheit genannt. Das Kriterium konnte also überwiegend erfüllt werden und wird als eher positiv gewertet.

Klärung der Evaluationszwecke

Hintergrund der Studie ist eine vorliegende Arbeit aus dem Jahr 1997^{xxv}, die zeigte, dass Studierende in einem virtuellen Klassenzimmer^{xxvi} signifikant höhere Lernergebnisse erzielten als Studierende in einer traditionellen Klasse. Laut dieser Studie von Jerald G. Schutte (1997) betrage dieser Unterschied sogar 20%. Nachdem es verschiedene methodische Kritik gegeben hatte, griffen Edward R. Jones et al. (1999) das Thema erneut auf und beschlossen, die Ergebnisse von Jerald G. Schutte in einer eigenen Untersuchung zu replizieren.

Dabei wurden drei Absichten verfolgt:

1. Es sollten die Leistungsunterschiede zwischen Studierenden in einem webbasierten und einem traditionellen Kurs vor dem Hintergrund evaluiert werden, dass die Studierenden selbst die Wahl gehabt hatten, welchen dieser beiden Kurse sie besuchen wollten.
2. Darüber hinaus sollte die Hypothese untersucht werden, dass der Hintergrund der Studierenden, die einen Webkurs vorziehen, keine Unterschiede zu dem Hintergrund der Studierenden aufweist, die einen traditionellen Kurs bevorzugen.
3. Es sollte identifiziert werden, inwiefern Vorteile der Technologie (falls vorhanden) in einer webbasierten Klasse den Lernerfolg in einem einführenden Statistikkurs erhöhen (vgl. Edward R. Jones et al. (1999:2).

Es wird nicht explizit erwähnt, ob es ein Ziel der Evaluation darstellte, die Ergebnisse für eigene Zwecke nutzbar zu machen. Da in der

^{xxv} Schutte, J.G. (1997): Virtual teaching in Higher Education: The new Intellectual Superhighway or Just Another Traffic Jam? Retrieved November 29, 1998 from WWW: <http://www.csun.edu/sociology/virexp.htm>.

^{xxvi} Im Original wird der Begriff „Virtual classroom“ verwendet.

Publikation jedoch der Schluss gezogen wird, dass virtuelle Lehre an der Universität stärker vorangetrieben werden soll, ist diese Frage sicherlich auch als Ziel anzusehen.

Das Kriterium wurde voll erfüllt und wird als positiv gewertet.

Klarheit der Veröffentlichung

Die Veröffentlichung folgt einer inhaltlichen Struktur und ist gut verständlich. Es bleiben jedoch inhaltliche Fragen, da in manchen Fällen Begründungen zu Forschungsentscheidungen fehlen.

Das Kriterium wurde überwiegend erfüllt und wird als eher positiv gewertet.

II.III.I.II Genauigkeit

Beschreibung des Evaluationsgegenstandes

Die untersuchte Maßnahme war das virtuelle Klassenzimmer im Vergleich mit traditionellem Unterricht. Über das Kursmaterial wird berichtet, dass anspruchsvolle Lektionen als PowerPoint Vortrag mit synchronisiertem Audio angeboten worden seien. Generell handelte es sich jedoch nur um „Übersetzungen“ traditionellen Materials. Da weitergehende Angaben fehlen, ist festzuhalten, dass das Kursmaterial an dieser Stelle nicht ausreichend beschrieben ist. Darüber hinaus werden keine Begründungen zu den Kursarten angeführt.

Das Kriterium wurde in vielen Bereichen nicht erfüllt und wird daher als eher negativ gewertet.

Kontextanalyse

Voraussetzung für die Einschreibung in den Kurs war das Bestehen eines Vortests^{xxvii} in Algebra. Ebenfalls als eine Voraussetzung der Untersuchung ist anzusehen, dass die Gruppenzuweisung auf freiwilliger Basis erfolgte und nicht forciert wurde.

Darüber hinaus ist es in Bezug auf die zu Rate gezogenen Abbruchquoten wichtig zu wissen, dass es Studierenden an der A&M University Corpus Christi ohne jegliche Sanktionen erlaubt ist, sich bis zum letzten Tag des Semesters aus einem Kurs wieder auszutragen, daher sind hohe Abbruchquoten in der Regel nicht ungewöhnlich. Im Gegenteil: Wer für den eigenen Notendurchschnitt noch ein gutes Ergebnis benötigt, trägt sich aus den belegten Kursen wieder aus, wenn absehbar ist, dass das gewünschte Ergebnis nicht erzielt werden kann. Insofern liefert ein hoher Notendurchschnitt in der Abschlussklasse eines Kurses hier ein sehr verzerrtes Bild.

Es wird davon ausgegangen, dass diejenigen, die sich in den virtuellen Kurs eingeschrieben hatten, vertrauter im Umgang mit Computer und Internet waren als die Studierenden im traditionellen

^{xxvii} Test A4/F der Mathematical Association of America.

Kurs. Da der Kontext gut beschrieben wird, ist das Kriterium als voll erfüllt und damit positiv zu werten.

Beschreibung von Zwecken und Vorgehen

Das Untersuchungsdesign entsprach im Wesentlichen der Vorgängerstudie von Jerald G. Schutte (1997). In beiden Fällen setzte sich die Stichprobe aus Studierenden eines einführenden Statistikkurses zusammen. Auch die Unterrichtsmethode war mit der Vorgängerstudie vergleichbar:

“Likewise, the web-based class used the same general means employed by Schutte - lecture notes, on-line quizzes, a discussion listserver, email and biweekly live web classes (chat sessions). [...] In addition, the lecture notes over the more difficult material were served as a PowerPoint lecture with synchronized audio; Real Audio compressed to server over a 28.8k modem connection” (Edward R. Jones et al. 1999:3).

Die Evaluation verlief summativ und bezog sich vor allem auf das Kriterium Lernerfolg. Detaillierte Kriterien werden in der vorliegenden Veröffentlichung nicht deutlich. Die Untersuchung zielte auf die Messung der Prozessqualität.

Es werden keine Begründungen bezüglich der Wahl des Evaluationsdesigns vorgetragen. Die eingesetzten Methoden werden jedoch überwiegend erläutert. Das Kriterium wurde teilweise erfüllt und somit als neutral gewertet.

Übertragbarkeit

Zur Übertragbarkeit der Ergebnisse werden keine Angaben gemacht. Es sind jedoch ähnliche Resultate unter gleichen Bedingungen zu erwarten (keine spezifische Nutzung der neuen Medien, gleiches Kursmaterial, unterschiedliche Gruppenzusammensetzungen). Das Kriterium wird daher als neutral gewertet.

II.III.I.III Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Nützlichkeit						
1.1	Identifizierung der Beteiligten und Betroffenen		x			
1.2	Klärung der Evaluationszwecke	x				
1.3	Klarheit der Veröffentlichung		x			
2. Genauigkeit						
2.1	Beschreibung des Evaluationsgegenstandes				x	
2.2	Kontextanalyse	x				
2.3	Beschreibung von Zwecken und Vorgehen			x		
2.4	Übertragbarkeit			x		
	Gesamtbewertung	2	2	2	1	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 1: Bewertungsraster P1

II.III.II P 2: Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr. (2001)

II.III.II.I Nützlichkeit

Identifizierung der Beteiligten und Betroffenen

Die Beteiligten und Betroffenen können mehr oder weniger identifiziert werden. Während nachvollziehbar ist, welche Kohorten die Zielgruppen der Untersuchung darstellten, können keine klaren Kurse identifiziert werden. Ebenfalls ist nicht deutlich, wer an der Studie beteiligt war.

Das Kriterium ist teilweise erfüllt und kann somit mit neutral bewertet werden.

Klärung der Evaluationszwecke

Vor dem Hintergrund eines Papiers der „Middle States Commission on Higher Education“ (2001), das die Analyse studentischer Zufriedenheit^{xxviii} als einen wichtigen Indikator für die Effektivität einer höheren Bildungseinrichtung beschreibt und in Bezug darauf, dass die Zahl der Online-Einschreibungen am University of Maryland University College (UMUC) seit dem Jahr 2000 ständig ansteigt, zielt die hier untersuchte Studie von Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr (2001) darauf ab, die Zufriedenheit von Studierenden mit Online- und Face-to-Face-Betreuung zu vergleichen.

Die Zwecke der Evaluation werden eindeutig geklärt, weshalb das Kriterium mit positiv gewertet wird.

Klarheit der Veröffentlichung

Der vorliegende Forschungsbericht ist klar formuliert, folgt einer inhaltlichen Struktur und enthält alle erforderlichen Daten. Die Erhebungsinstrumente werden beschrieben und im Anhang dargestellt. Das Kriterium wird voll erfüllt und kann als positiv gewertet werden.

II.III.II.II Genauigkeit

Beschreibung des Evaluationsgegenstandes

Der Evaluationsgegenstand ist grob beschrieben: Die Untersuchung bezog sich auf die Zufriedenheit der Studierenden mit der akademischen Qualität von Online- und Face-to-Face-Kursen. Als Messinstrument diente

^{xxviii} Im Original: "One tool for studying institutional effectiveness and quality is the university's alumni survey. In the Middle States Commission's Characteristics of Excellence, they list a number of such tools and documents to use in the analysis and documentation of an institution's excellence. "Analysis of student satisfaction survey results" is listed as an assessment tool to help an institution perform outcome assessment and evaluate its overall effectiveness" (zit. nach Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr 2001:43-44).

der Alumni-Survey^{xxix} des Jahres 1999, über den Selbstauskünfte der Studierenden eingeholt werden.

Es ist jedoch nicht nachvollziehbar, welche Kurse und Kursarten in die Untersuchung aufgenommen wurden. Ebenso wenig ist über das Kursmaterial bekannt. So bleibt unklar, ob es didaktische Unterschiede in der Lehre in Präsenz- und Onlineveranstaltungen gegeben hat. Das Kriterium ist als eher negativ zu werten.

Kontextanalyse

Die Untersuchung wurde am University of Maryland University College (UMUC) durchgeführt. Dies ist insofern wichtig zu erwähnen, als die Zahl der Einschreibungen im „undergraduate“-Bereich etwa zur Hälfte in Online-Seminare erfolgte. Die Studierenden konnten sowohl Online- als auch Face-to-Face-Kurse belegen und werden in der vorliegenden Studie als „mixed-format“ beschrieben. Damit bilden sie die spezifische Zielgruppe ab. Im Jahr 2000 wurden diese Studierenden in Bezug auf ihre Zufriedenheit befragt.

Die AutorInnen verweisen auf das Papier „Quality On the Line: Benchmarks for Success in Internet-Based Distance Education“ (Phipps & Merisotis 1999), welches als Bezugsrahmen empfiehlt, die Effektivität und den Lehr-Lernprozess mittels einer Evaluation zu messen, welche sich verschiedener Methoden und spezifischer Standards bedient.

Als eine dieser Methoden wird die Messung der Effektivität und Qualität von Präsenz- und Online-Lernen neben anderen Verfahren auch durch den Alumni-Survey angesehen.

„‘Analysis of student satisfaction survey results’ is listed as an assessment tool to help an institution perform outcome assessment and evaluate its overall effectiveness“ (Middle States Commission on Higher Education“ 2001:43-44, zit. nach Gail Wisan, Shirin Nazma & Charles P. Pscherer Jr 2001:4).

Auf das Vorwissen der ProbandInnen sowie auf pädagogische Rahmenbedingungen wird nicht eingegangen.

Das Kriterium wurde eher erfüllt und wird als eher positiv gewertet.

Beschreibung von Zwecken und Vorgehen

Zur Messung der Zufriedenheit mit der akademischen Qualität von Online- und Face-to-Face-Kursen und der Qualität des Services wurden alle Bachelor-AbsolventInnen des Jahres 1999 befragt.

Es handelte sich hierbei um eine summative Querschnittsmessung, die auf die Outputqualität ausgerichtet ist. Es wurden drei Ebenen von Studierenden untersucht, die nach Anzahl der belegten Online-Kurse eingeteilt wurden.

^{xxix} Dies entspricht der Empfehlung des bereits erwähnten Papiers der „Middle States Commission on Higher Education“ (2001).

Das Vorgehen wird dabei sehr detailliert beschrieben, so dass das Kriterium „Beschreibung von Zwecken und Vorgehen“ mit positiv bewertet werden kann.

Übertragbarkeit

Es wird in der Publikation nicht darüber berichtet, ob sich die Ergebnisse übertragen lassen. Beispiele aus anderen Bereichen werden ebenfalls nicht angegeben. Da jedoch anzunehmen ist, dass eine Replizierung der Befragung mit denselben Methoden ähnliche Ergebnisse liefern würde, ist dieses Kriterium mit neutral zu bewerten.

II.III.II.III Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Nützlichkeit						
1.1	Identifizierung der Beteiligten und Betroffenen			x		
1.2	Klärung der Evaluationszwecke	x				
1.3	Klarheit der Veröffentlichung	x				
2. Genauigkeit						
2.1	Beschreibung des Evaluationsgegenstandes				x	
2.2	Kontextanalyse		x			
2.3	Beschreibung von Zwecken und Vorgehen	x				
2.4	Übertragbarkeit			x		
	Gesamtbewertung	3	1	2	1	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 2 Bewertungsraster P2

II.III.III P 3: Rob Martens, Stanley J. Portier & Martin Valcke (1995)

II.III.III.I Nützlichkeit

Identifizierung der Beteiligten und Betroffenen

Die Studie wurde in Kooperation zwischen der Open University der Niederlande und der Universität Gent in Belgien durchgeführt. Dabei wirkten 502 Studierende mit, die einen sozialwissenschaftlichen Studiengang im ersten Semester belegten. Davon waren 342 ProbandInnen in der Kontrollgruppe, der Rest ist in 5 Untersuchungsgruppen aufgeteilt, die an späterer Stelle näher beschrieben werden.

Die Beteiligten und Betroffenen konnten ausreichend identifiziert werden. Das Kriterium ist erfüllt und wird als positiv gewertet.

Klärung der Evaluationszwecke

Die Hauptfragestellung der Untersuchung zielte darauf ab, ob unabhängiges Lernen in einer interaktiven Lernumgebung eine Alternative zu traditionellem Face-to-Face Lernen darstellte. Dies sollte vor allem vor dem Hintergrund einer kostengünstigen Ausbildung untersucht werden^{xxx}. Dazu wurden Prototypen interaktiver Lernumgebungen mit Printmaterialien verglichen:

“To measure the effectivity of the ILE, three prototype-versions are used in a research set-up. Moreover, two alternative learning environments based on printed learning materials (PLE, printed learning environments) and a control setting with lectures are incorporated into the research design. The learning environments are researched with university students taking a statistics course. This course focused on ‘inferential statistics’ which is a part of an Ou-course for the social sciences” (Rob Martens, Stanley J. Portier & Martin Valcke 1995:8).

Eine Schlüsselrolle in der Untersuchung spielten die sogenannten „embedded support devices (ESD)“, eine Art didaktische Zugabe zum Basiskursmaterial mit Lösungshinweisen, Glossaren, Selbsttests, Beispielen usw. Die ESDs waren sowohl in den online vermittelten als auch in den Printkursen vorhanden, allerdings entschied in beiden Fällen der Autor bzw. die Autorin der jeweiligen Kurse über ihre Verwendung.

Da die Evaluationszwecke klar beschrieben werden, ist das Kriterium voll erfüllt und kann positiv bewertet werden.

^{xxx} Dazu muss gesagt werden, dass die OpenUniversity ebenfalls eine FernUniversität ist, deren Lehrvermittlung traditionell über Printkurse erfolgt. Dies ist zumindest zur Zeit der Veröffentlichung der Fall, inzwischen hat sich die OU vor allem in Bezug auf die Entwicklung und Umsetzung interaktiver Lernangebote einen Namen gemacht.

Klarheit der Veröffentlichung

Auch in dieser Primärstudie ist der vorliegende Forschungsbericht klar formuliert, enthält alle erforderlichen Daten und folgt einer inhaltlich logischen Struktur. Die Erhebungsinstrumente werden beschrieben, jedoch nicht im Detail dargestellt. Das Kriterium wird erfüllt und kann als eher positiv gewertet werden.

II.III.III.II Genauigkeit

Beschreibung des Evaluationsgegenstandes

Von der Lernumgebung existierten verschiedene Versionen: in der ersten Version waren alle Möglichkeiten der Interaktivität geblockt, so dass Studierende die Umgebung als eine Art elektronisches Lernbuch ohne weitere Funktionen verwenden konnten. In zwei weiteren Versionen wurden diese angeboten, die Umgebungen unterschieden sich hier nur in der Art und Weise der angebotenen ESD:

“Considering the independent learning setting, the high-quality elaboration of the learning materials is of prime importance to support a learning process:

- We distinguish in the learning materials on the one hand `basic content that reflects the domain specific (scientific) information and on the other hand embedded support devices (ESD). Given the large amount of ESD in learning materials, these are clustered according to their basic functions and effects as suggested by Valcke, Martens, Poelmans and Daal (1993):
 - *orienting ESD*: learning objectives, references to other learning materials, references to required prior knowledge, and use of history.
 - *processing ESD*: additional learning materials, advance organizers, figures, glossary, introductions, study advices, summaries, tables, examples, extended learning materials.
 - *Testing ESD*: self-test items, exercise items (on know, insight and apply-level, answers” (ebd., Hervorhebungen im Original).

Je nach gestesteter Version, waren die ESD entweder mehr oder weniger sichtbar.

Das Printmaterial entsprach dem Standard von Lehrbriefen. Es handelte sich hierbei um didaktische aufbereitete Kurse, die von den Studierenden eigenständig bearbeitet wurden. Zu Forschungszwecken wurden zwei verschiedene Versionen von Printmaterial eingesetzt – einmal mit und einmal ohne Hilfestellungen bzw. didaktischen Anweisungen. Als Kontrollgruppe diente der Untersuchung eine traditionelle Face-to-Face-Vorlesung der Inferenzstatistik.

Der Evaluationsgegenstand wird im Grunde genommen ausreichend beschrieben, jedoch fehlen detailliertere Angaben zu den eingesetzten Instrumenten, weshalb das Kriterium wird nur als eher positiv gewertet werden kann.

Kontextanalyse

Bei der Beschreibung des theoretischen Hintergrundes werden verschiedene Variablen herausgestellt, die Einfluss auf das unabhängige Lernen haben und die somit in der Studie untersucht werden. Bei der Beschreibung von unabhängigem Lernen beziehen sich die AutorInnen unter anderem auf das Instructional Design Modell von Richard E. Snow & Judy Swanson (1992).

"These authors define the following interactions (represented by '**') between variables and processes:

*aptitudes * learning types * content domain * instructional design
* situations, populations*

Applying this matrix to our research we get the following:

<i>aptitude:</i>	cognitive and non-cognitive (motivation) aptitudes
<i>learning type:</i>	independent learning in interactive or printed learning environments
<i>content domain:</i>	statistics in social sciences
<i>instructional design:</i>	learning materials where ESD are separated from the basic content
<i>situation:</i>	independent learning
<i>population:</i>	adult students" (a.a.O.:9, Hervorhebungen im Original).

Darüber hinaus werden Variablen beschrieben, die Einfluss auf das Lernen ausüben. Dazu gehören kognitive Fähigkeiten (cognitive aptitudes), Lesekompetenzen (cognitive aptitudes related to reading) und Motivation als nicht-kognitive Dimension.

Da sich die Studierenden im ersten Studiensemester befanden, wurde das Vorwissen als sehr gering eingeschätzt, was in Bezug auf die Untersuchung hinsichtlich der Vergleichbarkeit vorteilhaft ist. Es wurde dennoch ein Pretest absolviert, der das Vorwissen der Studierenden geprüft hat und dieses als eine Einflussvariable behandelt werden kann.

Über die technischen Voraussetzungen werden keine Informationen gegeben. Die Navigation innerhalb der Lernumgebung scheint nachvollziehbar zu sein, ebenfalls die Benutzung der ESDs. Da eine Logfile-Analyse durchgeführt wurde, ist davon auszugehen, dass die Studierenden nicht zu Hause, sondern immer in von der Hochschule betreuten Computerräumen gearbeitet haben.

Durch die theoretische Fundierung wurde der Rahmen der Evaluation gut abgesteckt. Das Kriterium wurde erfüllt und wird mit positiv bewertet.

Beschreibung von Zwecken und Vorgehen

Die Beschreibungen von Zwecken und Vorgehen geschieht auf sehr detaillierte Weise.

Die Methoden der Untersuchung werden begründet. Auf das Evaluationsdesign wird nicht spezifisch eingegangen, jedoch wird dies aus dem Kontext klar. Das Kriterium wurde voll erfüllt und wird mit positiv bewertet.

Übertragbarkeit

Es wird in der Publikation nicht darüber berichtet, ob sich die Ergebnisse übertragen ließen. Beispiele aus anderen Bereichen werden nicht angegeben. Da jedoch die untersuchten Lernumgebungen sehr speziell sind, ist eine Übertragung vermutlich ausgeschlossen. Das Kriterium ist als neutral zu werten.

II.III.III.III Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Nützlichkeit						
1.1	Identifizierung der Beteiligten und Betroffenen	x				
1.2	Klärung der Evaluationszwecke	x				
1.3	Klarheit der Veröffentlichung		x			
2. Genauigkeit						
2.1	Beschreibung des Evaluationsgegenstandes		x			
2.2	Kontextanalyse	x				
2.3	Beschreibung von Zwecken und Vorgehen	x				
2.4	Übertragbarkeit			x		
	Gesamtbewertung	4	2	1	0	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 3: Bewertungsraster P3

II.III.IV P 4: Josef Smolle, Reinhard Staber, Heide Neges, Gilbert Reibnegger & Helmut Kerl (2005)

II.III.IV.I Nützlichkeit

Identifizierung der Beteiligten und Betroffenen

An der Untersuchung waren 26 Studierende beteiligt, die im Wintersemester 2003/2004 an einem speziellen Seminar im Studienmodul zur Dermatoonkologie teilnahmen. Dieses hatte die Dauer von 5 Wochen und umfasste insgesamt 30 Unterrichtsstunden.

Die Beteiligten und Betroffenen können daher gut identifiziert werden. Das Kriterium ist mit positiv zu werten, da es voll erfüllt wurde.

Klärung der Evaluationszwecke

Die Forschungsfrage lautete, ob der Erfolg beim Lernen mit Computerlehrprogrammen in Form von Computer Based Training geringer oder höher ist als der Lernerfolg, der mit „innovativen didaktischen Konzepten im Präsenzunterricht“ (Josef Smolle, Reinhard Staber, Heide Neges, Gilbert Reibnegger & Helmut Kerl 2005:883) erzielt wurde. Die AutorInnen beziehen sich bei dieser Fragestellung auf den wachsenden Einfluss neuer Medien im Bildungsbereich, speziell in der medizinischen Ausbildung.

Die Evaluationsziele werden deutlich dargestellt. Das Kriterium ist mit positiv zu bewerten.

Klarheit der Veröffentlichung

Die Veröffentlichung folgt einer klaren Struktur und ist auch für Fachfremde gut verständlich. Vereinzelt hätten die Forschungsentscheidungen detaillierter begründet werden können.

Das Kriterium wurde überwiegend erfüllt und ist mit eher positiv zu bewerten.

II.III.IV.II Genauigkeit

Beschreibung des Evaluationsgegenstandes

Der Kurs im Fach Dermatoonkologie umfasste insgesamt 30 Stunden. Wie bereits erwähnt, wurden 24 davon in Präsenzform und 6 virtuell abgehalten.

Bei den Lernunterlagen im Online-Kurs handelte es sich um Lernprogramme zu klinischen Fragestellungen, die Bildmaterial und eine differenzierte Rückmeldefunktion enthielten.

Dieser Kursteil wurde dementsprechend in Form von Computer-Based-Training abgehalten, der über das Internet zugänglich war.

Im Präsenzkurs wurde Team-Techning als didaktisches Modell angewendet. Dabei wurden klinische Themen kurz mit PowerPoint präsentiert und anschließend in Kleingruppen diskutiert und vorgestellt. Weitere Informationsvermittlung fand durch Mini-Lectures statt.

„Jeder Frame [bestand] aus einer Vignette und 5 Auswahloptionen. Auf jede Auswahloption, ob richtig oder falsch, gab es eine differenzierte elaborierte Rückmeldung. Bei falschen Antworten verzweigte das Programm auf den vorigen Frame zurück, bei richtigen Antworten wurde zum nächsten Frame weitergegangen [...]. Die 8 verwendeten CBT-Lernobjekte [...] umfassten zwischen 8 und 35 Frames“ (ebd.).

Bei der Bearbeitung der Fragestellungen wurde jeder Mausklick registriert und so die Trefferquote bzw. der Anteil an falschen Klicks aufgezeichnet.

„Zur Ermittlung der Trefferquote wurde jeder Auswahlvorgang, d. h. jeder Mausklick, registriert, und der Anteil an „falschen“ Klicks im Verhältnis zur Gesamtzahl der Klicks aufgezeichnet. Wenn z. B. jemand jede Frage auf Anhieb richtig beantworten konnte, ergab sich eine Trefferquote von 100%. Wenn dagegen jemand z. B. bei jedem Frame einen Fehlversuch aufwies, bevor er die richtige Antwort fand, dann ergab sich eine Trefferquote von 50 %“ (ebd.).

Für die Absolvierung der Aufgaben wurde im Vorfeld ein Termin festgesetzt. Zum Bestehen dieses Kursteils war eine Trefferquote von 90% erforderlich, was u.U. ein mehrfaches Durchlaufen einzelner Aufgaben erforderte.

Der Evaluationsgegenstand wird detailliert beschrieben. Das Kriterium wird mit positiv bewertet.

Kontextanalyse

Die Einbindung der Lernmaterialien im Online-Kurs erfolgte über den „Virtuellen Medizinischen Campus“ der Universität Graz.

„Der VMC Graz stellt ein Lern- und Informationssystem dar, das das gesamte Diplomstudium Humanmedizin inhaltlich abbildet und von den Studierenden mit großem Interesse genutzt wird. Von den Studierenden der Humanmedizin in Graz werden pro Monat bis zu 40000 Zugriffe auf Lernobjekte im VMC Graz getätigt“ (ebd.).“

Im Semester der Untersuchung waren etwa 300 Studierende in den Studiengang eingeschrieben. Das hier beschriebene Studienmodul konnte als eines unter 6 möglichen Modulen gewählt werden – die Zulassung war jedoch auf 26 Studierende beschränkt.

Der Kontext wird überwiegend erläutert. Auf das Vorwissen der ProbandInnen sowie auf pädagogische Rahmenbedingungen wird nicht eingegangen, weshalb das Kriterium nur mit eher positiv gewertet wird.

Beschreibung von Zwecken und Vorgehen

Der Studie liegt eine summative Evaluation zu Grunde. Obwohl zwei Messpunkte benannt werden, handelt es sich hierbei dennoch um eine Querschnittsevaluation, da zu jedem Messzeitpunkt nur ein Setting untersucht wurde und die Evaluation nicht auf den Verlauf ausgerichtet war.

Das Vorgehen ist verständlich beschrieben, die Untersuchungsmethoden werden dargestellt. Allerdings werden keine Begründungen für die Methodenauswahl geliefert. Das Evaluationsdesign wird nicht explizit genannt, wird aber aus dem Kontext deutlich. Das Kriterium ist überwiegend erfüllt und wird mit positiv bewertet.

Übertragbarkeit

In der Publikation wird auf den Aspekt der Übertragbarkeit nur insofern eingegangen, als die Überlegung angestellt wird, ob sich das CBT auch unter ökonomischen Gesichtspunkten als Alternative zum Präsenzlernen eignete. Dazu müssten ähnliche Ergebnisse in größeren Studien reproduziert werden:

„Beachtet man den verminderten Bedarf an Lehrbetreuung, der sich bei einer Anwendung an einer großen Zahl von Studierenden ergäbe, so könnte CBT - vorausgesetzt, dass sich ähnliche Ergebnisse in größeren Untersuchungen bestätigen lassen - als durchaus ökonomische Alternative erscheinen“ (a.a.O.:888).

Die AutorInnen weisen also selbst darauf hin, dass der Rahmen der Untersuchung zu klein war, um Aussagen zur Übertragbarkeit treffen zu können. Das Kriterium wurde kaum erfüllt und ist mit eher negativ zu werten.

II.III.IV.III Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Nützlichkeit						
1.1	Identifizierung der Beteiligten und Betroffenen	x				
1.2	Klärung der Evaluationszwecke	x				
1.3	Klarheit der Veröffentlichung		x			
2 Genauigkeit						
2.1	Beschreibung des Evaluationsgegenstandes	x				
2.2	Kontextanalyse		x			
2.3	Beschreibung von Zwecken und Vorgehen	x				
2.4	Übertragbarkeit				x	
	Gesamtbewertung	4	2	0	1	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 4: Bewertungsraster P4

II.III.V P 5: Randall S. Davies & Robert Mendenhall (1998)

II.III.V.I Nützlichkeit

Identifizierung der Beteiligten und Betroffenen

Zum Zeitpunkt der Untersuchung (1998) waren etwa 1000 Studierende in den Kurs HEPE 129 eingeschrieben. Abhängig von der Kapazität wurde dieser Kurs in 5 bis 6 Sektionen unterteilt, deren Klassengröße etwa jeweils 200 Studierende betrug.

Die Betroffenen konnten ausreichend identifiziert werden. Das Kriterium ist damit voll erfüllt und wird mit positiv bewertet.

Klärung der Evaluationszwecke

In der Evaluation wurden zwei Instruktionsmethoden miteinander verglichen: Onlinelernen vs. Lernen im Klassenraum. Wie bereits erwähnt, wurde dazu der Kurs HEPE 129 analysiert, der grundsätzlich in Präsenzform abgehalten wurde. Im Rahmen der Untersuchung wurden jedoch einzelne Kursteile in einer Online-Version über das WWW angeboten.

Ziel der Studie war es, Vor- und Nachteile beider Lehrformen auszumachen und den Grad der Erreichung der Kursziele darzustellen.

Es wird nicht explizit erwähnt, wer die Studie initiiert hatte, jedoch ist aus der folgenden Äußerung zu schließen, dass auch die Universität selbst ein Interesse an den Ergebnissen hatte:

„This is important for the department of Health and Human Performance because they hope to find that students taking the online course are being provided with a satisfactory learning experience“ (Randall S. Davies & Robert Mendenhall 1998:1).

Die Evaluationszwecke wurden detailliert dargestellt. Das Kriterium wird mit positiv bewertet, da es voll erfüllt wurde.

Klarheit der Veröffentlichung

Die Veröffentlichung folgt einer sinnvollen Struktur und ist mehrheitlich in verständlicher Form verfasst. Lediglich bei der Beschreibung der Zusammensetzungen der Testgruppen traten kleinere

Unklarheiten auf^{xxxi}. Das Kriterium wurde überwiegend erfüllt und wird mit eher positiv bewertet.

II.III.V.II Genauigkeit

Beschreibung des Evaluationsgegenstandes

Der Kurs „HEPE 129 - Fitness and Lifestyle Management“ wurde im Studiengang „Health Education/ Physical Education“ an der Brigham Young University durchgeführt. Ziel des Kurses war das Erlernen von Grundlagen zu Gesundheit und Fitness, die Studierende befähigen sollten, dieses erworbene Wissen beim Erstellen von personalisierten Fitness- und „lifestyle management“-Programmen anzuwenden.

Während der ursprüngliche Kurs im Klassenraum auf dem Campus unterrichtet wurde, war der Online-Kurs, der vom „Independent Studies Departement“ entwickelt wurde, über das WWW verfügbar. Beide Kurse bezogen sich jedoch auf die gleichen Inhalte.

„Both courses cover the same basic objectives in thirteen independent modules or lessons; both require students to complete similar additional learning activities; yet the methods for providing the instruction are quite different“ (a.a.O.:6.).

Darüber hinaus wurden die Studierenden aufgefordert, an einem Fitness-Check teilzunehmen, bei dem ein Bluttest und ein Prä-/Post-Fitness-Test zu Beginn und zum Ende des Semesters durchgeführt wurde. Während des gesamten Semesters fand ein verpflichtendes Fitness-Programm statt:

„Students participating in the classroom experience must sign up for an appointment to have the fitness appraisal completed. Fitness appraisals are completed outside of regular class time and are administered by trained staff. This appraisal requires that students give a blood sample for analysis. Students record the results of the fitness appraisal on a fitness questionnaire. The questionnaire is submitted and processed. Students receive the results approximately one week later. With the online experience, students complete the fitness appraisal on their own following instructions provided. They are not required to give a blood sample. The fitness results are recorded on a fitness questionnaire and processed immediately by the computer“ (a.a.O.:7).

Bei dem Kurs, der im Klassenraum unterrichtet wurde, handelte es sich um einen traditionellen Kurs, der in 18 Vorlesungsmodulen aufgeteilt war. Die Studierenden wurden in 13 dieser Module getestet und mussten darüber hinaus 3 von 5 Wahlpflichtmodulen (Marriage and Family Relations, Women's Health, Environmental Health, Maintaining a Healthy Weight, and Crisis Intervention) belegen, die jedoch nicht prüfungsrelevant waren.

^{xxxi} So wurde beispielsweise berichtet, dass die Studierenden, die sich freiwillig für die Untersuchung zur Verfügung stellten, 4 oder 5 Online-Stunden absolvieren sollten. Es wurde daraus zunächst nicht klar, ob es sich dabei um den vollen Kursumfang, oder nur um „Probestunden“ handelte. Im Rahmen der weiteren Beschreibungen konnte dies jedoch geklärt werden.

Das Material des Online-Kurses wurde von verschiedenen Lehrenden des HEPE 129-Kurses erstellt und diente vor allem der Reflexion. Es gab 13 Lerneinheiten, die zur Absolvierung des Kurses belegt werden mussten. Das gesamte Kursmaterial war online verfügbar und bestand aus Videoclips, Animationen, Sound, Grafiken und weiterführenden Links. Weiterhin gab es ein „did you know“-Feature, das verwandte Fakten präsentierte. Darüber hinaus bekamen die Studierenden vier CD-ROMs, die Videoclips und Animationen enthielten. Durch ein Bulletin-Board bestand die Möglichkeit zum Kontakt mit anderen Lernenden - dies war jedoch freiwillig. Im Online-Kurs wurde nach jeder Lektion ein Test durchgeführt, auf den direkte Rückmeldungen gegeben wurden. Ebenfalls war Feedback in Bezug auf den Lernprozess und die Fitnessprogramme vorgesehen. Im Online-Kurs gab es keine Wahlpflichtmodule.

Der Evaluationsgegenstand wurde sehr detailliert beschrieben. Das Kriterium wird positiv bewertet.

Kontextanalyse

Der Forschungsbericht setzt sich ausführlich mit den Theorien und dem Stand der Forschung in Bezug auf Lernen und Instruktion auseinander. Es werden diverse Referenzen angegeben. Darüber hinaus gehen die Autoren auf den Aspekt ein, dass Studien, die Online- und Präsenzinstruktion miteinander vergleichen, oftmals zu dem Schluss gelangen, dass sich Lernergebnisse durch Computernutzung verbessern. Dies sei aber keine Garantie für gute Instruktion. Aus diesem Grund werden in der vorliegenden Arbeit „Elements of good Instruction“ herausgearbeitet.

Das Vorwissen der Studierenden oder Vorraussetzungen an der Universität werden dagegen fast nicht beschrieben. Das Kriterium ist daher nur überwiegend erfüllt und wird mit eher positiv bewertet.

Beschreibung von Zwecken und Vorgehen

Die Evaluation wurde als Querschnittserhebung durchgeführt und hatte formativen Charakter.

Die Methoden der Untersuchung werden ausführlich dargestellt, jedoch nicht in allen Fällen näher begründet. Das Kriterium ist damit überwiegend erfüllt und wird mit eher positiv bewertet.

Übertragbarkeit

Es wird in der Publikation nicht darüber berichtet, ob sich die Ergebnisse übertragen ließen. Beispiele aus anderen Bereichen werden nicht angegeben. Da jedoch die untersuchten Lernumgebungen sehr speziell sind, ist eine Übertragung vermutlich ausgeschlossen. Das Kriterium ist als negativ zu werten.

II.III.V.III Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Nützlichkeit						
1.1	Identifizierung der Beteiligten und Betroffenen	x				
1.2	Klärung der Evaluationszwecke	x				
1.3	Klarheit der Veröffentlichung		x			
2 Genauigkeit						
2.1	Beschreibung des Evaluationsgegenstandes	x				
2.2	Kontextanalyse		x			
2.3	Beschreibung von Zwecken und Vorgehen		x			
2.4	Übertragbarkeit					x
	Gesamtbewertung	3	3	0	0	1
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 5: Bewertungsraster P5

II.III.VI P 6: Vassilis Hartzoulakis (2002)

II.III.VI.I Nützlichkeit

Identifizierung der Beteiligten und Betroffenen

An der Untersuchung wirkten 24 SchülerInnen im Alter von 15-17 Jahren des zweiten und dritten Jahrgangs des Lykeion in Peresteri mit. Zu diesem Zeitpunkt hatten diese bereits mindestens 6 Jahre Englischunterricht absolviert.

Weiterführende Angaben zu Beteiligten und Betroffenen wurden nicht gemacht. Ebensovienig wird ersichtlich, in welchem Zeitraum die Untersuchung durchgeführt wurde.

Das Kriterium wurde teilweise erfüllt und wird mit neutral gewertet.

Klärung der Evaluationszwecke

Die Studie untersuchte die Effektivität von computerunterstütztem Sprachenlernen im Vergleich zu traditionellem Lernen im Klassenraum. Die Forschungsfragen lauteten: Ist computervermittelte Instruktion mehr, weniger oder gleich effektiv, wie traditionelles Lehren? Ist computervermittelte Instruktion für Mädchen und Jungen gleichermaßen effektiv? Um diese Fragen zu klären, wurden die Lernergebnisse der verschiedenen Untersuchungsgruppen miteinander verglichen.

„The study compares the learning outcomes achieved through computer-assisted instruction with those obtained in a traditional classroom when teaching one of the formS - in this case, inversion to two groups of randomly selected students in the second and third grade of Lykeio“ (Vassilis Hartzoulakis 2002:7).

Zum Hintergrund ist zu sagen, sich der Autor vor allem auf das Lehren von „formS“ bezieht, was bedeutet, dass isolierte linguistische Merkmale und formale Sprachelemente aus dem Kontext extrahiert werden. Detailliertere Angaben zu „form“ und „formS“ sind bei der Beschreibung der Kontextanalyse zu finden.

Die Evaluationsziele werden deutlich dargestellt. Das Kriterium ist mit positiv zu bewerten.

Klarheit der Veröffentlichung

Positiv ist hervorzuheben, dass der Autor zunächst auf die relevanten Definitionen von Begriffen eingeht, die in der Arbeit benutzt werden.

„Lykeio: Greek Secondary Education Institution addressing adolescents from 15 to 17 years of age.

Traditional Class: Conventional term for the class where no computers are used in the process of the lesson, the visual stimulus

is the teacher and the blackboard, and the interaction patterns are between the teacher and the students, and between the students themselves.

Form: Overall engagement in meaning, encompassing elements of language studied in context.

FormS: Isolated linguistic features, formal elements of language extracted from context" (a.a.O.:8).

Die Veröffentlichung ist einer logischen Struktur entsprechend gegliedert. Allerdings ist sie nicht durchgängig verständlich. So hätte die Unterscheidung von „form“ und „formS“ deutlicher hervorgearbeitet bzw. nicht ausschließlich im Kontext anderer Kapitel bearbeitet werden können.

Das Kriterium wird überwiegend erfüllt und ist mit „eher positiv“ zu bewerten.

II.III.VI.II Genauigkeit

Beschreibung des Evaluationsgegenstandes

Die Software, die zur Präsentation der Unterrichtsstunden eingesetzt wurde, wurde vom Autor selbst entwickelt. Es handelte sich um eine Serie von Webseiten in html, die im Internet zur Verfügung standen. Das Thema der Lernreihe war „inversion“.

„The webpages presented the form ‘inversion’ in an interactive way, where the students would be able to explore -employing ‘point and click’ techniques- the various cases where inversion can be used“ (a.a.O.:28).

Auf den Webseiten wurden 6 Fälle von Inversion beispielhaft beschrieben und mit Macromedia Flash umgesetzt. Es folgten Beispiele und Übungen mit „Hot potatoe“^{xxiii}-tests. Die Webseiten waren jedoch nicht im engeren Sinn interaktiv, da sie keine Kooperation unter den Lernenden ermöglichten.

Im Präsenzunterricht wurden zwei verschiedene Stunden von derselben Lehrperson (dem Autor) zum Thema „inversion“ abgehalten. Die Beispielsätze und Übungen, die im Unterricht verwendet wurden, entsprachen denen, die auch am Computer präsentiert wurden. Das Material wurde an der Tafel im Stil einer Vorlesung präsentiert. Fragen von SchülerInnen wurden erst nach einer Zusammenfassung der gesamten Stunde beantwortet.

Der Evaluationsgegenstand wird mehr oder weniger präzise beschrieben. Die für die Experimentalgruppe erstellten Webseiten sind in der Arbeit enthalten. Das Kriterium wird hinreichend erfüllt und wird mit „eher positiv“ bewertet.

^{xxiii} Software zum Erstellen von Quizzes, Multiple-Choice-tests u.ä.

Kontextanalyse

Der Lehrplan des griechischen Ministeriums für Erziehung (Hellenic Ministry of Education) spricht sich für den Einsatz von computergestützten Lehrmethoden im Englischunterricht aus, jedoch ohne anzugeben, auf welchem Weg dies geschehen soll.

„What is specified is a detailed list of forms, functions, and structures that are to be taught throughout the three year span of the Hellenic Secondary Education. The teachers are, in that way, guided towards covering the material set -the specified **forms** and functions - in the best way they see fit for their students“ (a.a.O.:5, Hervorhebung AP).

Da dem Begriff „form“ verschiedene Definitionen zu Grunde liegen, wird in der vorliegenden Arbeit zwischen „form“ und „formS“ unterschieden (ebd.).

„Doughty and Williams (1998:4) clarified the issue by differentiating the term form from the term formS. They adopted Long's (1988) suggestions who defined isolated linguistic features (e.g. grammar) as formS and the overall engagement in meaning as form. Doughty and Williams (1998) stressed that the two terms are not polar opposites; instead, they are interwoven in the sense that form entails isolation or extraction of linguistic features from context, whereas formS is limited to isolated features. This approach has been adopted by the author throughout this study“ (ebd.).

Für das Kriterium der Kontextanalyse ist darüber hinaus bedeutsam, dass der Autor eine detaillierte Beschreibung der relevanten Literatur vornimmt und auf den technischen Hintergrund von computerunterstütztem Sprachenlernen eingeht. Darüber hinaus weist er auf bereits bestehende Vergleichsuntersuchungen in diesem Bereich hin.

Letztlich wird noch auf die besondere Situation am Lykeion Bezug genommen: Bevor die SchülerInnen das Lykeio besuchen, erhalten sie zuvor in der Primar- und Sekundarerbziehung insgesamt sechs Jahre Englischunterricht. Eine große Anzahl derer erhält weiteren Englischunterricht an privaten Institutionen. Dies führt zu einem Ungleichgewicht im Level der englischen Sprachkenntnisse der SchülerInnen am Lykeion. Daher musste vor der Ziehung der Stichprobe der Kenntnisstand im Einzelnen festgestellt werden.

Das Kriterium wird voll erfüllt, da pädagogische und technische Voraussetzungen der Evaluation beschrieben und Rahmenbedingungen deutlich gemacht werden. Darüber hinaus wird deutlich, welchen theoretischen Überlegungen die Evaluation folgt. Das Kriterium wird mit positiv bewertet.

Beschreibung von Zwecken und Vorgehen

Die Studie basierte auf einem Prätest-Posttest-Kontrollgruppen-Design, das in einer Querschnittserhebung angelegt wurde.

„This study is based upon a pretest-posttest randomized experimental design (Trochim, 2002) with the objective of comparing CALL's effectiveness in teaching forms with the effectiveness of traditional teaching, by analyzing the outcomes of a computer-based instruction with those of a traditional class" (a.a. O:6).

Wie bereits erwähnt, wurden zwei Untersuchungsgruppen gebildet – die Experimentalgruppe erhielt das Unterrichtsmaterial online, die Kontrollgruppe während des Präsenzunterrichts. Zuvor wurde in einem Prätest der Leistungsstand der Lernenden erhoben.

Zwei Tage nach dem Unterricht wurde ein Posttest durchgeführt. Dies wird damit begründet, dass die Evaluation der Effektivität nicht durch Retentionsfaktoren beeinflusst werden sollte. Der Posttest entsprach im Wesentlichen dem Prätest, mit dem Unterschied, dass alle „unnötigen“ Items – also Items, die sich nicht direkt auf das Thema „inversion“ bezogen – entfernt wurden.

Die Methoden der Untersuchung werden dargestellt und begründet. Das Kriterium wird mit positiv bewertet.

Übertragbarkeit

In der Studie wurde versucht, von einer geringen Stichprobe auf die Grundgesamtheit zu schließen. Insofern erhebt diese durchaus den Anspruch auf Übertragbarkeit. Allerdings sind damit zahlreiche Limitationen verbunden:

„Even though the differences could be attributed to the mentality of the students in a Greek State Lykeio, in order to come to more definitive conclusions, a study comparing CALL's learning outcomes when teaching various forms together with an investigation of the retention factor is required" (a.a.O.:i).

Es wird in der Publikation nicht darüber berichtet, ob sich die Ergebnisse übertragen lassen. Beispiele aus anderen Bereichen werden nicht angegeben. Das Kriterium wird als neutral gewertet.

II.III.VI.III Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Nützlichkeit						
1.1	Identifizierung der Beteiligten und Betroffenen			x		
1.2	Klärung der Evaluationszwecke	x				
1.3	Klarheit der Veröffentlichung		x			
2. Genauigkeit						
2.1	Beschreibung des Evaluationsgegenstandes		x			
2.2	Kontextanalyse	x				
2.3	Beschreibung von Zwecken und Vorgehen	x				
2.4	Übertragbarkeit			x		
	Gesamtbewertung	3	2	2	0	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 6: Bewertungsraster P6

II.III.VII P 7: William J. Ryan (2002)

II.III.VII.I Nützlichkeit

Identifizierung der Beteiligten und Betroffenen

Die Untersuchung wurde in einem Kurs zur Einführung in die Statistik während des Schulhalbjahres 1999-2000 am Lakeland Community College in Ohio, USA, durchgeführt. Beteiligt wurden dabei nur die Studierenden, die den Vortest und das Abschlussexamen absolviert hatten. Dabei handelte es sich um eine Kohorte von 78 Studierenden, von denen 31 unter und 47 über 22 Jahre alt waren.

„Of the students who completed the two tests 31, or 40%, were in the traditional^{xxxiii} age bracket (under the age of 22), and 47, or 60%, were in the non-traditional age bracket (22 and older). This age factor was different than other distance learning demographic research that had indicated that "the independent study population has shifted towards younger students" (Wallace, 1996, p. 1)" (William J. Ryan 2002:7).

Die Studierenden waren zu 59% weiblich und zu 41% männlich.

Die Beteiligten und Betroffenen konnten klar identifiziert werden. Das Kriterium ist mit positiv zu bewerten.

Klärung der Evaluationszwecke

In der Pilotevaluation wurde ein einzelner Kurs untersucht, der auf drei verschiedene Arten gelehrt wurde: als traditioneller Vorlesungskurs, als Telekurs und als webbasierter Kurs.

„This study examined the effectiveness of using a telecourse and a web-based course to teach an Introduction to Statistics course at Lakeland Community College with a classroom equivalent“ (ebd.).

Zusammengefasst ging es um die Effektivität von webbasierten Kursen und Telekursen. Dies ist insbesondere vor dem Hintergrund wichtig, dass US-Bildungsinstitutionen zum Zeitpunkt der Untersuchung gerade begannen, die Einführung dieser Kursarten zu diskutieren.

Ziel und Zweck der Untersuchung werden überwiegend nachvollziehbar dargestellt. Das Kriterium ist mit eher positiv zu bewerten.

Klarheit der Veröffentlichung

Die vorliegende Publikation ist klar strukturiert und überwiegend eindeutig formuliert. Die zum Verständnis erforderlichen Daten werden größtenteils geliefert. Die Erhebungsinstrumente werden beschrieben, jedoch nicht im Detail dargestellt. Das Kriterium wird somit nur teilweise erfüllt. Vor dem Hintergrund, dass an einen Tagungsbeitrag andere

^{xxxiii} „Traditionell“ bezieht sich in diesem Fall auf das Durchschnittsalter der Studierenden in Fernlehrkursen, nicht auf die Lehrmethoden.

Ansprüche gestellt werden müssen als an einen Forschungsbericht, wird das Kriterium unter Berücksichtigung dessen als eher positiv gewertet.

II.III.VII.II Genauigkeit

Beschreibung des Evaluationsgegenstandes

Der Evaluationsgegenstand selbst wird nicht näher erläutert. Es wird nur deutlich, dass es sich um einen Kurs zur Einführung in die Statistik handelte, der als traditioneller Vorlesungskurs, als Telekurs und als webbasierter Kurs angeboten wurde. Über das eingesetzte Material liegen keine Angaben vor, ebenso wird die Auswahl des Kurses nicht begründet. Das Kriterium wird nicht erfüllt und müsste damit eine negative Bewertung erhalten. Da jedoch berücksichtigt werden muss, dass es sich bei der Veröffentlichung jedoch um einen Tagungsbeitrag und nicht um einen streng wissenschaftlichen Text handelt, wird das Kriterium nur mit eher negativ bewertet.

Kontextanalyse

Der Kontext wird sehr detailliert beschrieben. Insbesondere der Stand der Forschung in Bezug auf Fernlehre und E-Learning wird genau dargestellt. Darüber hinaus werden die Bezüge der vorliegenden Studie zur durchführenden Institution deutlich gemacht.

„This research is taking place at a community college that has met the needs of the traditional as well as non-traditional student using classical, classroom-based, teaching methods. As the older student cohort begins to impact student enrollment [...], these institutions are facing challenges in meeting the needs of these students in terms of facilities and faculty resources“ (a.a.O.:13).

Auf das Vorwissen der ProbandInnen sowie auf pädagogische Rahmenbedingungen wird nicht eingegangen.

Das Kriterium wird überwiegend erfüllt und wird mit eher positiv bewertet.

Beschreibung von Zwecken und Vorgehen

Die Evaluation verlief summativ^{xxxiv} und bezog sich vor allem auf das Kriterium Lernerfolg. Detaillierte Kriterien werden in der vorliegenden Veröffentlichung nicht deutlich. Die Untersuchung zielte auf die Messung der Outputqualität.

Es werden keine Begründungen bezüglich der Wahl des Evaluationsdesigns vorgetragen. Die eingesetzten Methoden werden teilweise erläutert. Das Kriterium wurde kaum erfüllt und somit als eher negativ gewertet.

^{xxxiv} Nach Durchführung der Pilotstudie wurden weitere Evaluationen durchgeführt. Da diese jedoch nicht im Detail dargestellt werden, ist die hier vorgestellte Untersuchung als summative Evaluation anzusehen. Die vorgestellte Auswertungsergebnisse beziehen sich jeweils auch vor auf die Pilotuntersuchung.

Übertragbarkeit

Der Autor gibt an, dass die Daten, die in dieser Studie gesammelt wurden, von anderen Institutionen genutzt werden können. Weitere Forschungen könnten seiner Ansicht nach hier angeschlossen werden. Von einer Übertragbarkeit der Ergebnisse wird somit ausgegangen.

Eine Reproduktion der Untersuchung würde vermutlich ähnliche Ergebnisse liefern. Da zur Übertragbarkeit jedoch keine expliziten Angaben gemacht werden, wird das Kriterium mit eher positiv bewertet, da es überwiegend erfüllt wurde.

II.III.VII.III Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Nützlichkeit						
1.1	Identifizierung der Beteiligten und Betroffenen	x				
1.2	Klärung der Evaluationszwecke		x			
1.3	Klarheit der Veröffentlichung		x			
2. Genauigkeit						
2.1	Beschreibung des Evaluationsgegenstandes				x	
2.2	Kontextanalyse		x			
2.3	Beschreibung von Zwecken und Vorgehen				x	
2.4	Übertragbarkeit		x			
	Gesamtbewertung	1	4	0	2	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 7: Bewertungsraster P7

II.III.VIII P 8: Thommy Eriksson, Alois Goller & Sergei Muchin (2001)

II.III.VIII.I Nützlichkeit

Identifizierung der Beteiligten und Betroffenen

Der Kurs „Planning and design of web pages“ war an Studierende gerichtet, die an kleinen oder mittelständischen Unternehmen arbeiteten. Diese Zielgruppe macht etwa 60% der Studierenden an der Chalmers Tekniska Högskola aus.

„Generally, the students participating in the course had diverging academic backgrounds. About 70% had previously produced simple web pages. The number of participating students each semester ranged from 25 to 40. For 90% of the distance education students this was the first time they participated in a distance education. Despite the fact that this was an education in web production, quite few students (8%) had previous experience in CMC technologies other than e-mail. Thus they should be comparable to the general population of web users when it comes to CMC experience. Both the distance education and the conventional education had students from the same geographical area, primarily the city of Göteborg (population 500,000) with suburbs and neighboring towns (Thommy Eriksson, Alois Goller & Sergei Muchin 2001:1).

Die Untersuchung wurde im Wintersemester 1999/2000 durchgeführt.

Das Kriterium wird voll erfüllt und wird mit positiv bewertet.

Klärung der Evaluationszwecke

In erster Linie zielte die Evaluation auf die Untersuchung der Kommunikation in einer virtuellen Lernumgebung ab. Weiterhin wurde die Zufriedenheit der Studierenden mit dem Kurs untersucht und es wurden Vergleiche in Bezug auf die Lernergebnisse angestellt.

Ziel und Zweck der Untersuchung werden angegeben. Das Kriterium wird mit positiv bewertet.

Klarheit der Veröffentlichung

Die vorliegende Publikation ist mehrheitlich strukturiert. Die zum Verständnis erforderlichen Daten werden größtenteils geliefert. Die Erhebungsinstrumente werden grob beschrieben, jedoch nicht im Detail dargestellt. Das Kriterium wird somit kaum erfüllt. Vor dem Hintergrund, dass an einen Tagungsbeitrag andere Ansprüche gestellt werden müssen als an einen Forschungsbericht, wird das Kriterium unter Berücksichtigung dessen als neutral gewertet.

II.III.VIII.II Genauigkeit

Beschreibung des Evaluationsgegenstandes

Der Kurs wurde während neun Wochen in Teilzeit angeboten. Ziel war eine holistische Annäherung zu Wissen und Fähigkeiten, die erfordert werden, um Webseiten erstellen zu können sowie Aspekte von Internet-Marketing, Projektmanagement, Webdesign, HTML und Programmierung.

1997 wurde dieser Kurs erstmalig eingesetzt und seitdem 7 Mal durchgeführt. Jedes Semester wurde er stufenweise durch neue Ansätze, neuen Inhalt oder neue Technologien weiterentwickelt. Die Kurswebseite verfügte über ein „schwarzes Brett“, Zeitpläne, Glossare, ein netzbasiertes Lehrbuch, Quizfragen, Anweisungen und Tests, einen Überblick über deren Ergebnisse und eine Newsgroup.

Der Präsenzkurs bestand aus Vorträgen und praktischen Übungen. Eine Unterrichtseinheit hatte die Dauer von 3 Stunden.

Den Kursteilnehmenden beider Gruppen wurde jeweils am Mittwochabend eine Aufgabe gestellt, die bis zum darauf folgenden Montag zu lösen war. Bevor diese abgegeben werden konnte, musste ein Multiple-Choice-Quiz absolviert werden. Beide Kurse partizipierten darüber hinaus an der gleichen Newsgroup. Die Debatten dort wurden in der Regel von der Lehrperson initiiert. Es war den Studierenden jedoch vorgeschrieben, mindestens eine Antwort zu geben.

Der Evaluationsgegenstand ist klar beschrieben worden. Das Kriterium wurde voll erfüllt und wird mit positiv bewertet.

Kontextanalyse

Der Kontext der Hochschule wird detailliert beschrieben. Darüber hinaus werden einige Bezüge zu E-Learning und Kommunikation hergestellt. Dabei wird der Stand der Forschung dargestellt und andere Publikationen in den Blick genommen.

Da die meisten Punkte jedoch nicht in aller Ausführlichkeit beschrieben werden und auf das Vorwissen der ProbandInnen sowie auf pädagogische Rahmenbedingungen nicht eingegangen wird, wird das Kriterium nur teilweise erfüllt und mit neutral bewertet.

Beschreibung von Zwecken und Vorgehen

Bei der vorliegenden Evaluation handelt es sich um eine summative Querschnittsuntersuchung.

Die eingesetzten Methoden werden beschrieben, aber nicht im Detail begründet. In der Primärstudie wird begründet, welches Evaluationsdesign zu Grunde gelegt wurde (Längs- oder Querschnitt, summative oder formative Evaluation).

Das Kriterium wurde kaum erfüllt und wird mit eher negativ bewertet.

Übertragbarkeit

Eine Reproduktion der Untersuchung würde vermutlich ähnliche Ergebnisse liefern. Zwar werden zur Übertragbarkeit keine expliziten Angaben gemacht, jedoch bezieht sich das Review der Literatur auf Beispiele anderer Studien.

Das Kriterium wird mit neutral bewertet, da es teilweise erfüllt wurde.

II.III.VIII.III Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Nützlichkeit						
1.1	Identifizierung der Beteiligten und Betroffenen	x				
1.2	Klärung der Evaluationszwecke	x				
1.3	Klarheit der Veröffentlichung			x		
2. Genauigkeit						
2.1	Beschreibung des Evaluationsgegenstandes	x				
2.2	Kontextanalyse			x		
2.3	Beschreibung von Zwecken und Vorgehen				x	
2.4	Übertragbarkeit			x		
	Gesamtbewertung	3	0	3	1	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 8: Bewertungsraster P8

II.III.IX P 9: Ruth Green & Karen Gentemann (2001)

II.III.IX.I Nützlichkeit

Identifizierung der Beteiligten und Betroffenen

Die Studierenden, die an der Untersuchung teilnahmen, waren im Wintersemester 1999/2000 in drei Online-Kurse „advanced English“ an der George Mason University eingeschrieben.

Die Beteiligten konnten mehr oder weniger identifiziert werden. Das Kriterium ist eher positiv zu bewerten.

Klärung der Evaluationszwecke

Aufbauend auf bereits vorhandenen Untersuchungsergebnissen in der Literatur möchten die Autorinnen die Effektivität von Online-Kursen durch ein Evaluationsdesign untersuchen, das nicht nur auf den reinen Vergleich der Leistung ausgerichtet ist.

„The study examines differences between the OLC students at the beginning and end of the semester (e.g., attitudes about OLC and FTF courses, perceived levels of technological expertise, use of technology) and differences in academic outcomes (e.g., course completion, grade) for students in the OLC vs. FTF (Ruth Green & Karen Gentemann 2001:3).

Ziel und Zweck der Untersuchung wurden ausführlich dargestellt. Das Kriterium wird positiv bewertet.

Klarheit der Veröffentlichung

Die Veröffentlichung folgt einer klaren Struktur und ist gut verständlich. Vereinzelt hätten die Forschungsentscheidungen detaillierter begründet werden können.

Das Kriterium wurde überwiegend erfüllt und ist mit eher positiv zu bewerten.

II.III.IX.II Genauigkeit

Beschreibung des Evaluationsgegenstandes

Die Untersuchung zielte darauf ab, die Leistungen und Einstellungen der Studierenden im Online-Kurs „advanced English“ zu zwei unterschiedlichen Zeitpunkten im laufenden Semester festzustellen und miteinander zu vergleichen. Als Kontrollgruppen diente der reguläre Kurs in Präsenzform. Beide hatten den gleichen Umfang (3 Kreditpunkte).

In Präsenzform bestand der Kurs in erster Linie darin, das Schreiben und Analysieren von Texten zu trainieren.

„English 302: Advanced Composition is a required 3-credit course with extensive pre-requisites (Completion of 45 credit hours, three credits of composition, and 6 credits of literature). The course provides intensive practice in writing and analyzing such expository forms as the essay, article, proposal, and technical and scientific report. Students must maintain a C average to fulfill degree requirements“ (ebd.).

Als Online-Angebot wurde der Kurs durch Kurs-Management-System mit einer Web-CT-Installation angeboten.

„The OLC site provided a course syllabus, academic rules and regulations, assignments and links to materials needed for each assignment. The students enrolled in the OLC were also connected via a class e-mail system (file exchange). Assignments varied but were typically scenario/case studies that required students to provide on-line, research-based responses using the university data bases and other materials identified on the website. For example, one assignment provided students a mock memo with an identified problem from a work supervisor. Students were asked to respond via electronic memo with a research-based response and recommendations for the supervisor“ (a.a.O.:4).

In zwei der drei Online-Klassen war ein Face-to-Face-Treffen zu Beginn und am Ende des Kurses vorgesehen, eine Klasse führte dies nur zu Beginn durch.

Zu den eingesetzten Materialien werden keine Angaben gemacht. Daher kann das Kriterium – wenngleich alle anderen Aspekte gut dargestellt wurden – lediglich mit neutral bewertet werden.

Kontextanalyse

Zunächst findet ein Review der relevanten Literatur statt. Hierbei wird insbesondere auf die Ergebnisse von Vergleichsstudienforschung sowie auf die didaktische Qualität von Online-Lernen Bezug genommen.

Darüber hinaus werden die Rahmenbedingungen der Universität und die typische Zielgruppe erläutert.

Auf pädagogische Voraussetzungen der Evaluation oder Vorwissen der Studierenden wird nicht speziell eingegangen.

Das Kriterium wurde überwiegend erfüllt und wird mit eher positiv bewertet.

Beschreibung von Zwecken und Vorgehen

Bei der vorliegenden Evaluation handelt es sich um eine summative Längsschnittuntersuchung.

Zur Messung der Effektivität wurden verschiedene Methoden miteinander kombiniert. Zum einen wurden schriftliche Befragungen zu Beginn und am Ende des Kurses durchgeführt^{xxxv}. Diese betrafen die Bereiche „educational strategies“, „experiences with technology“,

^{xxxv} Die Umfragen wurden mit Hilfe des Flashlight-Inventory erstellt. Hierbei handelt es sich um ein Evaluationswerkzeug, das mit externer Unterstützung entwickelt wurde.

„student academic and demographic information“ und schlossen darüber hinaus Interviews und Gruppenprotokolle ein. Hierbei ging es primär um den persönlichen Hintergrund und Erwartungen gegenüber dem Unterricht.

In den Fragebögen waren verschiedene Skalen enthalten, die auf die Bereiche „Importance of academic/social interactions“, „Expertise with technology“, „Student expectations for instructor priorities“, „Use of Seven Principles of Good Practice“, „Relationships and contacts with other students“, „Course characteristics“, „Course procedures“, „Course feedback“, „Benefits of OLC“ und „Benefits of FTF“ abzielten.

Die Antworten wurden mit Hilfe einer statistischer Methoden ausgewertet. Signifikanztests wurden in Bezug auf die Abschlussnote, Bestehensquote, Geschlecht und ethnischer Herkunft gegenüber der Präsenzklasse durchgeführt.

„For the variables that compared the students in the OLC with the FTF, data were obtained using institutional databases. OLC and FTF groups were compared based on gender, age, cumulative and course GPA and course retention/attrition rates. [...] Average course grade and retention/attrition rates were compared for all OLC and FTF students enrolled in fall 1999 in the 15 sections of English 302 designated for business majors (OLC courses n=3, FTF n=12 courses)“ (a.a.O.:12f.).

Die eingesetzten Methoden werden ausführlich beschrieben, jedoch nicht im Detail begründet.

Das Kriterium wurde teilweise erfüllt und wird mit neutral bewertet.

Übertragbarkeit

Zwar werden zur Übertragbarkeit keine expliziten Angaben gemacht, jedoch bezieht sich das Review der Literatur auf Beispiele anderer Studien. Eine Übertragung der Untersuchung an andere Hochschulen würde vermutlich ähnliche Ergebnisse liefern.

Das Kriterium wird mit neutral bewertet, da es teilweise erfüllt wurde.

II.III.IX.III Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Nützlichkeit						
1.1	Identifizierung der Beteiligten und Betroffenen		x			
1.2	Klärung der Evaluationszwecke	x				
1.3	Klarheit der Veröffentlichung		x			
2. Genauigkeit						
2.1	Beschreibung des Evaluationsgegenstandes			x		
2.2	Kontextanalyse		x			
2.3	Beschreibung von Zwecken und Vorgehen			x		
2.4	Übertragbarkeit			x		
	Gesamtbewertung	1	3	3	0	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 9: Bewertungsraster P9

II.III.X P 10: David J. Pucel & Thomas F. Stertz (2005)

II.III.X.I Nützlichkeit

Identifizierung der Beteiligten und Betroffenen

Die Untersuchung wurde zwischen den Jahren 2001 und 2003 an insgesamt 84 Lehramtsstudierenden in (44 in Online-Kursen und 40 im Präsenzunterricht) durchgeführt, die zu dieser Zeit die Kurse „History and Philosophy of Career and Technical Education“ und „Instructional Methods for Business and Industry“ an der George Mason University belegt hatten. Diese Kurse wurden von zwei Doktoranden des Faches konzipiert und durchgeführt.

Die Beteiligten konnten gut identifiziert werden. Das Kriterium ist positiv zu bewerten.

Klärung der Evaluationszwecke

Die vorliegende Studie zielte darauf ab, die Zufriedenheit der Studierenden mit zwei Online-Kursen und deren Leistung im Vergleich zu zwei gleichwertigen Präsenzkursen zu untersuchen. Dabei wurde das Ziel verfolgt, ein generalisierbares Modell zur Evaluation von Online-Kursen im Vergleich zu traditionellen Kursen zu entwickeln und gleichzeitig festzustellen, ob diese gleiche Leistungen der Studierenden hervorbringen.

Die Autoren geben darüber hinaus an, generell mehr Forschung in Bezug auf die Messung der Effektivität von Online-Kursen durchführen zu wollen. Dies könne durch die Lernleistung und die Zufriedenheit der Studierenden realisiert werden.

Ziel und Zweck der Untersuchung werden benannt. Das Kriterium ist mit positiv zu bewerten.

Klarheit der Veröffentlichung

Die Veröffentlichung folgt einer nachvollziehbaren Struktur und ist gut verständlich. Vereinzelt hätten die Forschungsentscheidungen detaillierter begründet werden können.

Das Kriterium wurde überwiegend erfüllt und ist mit eher positiv zu bewerten.

II.III.X.II Genauigkeit

Beschreibung des Evaluationsgegenstandes

Die Evaluation wurde an Hand von 4 Kursen durchgeführt, die über einen Zeitraum von sieben Wochen angeboten wurden:

1. „History and Philosophy of Career and Technical Education“ als Online-Kurs

2. „Instructional Methods for Business and Industry“ als Online-Kurs
3. „History and Philosophy of Career and Technical Education“ als Präsenzkurs
4. „Instructional Methods for Business and Industry“ als Präsenzkurs

Dabei waren Ziele und Inhalte der jeweiligen Online- und Präsenzversionen nahezu identisch, da die Inhalte der traditionellen Kurse übertragen wurden:

„To ensure that the instruction in the WBI-enhanced versions of the courses was as nearly equivalent as possible to the past traditional versions, two instructors with Ph.D.'s in education who had long histories of teaching the courses in a traditional manner, and who were interested in online instruction, re-formatted the courses for WBI. Specialists in the development of online instruction from the university's College of Education and Human Development provided guidance in using WebCT to assemble the material for online delivery“ (David J. Pucel & Thomas F. Stertz 2005:12).

Auch die Online-Kurse hatten einen Präsenzanteil von insgesamt sechs Stunden, in denen beispielsweise die Präsentationen erstellt oder die Prüfungen absolviert wurden.

Die technische Umsetzung wurde mit Web-CT realisiert:

„Each WBI course contained a detailed syllabus, instructional modules designed to lead students through the learning activities and related resources, project guidelines and expectations, and exercises that required students to participate in cooperative learning activities. In addition, the course materials included Web links, video clips, audio clips, and PowerPoint presentations“ (a.a.O.:12).

Der Präsenzkurs in Philosophie wurde bereits während des Wintersemesters 2002 untersucht, der Präsenzkurs in Methoden im Sommer 2001. Nähere Angaben werden zu diesen Kursen nicht gemacht. Auch über das Kursmaterial ist nichts bekannt.

Das Kriterium wurde teilweise erfüllt und wird mit neutral bewertet.

Kontextanalyse

Die Autoren führen zunächst ein Review der relevanten Literatur durch, welche sich primär auf die Lehrerbildung sowie Web-Based-Instruction und Lernleistungen von Studierenden bezieht. Ferner wird auf die Spezifika von Vergleichsstudienforschung eingegangen.

Darüber hinaus werden die Rahmenbedingungen Lehrerbildung an der Georg Mason University erläutert:

„Due to a continuing lack of qualified individuals who are both occupationally competent and possess teaching degrees, schools hire occupationally competent people and provide them with teacher education as they are teaching. The occupational teachers teach in fields such as communications technology, construction

careers, creative design careers, early childhood careers, hospitality service careers, manufacturing careers, medical careers, and transportation careers. These new teachers are sparsely spread out throughout the state of Minnesota [...]Therefore, the Minnesota Department of Education asked the University of Minnesota to develop online- enhanced WBI course alternatives for some of its basic pedagogical courses [...]. It was hoped that offering the classes in a WBI format would reduce the amount of travel and in-classroom time and provide flexibility for students completing course requirements without compromising either student satisfaction or performance“(a.a.O.:11f.).

Auf pädagogische Voraussetzungen der Evaluation oder Vorwissen der Studierenden wird nicht speziell eingegangen.

Das Kriterium wurde überwiegend erfüllt und wird mit eher positiv bewertet.

Beschreibung von Zwecken und Vorgehen

Bei der vorliegenden Evaluation handelt es sich um eine summative Querschnittsuntersuchung, wenngleich die Messungen zu unterschiedlichen Zeitpunkten durchgeführt worden waren. Die Gründe für die Wahl dieses Evaluationsdesigns werden nicht begründet.

Die Zufriedenheit und Leistungen der Studierenden wurden sowohl in den Online- als auch in den Präsenzversionen der Kurse mittels derselben Kriterien evaluiert. Die Untersuchungsmethoden werden beschrieben (siehe „valide und reliable Informationen“), aber nicht im Detail begründet.

Das Kriterium wurde teilweise erfüllt und somit als neutral gewertet.

Übertragbarkeit

Die Autoren äußern sich zufrieden über das entwickelte Evaluationskonzept:

„In addition, it appears that the general model used to evaluate these courses yielded meaningful results and is feasible to implement. Therefore, this model may provide an effective method of evaluating other similar WBI courses“ (a.a.O.:21).

Zur Übertragbarkeit der Ergebnisse werden keine Angaben gemacht. Es sind jedoch ähnliche Resultate unter gleichen Bedingungen zu erwarten (keine spezifische Nutzung der neuen Medien, gleiches Kursmaterial, unterschiedliche Gruppenzusammensetzungen). Das Kriterium wird daher als neutral gewertet.

II.III.X.III Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Nützlichkeit						
1.1	Identifizierung der Beteiligten und Betroffenen	x				
1.2	Klärung der Evaluationszwecke	x				
1.3	Klarheit der Veröffentlichung		x			
2. Genauigkeit						
2.1	Beschreibung des Evaluationsgegenstandes			x		
2.2	Kontextanalyse		x			
2.3	Beschreibung von Zwecken und Vorgehen			x		
2.4	Übertragbarkeit			x		
	Gesamtbewertung	2	2	3	0	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 10: Bewertungsraster P10

II.III.XI P11: Larry S. Tinnerman (2006)

II.III.XI.I Nützlichkeit

Identifizierung der Beteiligten und Betroffenen

An der im Jahr 2003 durchgeführten Untersuchung nahmen 47 Studierende teil, die in einen Kurs zu Lernbehinderungen im Abschlussjahrgang eingeschrieben waren. Diese waren zwischen 18 und 25 Jahre alt und hatten unterschiedliche sozioökonomische Hintergründe.

Die Untersuchung wurde vermutlich an der Indiana University of Pennsylvania durchgeführt.

Die Beteiligten und Betroffenen konnten überwiegend identifiziert werden. Das Kriterium wird mit eher positiv bewertet.

Klärung der Evaluationszwecke

Es sollten zwei Kurse miteinander verglichen werden, die von der gleichen Lehrperson sowohl in einem Präsenzkurs als auch online gelehrt wurden. Das Forschungsinteresse lag hierbei sowohl auf den Lernleistungen der Studierenden als auch auf ihren Meinungen und Haltungen bezüglich der Kursdidaktik und –organisation.

Beispielweise wurde zusätzlich erhoben:

- „The awkwardness of the communication effectively using a nonverbal typed method such as chat or discussion forum
- The compatibility of various software packages
- The difficulty of teacher modeling for the students
- The demand for the instructor to be both a content expert and a technical expert“ (Larry S. Tinnerman 2006:o.S.)

Ziel und Zweck der Untersuchung wurden benannt. Das Kriterium ist mit positiv zu bewerten.

Klarheit der Veröffentlichung

Die Veröffentlichung folgt einer nachvollziehbaren Struktur und ist gut verständlich. Vereinzelt hätten die Forschungsentscheidungen detaillierter begründet werden können.

Das Kriterium wurde überwiegend erfüllt und ist mit eher positiv zu bewerten.

II.III.XI.II Genauigkeit

Beschreibung des Evaluationsgegenstandes

Die Untersuchung wurde an Hand von zwei Kursen zum Thema Lernbehinderungen durchgeführt. Dabei wurde die Instruktion entweder im Klassenraum, oder online gegeben. In beiden Fällen handelte es sich um dieselbe Lehrperson.

Über das Kursmaterial ist nichts bekannt.

Das Kriterium wurde teilweise erfüllt und ist mit neutral zu bewerten.

Kontextanalyse

Der Kontext der Untersuchung wird durch Referenzen zum Stand der Forschung in Bezug auf E-Learning und Fernlehre dargestellt. Diese werden in Beziehung zum Wandel der akademischen Ausbildung gesetzt. Insbesondere wird auf die Relevanz der Forschung hinsichtlich Behinderung und Rehabilitation hingewiesen.

Die Rahmenbedingungen der Studie werden erläutert. In einer demographischen Erhebung wird auf die speziellen persönlichen Hintergründe der Studierenden eingegangen.

Das Kriterium wurde überwiegend erfüllt und wird mit eher positiv bewertet.

Beschreibung von Zwecken und Vorgehen

Bei der vorliegenden Evaluation handelt es sich um eine summativ Längsschnittuntersuchung. Die Gründe für die Wahl dieses Evaluationsdesigns werden nicht begründet.

Die Zufriedenheit und Leistungen der Studierenden wurden sowohl in den Online- als auch in den Präsenzversionen der Kurse mittels derselben Kriterien evaluiert. Die Untersuchungsmethoden werden – teilweise sehr detailliert – beschrieben (siehe „valide und reliable Informationen“), aber nicht im Detail begründet.

Das Kriterium wurde teilweise erfüllt und somit als neutral gewertet.

Übertragbarkeit

Auf den Aspekt der Übertragbarkeit wird explizit eingegangen.

„While this study is particularly restricted to one course by one professor teaching in the two investigated instructional methodologies, it is believed that certain aspects of this study will generalize to other sites and similar subjects [...].

This study can be used as a tool in conjunction with other like investigations through other instructional venues in order to

provide a cross comparison of similarities and differences. In particular, this research should help in the planning and construction of future courses offered by universities at the graduate level to make the course offering the most effective experience for the student that is possible in both the traditional and distance educational environment. This comparative study can supply hints as to what is instructionally effective regardless of the method of delivery" (ebd.).

Der Autor arbeitet darüber hinaus Empfehlungen heraus, die bei zukünftigen Studien Berücksichtigung finden sollten. Dazu gehört beispielsweise das Verhältnis zwischen der Geschwindigkeit der Instruktion und der Strenge im Präsenzunterricht und die Auswirkungen der Regulation im Online-Kurs.

Das Kriterium wird mit eher positiv bewertet, da es überwiegend erfüllt wurde.

II.III.XI.III Bewertungsraster

Nr.	Kriterium	*	#	+	0	–
1. Nützlichkeit						
1.1	Identifizierung der Beteiligten und Betroffenen	x				
1.2	Klärung der Evaluationszwecke	x				
1.3	Klarheit der Veröffentlichung		x			
2. Genauigkeit						
2.1	Beschreibung des Evaluationsgegenstandes			x		
2.2	Kontextanalyse		x			
2.3	Beschreibung von Zwecken und Vorgehen			x		
2.4	Übertragbarkeit		x			
	Gesamtbewertung	2	3	2	0	0
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 11: Bewertungsraster P11

II.IV Vergleich der Primärstudien im Hinblick auf die übrigen Kriterien

An dieser Stelle sollen die Bewertungen zusammengefasst werden, die die Einzelstudien auf Grund der formalen Kriterien erhalten haben. Diese wurden – um nicht den Rahmen der Arbeit zu sprengen – im Anhang behandelt, da sie sich nicht unmittelbar auf die Forschungsfrage bezogen.

II.IV.XI.I Auswertung der Einzelkriterien

In der folgenden Tabelle sind die Einzelbewertungen übersichtlich dargestellt. Markierungen geben an, dass diese Bewertungen in den meisten Fällen vergeben wurden.

Nr.	Kriterium	*	#	+	0	–
1. Nützlichkeit						
1.1	Identifizierung der Beteiligten und Betroffenen	7	2	2	0	0
1.2	Klärung der Evaluationszwecke	10	1	0	0	0
1.3	Klarheit der Veröffentlichung	1	9	1	0	0
2. Genauigkeit						
2.1	Beschreibung des Evaluationsgegenstandes	3	2	3	3	0
2.2	Kontextanalyse	3	7	1	0	0
2.3	Beschreibung von Zwecken und Vorgehen	4	1	4	2	0
2.4	Übertragbarkeit	0	2	7	1	1
*positiv		# eher positiv		+ neutral		0 eher negativ
						- negativ

Tabelle 12: zusammenfassende Bewertungen der Kriterien

Es soll nun genauer auf die Einzelkriterien eingegangen werden:

Identifizierung der Beteiligten und Betroffenen

Es wird deutlich, dass die Beteiligten und Betroffenen in den meisten Fällen sehr gut identifiziert werden konnten. 7 von 11 Studien erhielten hier die höchste Bewertung. Keine der Studien wurde im negativen Bereich bewertet.

Klärung der Evaluationszwecke

Ziel und Zweck der Untersuchungen wurden in allen Studien benannt. Das Kriterium wurde nur in einem Fall als eher positiv, sonst durchgängig als sehr positiv bewertet.

Klarheit der Veröffentlichung

Die Veröffentlichungen folgen in der Regel einer nachvollziehbaren Struktur und sind gut verständlich. In einigen Fällen wären zusätzliche Begründungen hilfreich gewesen.

Das Kriterium wurde überwiegend als eher positiv bewertet. Keine Bewertung lag im negativen Bereich.

Beschreibung des Evaluationsgegenstandes

Dieses Kriterium weist die breiteste Streuung auf, wurde aber überwiegend positiv bewertet. Jedoch müssen Aussagen im Hinblick auf dieses Kriterium immer auf der Basis der Einzelstudie getroffen werden. Auffällig ist, dass oftmals wenig über das Kursmaterial bekannt ist. Auch werden zusätzliche Informationen zu den Kursen (z.B. in Bezug auf die Lehrperson) oft nicht gegeben.

Kontextanalyse

Der Kontext der Untersuchung wird in den meisten Fällen hergestellt. Dabei beziehen sich die Darstellungen oftmals auf den Stand der Forschung im Fachgebiet – es werden andere AutorInnen oder Diskussionen referiert. Auch wird auf die Rahmenbedingungen der Hochschule Bezug genommen. So werden beispielsweise das Einzugsgebiet oder Einschreibungs Voraussetzungen thematisiert. Seltener wird auf die Rahmenbedingungen der untersuchten Kurse eingegangen, z.B. im Hinblick auf das Vorwissen der Untersuchungspersonen. Dennoch kann das Kriterium zusammengefasst als erfüllt betrachtet werden.

Beschreibung von Zwecken und Vorgehen

Auch dieses Kriterium wird unterschiedlich beurteilt und ist wenig einheitlich. Dies ist an der stark streuenden Verteilung der Bewertungen zu erkennen. Die Untersuchungsmethoden werden in den meisten Fällen sehr umfassend dargestellt, aber es fehlen detaillierte Begründungen. Das gleiche gilt für die Gründe für die Wahl des Evaluationsdesigns. Das Kriterium wurde jedoch überwiegend positiv bewertet.

Übertragbarkeit

Die Übertragbarkeit ist in den meisten Fällen nicht ohne weiteres gegeben. Die Bewertungen von 7 der 11 Studien liegen hier im neutralen, 2 im negativen Bereich. M.E. ist dies damit zu begründen, dass die

Einzelstudien zu speziell auf die spezifischen Rahmenbedingungen der Untersuchung zugeschnitten waren.

II.IV.XI.II Zusammenfassung

Es wurde deutlich, dass die Primärstudien die formalen Kriterien überwiegend erfüllen konnten. Insbesondere die Klarheit der Veröffentlichung und die Klärung der Evaluationszwecke wurden in allen Primärstudien überzeugend dargestellt.

Bei der Beschreibung des Evaluationsgegenstandes wären noch weitere Angaben hilfreich gewesen. Darüber hinaus hätten die Entscheidungen für oder gegen ein Forschungsdesign oder die methodische Ausrichtung detaillierter begründet werden können.

Ebenfalls war die Übertragbarkeit der Ergebnisse nicht durchgängig gewährleistet. Dies lag einerseits daran, dass die Untersuchungen selbst zu speziellen Charakter hatten, andererseits hinaus führten Abstriche bei der Bewertung der anderen Kriterien dazu, dass die Verallgemeinerbarkeit der getroffenen Aussagen relativiert werden musste.

III Übersicht über die Operationalisierung der kognitiven Prozesse nach Lorin W. Anderson und David R. Krathwohl (2001)

CATEGORIES & COGNITIVE PROCESSES	ALTERNATIVE NAMES	DEFINITIONS AND EXAMPLES
1. REMEMBER—Retrieve relevant knowledge from long-term memory		
1.1 Recognizing	Identifying	Locating knowledge in long-term memory that is consistent with presented material (e.g., Recognize the dates of important events in United States history)
1.2 Recalling	Retrieving	Retrieving relevant knowledge from long-term memory (e.g., Recall the dates of important events in United States history)
2. UNDERSTAND—Construct meaning from instructional messages, including oral, written, and graphic communication		
2.1 Interpreting	Clarifying, paraphrasing, representing, translating	Changing from one form of representation (e.g., numerical) to another (e.g., verbal) (e.g., Paraphrase important speeches and documents)
2.2 Exemplifying	Illustrating, instantiating	Finding a specific example or illustration of a concept or principle (e.g., Give examples of various artistic painting styles)
2.3 Classifying	Categorizing, subsuming	Determining that something belongs to a category (e.g., Classify observed or described cases of mental disorders)
2.4 Summarizing	Abstracting, generalizing	Abstracting a general theme or major point(s) (e.g., Write a short summary of events portrayed on a videotape)
2.5 Inferring	Concluding, extrapolating, interpolating, predicting	Drawing a logical conclusion from presented information (e.g., In learning a foreign language, infer grammatical principles from examples)
2.6 Comparing	Contrasting, mapping, matching	Detecting correspondences between two ideas, objects, and the like (e.g., Compare historical events to contemporary situations)
2.7 Explaining	Constructing models	Constructing a cause-and-effect model of a system (e.g., Explain the causes of important 18th Century events in France)
3. APPLY—Carry out or use a procedure in a given situation		
3.1 Executing	Carrying out	Applying a procedure to a familiar task (e.g., Divide one whole number by another whole number, both with multiple digits)
3.2 Implementing	Using	Applying a procedure to an unfamiliar task (e.g., Use Newton's Second Law in situations in which it is appropriate)

Abbildung 1: Die kognitive Prozessdimension nach Lorin W. Anderson & David R. Krathwohl (2001), Teil 1

CATEGORIES & COGNITIVE PROCESSES	ALTERNATIVE NAMES	DEFINITIONS AND EXAMPLES
4. ANALYZE—Break material into its constituent parts and determine how the parts relate to one another and to an overall structure or purpose		
4.1 Differentiating	Discriminating, distinguishing, focusing, selecting	Distinguishing relevant from irrelevant parts or important from unimportant parts of presented material (e.g., Distinguish between relevant and irrelevant numbers in a mathematical word problem)
4.2 Organizing	Finding coherence, integrating, outlining, parsing, structuring	Determining how elements fit or function within a structure (e.g., Structure evidence in a historical description into evidence for and against a particular historical explanation)
4.3 Attributing	Deconstructing	Determine a point of view, bias, values, or intent underlying presented material (e.g., Determine the point of view of the author of an essay in terms of his or her political perspective)
5. EVALUATE—Make judgments based on criteria and standards		
5.1 Checking	Coordinating, detecting, monitoring, testing	Detecting inconsistencies or fallacies within a process or product; determining whether a process or product has internal consistency; detecting the effectiveness of a procedure as it is being implemented (e.g., Determine if a scientist's conclusions follow from observed data)
5.2 Critiquing	Judging	Detecting inconsistencies between a product and external criteria; determining whether a product has external consistency; detecting the appropriateness of a procedure for a given problem (e.g., Judge which of two methods is the best way to solve a given problem)
6. CREATE—Put elements together to form a coherent or functional whole; reorganize elements into a new pattern or structure		
6.1 Generating	Hypothesizing	Coming up with alternative hypotheses based on criteria (e.g., Generate hypotheses to account for an observed phenomenon)
6.2 Planning	Designing	Devising a procedure for accomplishing some task (e.g., Plan a research paper on a given historical topic)
6.3 Producing	Constructing	Inventing a product (e.g., Build habitats for a specific purpose)

Abbildung 2: Die kognitive Prozessdimension nach Lorin W. Anderson & David R. Krathwohl (2001), Teil 2

Erklärung

Hiermit versichere ich, dass ich die vorliegende Dissertation

„Wir evaluieren uns zu Tode.
Möglichkeiten und Grenzen der
Bewertung von Online-Lernen.
Eine Meta-Evaluation.“

selbstständig und ohne unerlaubte fremde Hilfe angefertigt und andere als die in der Dissertation angegebenen Hilfsmittel nicht benutzt habe.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Schriften entnommen sind, habe ich als solche kenntlich gemacht.

Die vorliegende Dissertation hat zuvor keiner anderen Stelle zur Prüfung vorgelegen. Es ist mir bekannt, dass wegen einer falschen Versicherung bereits erfolgte Promotionsleistungen für ungültig erklärt werden und eine bereits verliehene Doktorwürde entzogen wird.

Ort, Datum

Annabell Preußler