Universität Trier

Fachbereich I – Psychologie

Studentische Lehrveranstaltungsevaluation im Internet

Diplomarbeit

vorgelegt von

Joachim Schroer

im August 2003

Gutachter und Betreuer:

Dr. Ewald Naumann

Prof. Dr. Dieter Bartussek

Inhaltsverzeichnis

Studentische Lehrveranstaltungsevaluation im Internet

> *It is readily conceivable that an instructor may be using [...] a rating scheme [to] improve his classroom procedure and this is after all the ultimate aim. Of course such improvement [...] would be based on the assumption that students' ratings in the main have some degree of validity. And this, we believe, is not a dangerous assumption.*
>
> (Brandenburg & Remmers, 1927, S. 402)

1 Einleitung

Spätestens seit den Empfehlungen des Wissenschaftsrates (1996) „zur Stärkung der Lehre in den Hochschulen durch Evaluation“, die sich zum großen Teil mit denen des amerikanischen Joint Committee on Standards for Educational Evaluation (1994) decken, beschäftigt das Thema *Lehrevaluation* neben den betroffenen Studenten und Dozenten an den Hochschulen (vgl. Krampen & Zayer, 2000) auch die politische Diskussion (el Hage, 1996). Ziel der Lehrevaluation ist es, die Qualität der Hochschulausbildung zu gewährleisten. Der Wissenschaftsrat räumt allerdings ein, dass der damit verbundene Begriff der *Qualität der Lehre* seinerseits einer Festlegung bedarf:

> In engem Zusammenhang mit der Zielsetzung eines Lehrevaluationsverfahrens steht die Frage, was als „Qualität“ der „Lehre“ zu gelten hat. Der Begriff „Qualität“ entzieht sich einer allgemein gültigen inhaltlichen Definition. „Qualität“ ist ein multidimensionaler Begriff, der je nach Erkenntnisinteresse verschiedenartige Ausprägungen erfahren kann. (S. 62)

Der Wissenschaftsrat (1996) unterscheidet fünf Aspekte von Qualität:

1. Das *Ausbildungsprofil der Absolventen* sollte zuvor festgesetzten Normen und Standards entsprechen.
2. *Konsistenz und Kohärenz* der Ausbildung sollten zu einer stimmigen Organisation der Ausbildung führen.
3. *Studienziele und Ausbildungsprofile* sollten die Absolventen möglichst gut für die spätere Berufstätigkeit vorbereiten.
4. Die *effiziente Verwendung von Geldern* als „günstige[s] Verhältni[s] zwischen eingesetzten Mitteln und dem damit erzielten Ergebnis“ betrifft die Frage der materiellen und personellen

Ausstattung von Hochschulen.

5. Schließlich soll das Studium als *Qualifizierung* auch Bildung, Ausbildung, Persönlichkeitsentwicklung und Wissenserwerb vermitteln.

Diese Aufzählung zeigt, dass Lehrevaluation ein breites Spektrum an Bewertungskriterien umfasst (vgl. Wottawa & Thierau, 1998), die sich zum größten Teil auf die Rahmenbedingungen und das Curriculum beziehen. Die Evaluation der Qualität einer einzelnen Lehrveranstaltung wird in den Vorschlägen des Wissenschaftsrates (1996) kaum berücksichtigt. An der Selbstevaluation eines Faches sollen jedoch alle Gruppen beteiligt werden:

> Die erste Stufe der Evaluation (Selbstevaluation) [...] setzt voraus, dass die Ermittlung der Befunde und ihre Zusammenführung in einem Ergebnisbericht im Fachbereich auf breiter Basis erfolgt und alle am Ausbildungsprozess beteiligten Gruppen einbezieht. Hierzu gehört insbesondere, dass die Studierenden eines Fachbereichs von Anfang an in angemessener Weise in den Abstimmungsprozess eingebunden werden. (S. 70)

Diese Sichtweise ist seit 1998 auch gesetzlich verankert: das Hochschulrahmengesetz (HRG) schreibt in § 6 vor, dass „[d]ie Studierenden [...] bei der Bewertung der Qualität der Lehre zu beteiligen“ sind (Deutscher Bundestag, 1998). Die *studentische Evaluation von Lehrveranstaltungen*[1] kann damit ein Element des umfassenderen Prozesses der *Lehrevaluation* sein, aber daneben und unabhängig davon auch als Rückmeldung an den Dozenten dienen: die Qualität einer Lehrveranstaltung versteht man hier als Effektivität der Wissensvermittlung (*teaching effectiveness*, vgl. Marsh, 1987, S. 259).

Im Vergleich zum heftig umstrittenen Begriff der allgemeinen Lehrevaluation kann die Forschung zur studentischen Evaluation von Lehrveranstaltungen auf einen umfangreichen Schatz empirischer Studien zurückgreifen, die bis in die zwanziger Jahre zurückgehen (Brandenburg & Remmers, 1927). Marsh beschreibt in seiner Monographie von 1987 den Umfang der Studien zu diesem Thema bereits als kaum überschaubar: Während viele Studien aus den siebziger Jahren jedoch zum Teil erhebliche methodische Mängel aufwiesen, seien seitdem sowohl methodische als auch theoretische Fortschritte gemacht worden (vgl. dazu auch Rindermann, 2001). Abschnitt 2.1 skizziert den gegenwärtigen Stand der Diskussion.

[1] Cashin (1995) bevorzugt den Ausdruck “student ratings” statt “student evaluations”: Dadurch werde deutlicher, dass es sich um Daten handle, die interpretiert werden müssten. Da in der Literatur ansonsten aber die Begriffe *student evaluation of teaching* (SET) oder *studentische Lehrveranstaltungsevaluation* verbreitet sind, werde ich sie in dieser Arbeit ebenfalls verwenden.

Trotz der breiten Forschung zu diesem Thema scheitert der praktische Einsatz von Evaluationsinstrumenten in der Hochschullehre allzu häufig am hohen Aufwand, der mit der Datenerhebung und -auswertung verbunden ist. Als Lösung bietet es sich an, die Befragung der Studenten computergestützt im Internet oder Campusnetz durchzuführen (Abschnitt 2.2), da sich so die Erfassung der Antworten und ihre Auswertung weitgehend automatisieren lassen (Hänsgen, 1999). Eine entscheidende Rolle spielt allerdings das *Problem der Äquivalenz* (Abschnitt 2.2.1.1) zwischen der papierbasierten und computergestützten Erhebungsform.

2 Theoretischer Hintergrund

Lassen sich die bisherigen Ergebnisse zur Lehrveranstaltungsevaluation auf die Datenerhebung im Internet übertragen? Welche psychometrischen Unterschiede gibt es zwischen den beiden Erhebungsformen? Diesen Fragen geht die vorliegende Untersuchung nach. Neben der Äquivalenz von Papier- und Internetfragebögen zur Lehrveranstaltungsevaluation als Hauptfragestellung werden als Ergänzung auch eine Reihe von möglichen Störeinflüssen näher betrachtet.

Die Frage der Äquivalenz zwischen papierbasierten Fragebögen und dem Einsatz computerbasierter Instrumente berührt viele Bereiche der psychologischen Forschung; Abschnitt 2.2 stellt die Ergebnisse einiger bisher vorliegender Studien und ihre Implikationen vor. Im Zusammenhang mit dem Einsatz bei der studentischen Evaluation von Lehrveranstaltungen (2.1) führen diese schließlich zu den Hypothesen dieser Untersuchung (2.3).

2.1 Studentische Lehrveranstaltungsevaluation

Sowohl die studentische Evaluation von Lehrveranstaltungen (engl. *student evaluation of teaching*, SET) als auch die dahinter stehenden Konstrukte, Messinstrumente (2.1.1) und Gütekriterien (2.1.2), vor allem aber ihre Interpretation (2.1.4) sind in der Forschung heftig umstritten (Greenwald, 1997). Dennoch erlaubt es die breite Diskussion über diese Themen ein Fazit zu ziehen (2.1.5).

2.1.1 Merkmale guter Lehre und ihre Messung

Die meisten Autoren sind sich einig, dass Lehre eine komplexe Aktivität ist, die nur mit multidimensionalen Instrumenten angemessen bewertet werden kann (Marsh & Roche, 1997; Cashin, 1995; für eine Übersicht vgl. Theall & Franklin, 1991, S. 84). Es gibt daher eine Vielzahl solcher Instrumente zur studentischen Lehrveranstaltungsevaluation, die mit unterschiedlichen theoretischen und methodischen Ansätzen entwickelt wurden. Da sich die Konstruktentwicklung auf empirische, vor allem Faktoren- und Skalen-Analysen stützt, möchte ich einige Erhebungsinstrumente vorstellen. International ist das SEEQ-Inventar weit verbreitet (2.1.1.1), in Deutschland werden unter anderem das HILVE (2.1.1.2) und das TRIL (2.1.1.3) verwendet.

2.1.1.1 Students' Evaluation of Educational Quality (SEEQ)

Das Inventar *Students' Evaluation of Educational Quality* (SEEQ) wurde zu Beginn der achtziger Jahre an der University of Southern California zur Evaluation von Lehrveranstaltungen durch Studierende entwickelt (Marsh, 1982, 1987). In einem mehrstufigen Prozess der Item- und Skalenanalyse wurden folgende Faktoren gebildet (vgl. Anhang A):

1. *Learning/Value*
2. *Instructor Enthusiasm*
3. *Organization/Clarity*
4. *Group Interaction*
5. *Individual Rapport*
6. *Breadth of Coverage*
7. *Examinations/Grading*
8. *Assignments/Readings*
9. *Workload/Difficulty*

Marsh und Roche (1997) berichten von zahlreichen Belegen für die Stabilität dieser Faktorenstruktur: So konnten die neun Faktoren bei einer Kreuzvalidierung an Daten aus 50000 Kursen mit fast einer Million einzelner Fragebögen belegt werden, und zwar sowohl für die Gesamtgruppe, als auch bei 21 einzelnen Analysen verschiedener Untergruppen aus unterschiedlichen Fächern (Marsh & Hocevar, 1991). Seit seiner Veröffentlichung (Marsh, 1982) wurde der Fragebogen in mehrere Sprachen übersetzt und auch außerhalb der USA eingesetzt (z. B. Marsh, Hau, Chung & Siu, 1997). Damit kann es als eines der weltweit verbreitetsten Instrumente studentischer Lehrveranstaltungsevaluation gelten. An deutschen Hochschulen werden jedoch vorwiegend selbst entwickelte oder bewährte deutschsprachige Inventare eingesetzt (für eine Auswahl vgl. el Hage, 1996, Kap. 8), die für den deutschsprachigen Raum konstruiert wurden.

2.1.1.2 Heidelberger Inventar zur Lehrveranstaltungs-Evaluation (HILVE)

Das *Heidelberger Inventar zur Lehrveranstaltungs-Evaluation* (HILVE) war seit seiner Veröffentlichung (Rindermann & Amelang, 1994) Gegenstand heftiger Diskussionen (z. B. Kromrey, 1996a, 1996b). Bei seiner Entwicklung stand der Einsatz in Vorlesungen und Seminaren im Vordergrund. Der Fragebogen sollte neben der Lehre des Dozenten und Merkmalen der Veranstaltung auch das Verhalten der Studenten erfassen; ausdrücklich nicht erfragt wurden studentische Bewertungen der inhaltlichen Qualität der Vorlesung. Rindermann und Amelang (1994) verfolgten als Nahziel des Instrumentes, dass Dozenten eine Rückmeldung zur eigenen Lehre erhalten sollten, die dann zur Verbesserung des Unterrichts beitragen solle. Als Fernziel könne damit die Lehre

neben der Forschung an Bedeutung gewinnen. Die Autoren sahen einen Fragebogen als am besten geeignet für diese Aufgabe an, der „breit gefächerte Aspekte der Lehre“ unter Beachtung von Gütekriterien wie Reliabilität und Urteilerübereinstimmung erhebt.

Das Ziel der Autoren war es daher, einen solchen Fragebogen zu entwickeln, da zum Zeitpunkt der Entwicklung noch keine deutschsprachigen, nach psychometrischen Kriterien entwickelten Verfahren dieser Art vorlagen. Rindermann (1996b, S. 25; vgl. auch Rindermann & Amelang, 1994) gibt vier Quellen an, aus denen Items für diesen Zweck entnommen oder entwickelt wurden:

1. Befragung von Studenten nach Merkmalen guter Lehre ($n = 125$; *explorative* Auswahlstrategie, vgl. Gollwitzer und Schlotz, 2003),
2. Befragungen von Dozenten nach Merkmalen guter Lehre an Hochschulen ($n = 9$);
3. bereits vorliegende deutsche und englische Fragebögen (*internale* Auswahlstrategie, Fisseni, 1997) sowie
4. Konstrukte guter universitärer Lehre aus der pädagogischen und didaktischen Literatur (*rationale* Strategie).

Rindermann (1996a, S. 131) sieht den „Einbezug handlungsnahen Wissens in wissenschaftliche Konstruktionsverfahren [als] bewährtes Verfahren zur Sicherstellung von Praxisadäquanz entwickelter Verfahren“, denn dadurch werde der Einfluss einseitiger theoretischer Modelle vermieden. Er weist zudem darauf hin, dass sich die empirisch durch Befragung der Praktiker (d.h. Studenten und Dozenten) ermittelten Dimensionen guter Lehre „nicht gravierend“ von theoretischen Unterrichtsprinzipien unterscheiden. Anders als in amerikanischen Inventaren (etwa dem SEEQ, vgl. 2.1.1.1) wird der Aspekt der *Notenfairness* als Bewertungskriterium nur durch optio-

Tabelle 1: Skalen des HILVE-I (nach Rindermann, 1996a). Mögliche Biasvariablen stehen in Klammern.

Dozenten- und Lehreffektivitätsskalen		**Studentische Skalen und Rahmenbedingungen**	
Lehrverhalten des Dozenten	**Lehreffektivitätsskalen**	**Studentische Skalen**	**Anforderungen**
Struktur(ierung) Breite Verarbeitung Lehrkompetenz Dozentenmanagement Klima (Doz.-Stud.) (Popularität)	Interessantheit Lernen Allgemeinbeurteilung	Referate Beteiligungen Diskussion (Thema) (Besuchsgrund) (Besuchszahl)	Überforderung Fleiß

Anmerkung: Breite und Verarbeitung bilden die Skala Auseinandersetzung

nale Zusatz-Items erfasst, da an deutschen Universitäten nur selten veranstaltungsbezogene Prüfungen unmittelbar am Ende des Semesters stehen. Tabelle 1 zeigt die per Faktoren- und Skalenanalyse aus dem Itempool gewonnenen Dimensionen guter Lehre, Anhang B enthält die Items der einzelnen Skalen des HILVE-I, Anhang C die überarbeiteten Skalen des HILVE-II.

Rindermann und Amelang (1994) berichten von einer einfachen Messwiederholung nach zwei Monaten, bei der sich bei keiner Skala und keinem Item bedeutende Messzeitpunktdifferenzen zeigten. Die Varianzaufklärung durch den Messzeitpunkt erreichte höchstens 2%: Die Mittelwerte der studentischen Einschätzungen können also als zeitlich sehr stabil angesehen werden (dazu ausführlich Rindermann, 1996b, Kap. 16).

Die erhobenen Biasvariablen (vgl. Tabelle 1) weisen nur zum Teil Korrelationen mit den erhobenen Dimensionen auf. Am wichtigsten ist der *Besuchsgrund*. Generell werden solche Veranstaltungen besser bewertet, die man aus Interesse besucht, als diejenigen, die in der Prüfungsordnung vorgeschrieben sind oder in denen die Teilnehmer nur einen Schein erwerben möchten.

Rindermann und Amelang (1994) konnten allerdings keine Einflüsse der Studienzufriedenheit, der Abitur- oder Vordiplomsnote, des Geschlechts, des Alters oder der Semesterzahl auf die Bewertung der Vorlesung finden (vgl. 2.1.4.3). Der Faktor *Popularität* hängt stark mit vielen anderen Skalen zusammen, insbesondere mit solchen, die den Dozenten beschreiben ($r > .60$ bei *Lehrkompetenz* und *Dozentenengagement*; ebd., S. 144). Über die Richtung des Zusammenhangs ist damit allerdings noch keine Aussage möglich. Die Arbeitshaltung (Skala *Fleiß*) der Studierenden korreliert zwar nicht mit anderen Skalen des HILVE, ist jedoch eine „wichtige Determinante des individuellen Lernerfolgs“ (Rindermann, 2001, S. 70).

Das HILVE ist für einen mittleren Differenzierungsgrad entwickelt worden (Rindermann, 1996a). Es soll aber erlauben, durch optionale zusätzliche Items und offene Fragen das Inventar an die Erfordernisse verschiedener Veranstaltungen anzupassen.

2.1.1.3 Trierer Inventar zur Lehrveranstaltungsevaluation (TRIL)

Gollwitzer und Schlotz (2003) weisen auf ein Dilemma standardisierter Inventare zur Beurteilung von Lehrveranstaltungen hin: Durch breite Konstrukte ließen sich zwar verschiedene Veranstaltungen und Veranstaltungstypen (Seminare, Vorlesungen oder Tutorien) mit dem gleichen Inventar beurteilen und damit auch untereinander vergleichen, gleichzeitig fielen aber „einzelne

Facetten der Rückmeldung“ weg (ebd., S. 117). Gerade diese spezifischen Aspekte seien als konkretes, handlungsnahes Feedback an den Dozenten zur Verbesserung der Lehre aber wichtig. Eine Arbeitsgruppe an der Universität Trier entwickelte daher ein Instrument zur Lehrveranstaltungsevaluation, das folgende Kriterien erfüllen sollte:

1. Das Erfassen breiter Indikatoren für Qualität von Lehrveranstaltungen,
2. das Verwenden von Items, die eine konkrete, handlungsnahe Rückmeldung an den Dozenten erlauben,
3. einen breiten Einsatzbereich, der die Verwendung des Fragebogens in verschiedenen Veranstaltungstypen erlaubt,
4. Ökonomie und einfache Handhabung.

Zur Auswahl passender Items wurden bestehende Fragebögen durchgesehen und auf ihre konzeptuellen Gemeinsamkeiten untersucht. Daneben wurden solche Items ausgewählt, die Informationen „über einen bestimmten, für die konkrete Verbesserung der Lehre zielführenden Aspekt der Qualität einer Lehrveranstaltung“ liefern sollten (Gollwitzer & Schlotz, 2003, S. 118).

Die ausgewählten Items waren Aussagen zur Lehrveranstaltung, die zum Teil als allgemeine Aussagen, zum Teil in der Ich-Form formuliert waren. Ein mehrstufiger Prozess der Item- und Skalenanalyse führte schließlich zum *Trierer Inventar zur Lehrveranstaltungsevaluation* (TRIL), das Items auf sechs Dimensionen enthält (siehe auch Anhang D):

- *Struktur und Didaktik* (sechs Items),
- *Anregung und Motivation* (fünf Items),
- *Interaktion und Kommunikation* (vier Items),
- *Anwendungsbezug* (vier Items),
- *Persönlicher Gewinn durch die Veranstaltung* (vier Items)
- sowie eine *Gesamtbeurteilung* (drei Items).

Zusätzlich erfragen vier Items die Bewertung der Qualität studentischer Referate, falls die Veranstaltung Raum dafür bietet. Dazu kommen spezifische Items, die besondere Aspekte einzelner Veranstaltungstypen erfassen und sich darum keiner der breiteren Skalen zuordnen lassen. Unter Auslassung der sechsten Skala (*Gesamtbeurteilung*) ergaben Faktorenanalysen eine Varianzaufklärung von 48%; sowohl bei Berechnung von Varimax- als auch von Oblimin-Rotationen luden alle Items eindeutig auf den fünf Faktoren. Dabei ergab die Oblimin-Rotation eine bessere Anpassung an die Einfachstruktur, da die Faktoren „zum Teil hoch miteinander kovariierten“ (Gollwitzer & Schlotz, 2003, S. 122).

Diese Kovariation könnte die Frage aufwerfen, ob das Inventar statt verschiedener Skalen tatsächlich nur einen einzigen, globalen Faktor misst (was manche Autoren der studentischen Evaluation im Allgemeinen vorwerfen, siehe Tabelle 3 auf S. 16; Greenwald, 1997). Gollwitzer und Schlotz (2003) haben deshalb die Modellpassung dreier alternativer Modelle verglichen: Modell 1 beschreibt einen latenten globalen Faktor, Modell 2 fünf korrelierte Faktoren und Modell 3 fünf Faktoren erster Ordnung und einen Faktor zweiter Ordnung:

> Ein Vergleich der Anpassungsmaße [...] zeigt, dass das zweite Modell [...] die beste Anpassung an die Daten hat. Modell 3 [...] führt zu einer geringfügig schlechteren Anpassung. Modell 1 [...] ist in jedem Fall zu verwerfen. [...] Die Analysen zeigen, dass man davon ausgehen kann, dass es sich bei den im TRIL erfassten Skalen um unterschiedliche, aber nicht unabhängige Facetten der Veranstaltungsqualität handelt. (S. 124)

2.1.2 Allgemeine Gütekriterien studentischer Evaluation

Über die Diskussion der einzelnen Erhebungsinstrumente hinaus gibt es zahlreiche Studien zu den verschiedenen Gütekriterien der studentischen Lehrveranstaltungsevaluation. Von den Hauptgütekriterien werden vor allem die Reliabilität (2.1.3) und Validität (2.1.4) diskutiert – oder ihre mögliche Gefährdung durch Verzerrungen. Die Objektivität wird indirekt mitberücksichtigt, da man die errechneten Reliabilitätskennwerte auch als Indikatoren für Objektivität verstehen kann (Rindermann, 1998).

2.1.3 Reliabilität

In der Literatur wird die Reliabilität üblicherweise als interne Konsistenz (2.1.3.1), zeitliche Stabilität (2.1.3.2) und Generalisierbarkeit (2.1.3.3) behandelt.

2.1.3.1 Konsistenz

Ein Maß für die Beobachterübereinstimmung ist die *Interraterreliabilität* als Reliabilität des gemittelten Urteils (vgl. Rindermann, 2001, S. 119). Sie gibt an, wie hoch die Einschätzungen einer Stichprobe mit den Einschätzungen einer anderen Stichprobe aus derselben Population korrelieren würde – und nimmt mit steigender Zahl der Beobachter zu[2]. Rindermann (1996a) gibt als Untergrenze für ausreichend reliable Schätzungen eine Stichprobengröße von $n = 15$ an, Cashin

[2] Z.B. fand Marsh (1982), dass die durchschnittliche (Halbierungs-) Reliabilität des SEEQ bei .74 liegt, wenn sie auf der Einschätzung von nur zehn oder weniger Studenten beruht, aber auf .90 steigt, sobald 25 Antworten vorliegen.

(1995) empfiehlt als Daumenregel, Items mit einer Reliabilität kleiner .70 nur mit besonderer Vorsicht zu interpretieren. Skalen erweisen sich als reliabler als Rohwerte der einzelnen Items, da sich individuelle Differenzen ausmitteln (Rindermann, 1998).

Neben der Interraterreliabilität lassen sich auch *individuelle Übereinstimmungen* zwischen Studenten berechnen, die naturgemäß erheblich niedriger liegen als die Übereinstimmung der gemittelten Ratings (March & Roche, 1997). Rindermann (2001, S. 125) weist darauf hin, dass diese Übereinstimmungen „den zum Vergleich herangezogenen Werten [entsprechen], die für Gutachter-Übereinstimmungen wissenschaftlicher Publikationen erzielt werden“[3]. Individuelle Übereinstimmung und Interraterreliabilität kann man mit Hilfe der Spearman-Brown-Formel (Fisseni, 1997, S. 81ff.) ineinander überführen.

2.1.3.2 Stabilität

Die Interraterreliabilität ist über ein Semester hinweg sehr stabil (Cashin, 1995), ebenso wie die individuellen Übereinstimmungen. Das bedeutet, dass die Gespräche zwischen den Studenten über die Veranstaltung sich kaum auf die Beobachterübereinstimmung auswirken (Rindermann, 2001, S. 125).

2.1.3.3 Generalisierbarkeit

Wie genau bilden die studentischen Beurteilungen einer einzelnen Veranstaltung die allgemeine Lehreffektivität eines Dozenten ab? Marsh (1982) fand, dass vor allem der Dozent und nicht die

Tabelle 2: Mittlere Korrelationskoeffizienten zwischen den Beurteilungen verschiedener Veranstaltungen nach Rindermann (2001): Ein Dozent erhält in verschiedenen Veranstaltungen ähnliche Ratings.

Dozent	**Skalen**	**Gleiche Veranstaltung** (in einem anderen Semester)	**Andere Veranstaltung**
Gleicher Dozent	Lehrverhaltensskalen	.71	**.52**
	Lehreffektivitätsskalen	.61	**.31**
	studentische Skalen	.67	**.25**
	Rahmenbedingungen	.49	**.31**
Anderer Dozent	Lehrverhaltensskalen	.15	.04
	Lehreffektivitätsskalen	.16	.08
	studentische Skalen	.19	.09
	Rahmenbedingungen	.20	.10

[3]Er räumt jedoch ein, dass manche Zeitschriften bewusst Gutachter berufen, die verschiedene Standpunkte vertreten, um die Breite des Faches zu repräsentieren.

Veranstaltung die Beurteilung bestimmt. Neuere Studien (Rindermann, 2001, S. 157, vgl. Tabelle 2) bestätigen diese Ergebnisse eindrucksvoll, obwohl sie eben auch verdeutlichen, dass neben Dozentenmerkmalen auch studentische Charakteristika, Rahmenbedingungen und Wechselwirkungen zwischen diesen Größen bei der Beurteilung der Lehre eine Rolle spielen.

Aus diesem Grund weisen die meisten Autoren (z. B. Rindermann, 2001; Cashin, 1995) darauf hin, dass immer mehrere Veranstaltungen eines Dozenten berücksichtigt werden müssen, bevor man verlässliche Aussagen zum Lehrverhalten dieses Dozenten treffen kann. Das ist vor allem dann wichtig, wenn die Evaluationsergebnisse bei Personalentscheidungen verwendet werden, wie es etwa in den Vereinigten Staaten gehandhabt wird. Cashin (1995) empfiehlt:

> For most instructors [...], use ratings from a variety of courses, for two or more courses from every term for at least two years, totaling at least five courses. If there are fewer than fifteen raters in any of the classes, data from additional classes are recommended. (S. 2)

2.1.4 Validität

Verfahren zur studentischen Lehrevaluation müssen zwei Aspekten der Validität genügen (Rindermann, 2001, S. 163f.): Das Instrument muss

1. die Meinung der Studierenden ohne Verfälschung wiedergeben und
2. die Lehrveranstaltung genau beschreiben (beziehungsweise die einzelnen Teilaspekte der Lehreffektivität erfassen).

Punkt 1 kann man durch flexible, an die Veranstaltung angepasste Fragebögen einschließlich offener Fragen erreichen. Beim zweiten Punkt stößt man schnell auf das Problem, dass es kein allgemein anerkanntes externes Kriterium gibt, anhand dessen man die Validität der studentischen Ratings und damit ihrer Einschätzung der Lehreffektivität überprüfen könnte. Eine Alternative besteht darin, verschiedene Datenquellen zu verwenden, um diese Frage abschätzen zu können (Cashin, 1995). Häufig werden dazu der Lernerfolg der Studenten nach dem Besuch der Veranstaltung (2.1.4.1) sowie Selbstbeurteilungen der Dozenten oder Fremdeinschätzungen (2.1.4.2) verwendet. Für diese Studie ist der letzte Punkt weniger wichtig und wird daher nur kurz dargestellt. Zahlreiche Untersuchungen beschäftigen sich umgekehrt auch mit Biasvariablen, welche die Validität gefährden könnten (2.1.4.3). Tabelle 3 verdeutlicht in einer Übersicht einige grundsätzliche Positionen zur Frage der Validität studentischer Urteile.

Tabelle 3: Zentrale Positionen verschiedener Forscher zu Konzepten und Gütekriterien der studentischen Lehrevaluation nach Greenwald (1997; Übers. d. Verf.). Im deutschsprachigen Raum deckt sich die Ansicht von Rindermann (z.B. 1996b; 2001) weitgehend mit der von Marsh und Roche (1997).

	Validitätsbedenken und zentrale Fragen			
Autoren	**konzeptuelle Struktur: Sind Bewertungen konzeptuell ein- oder mehrdimensional?**	**konvergente Validität: Wie hoch korrelieren die Bewertungen mit anderen Indikatoren für effektive Lehre?**	**diskriminante Validität: Gibt es Verzerrungen durch Variablen, die nicht mit effektiver Lehre zusammenhängen?**	**daraus folgernde Validität: Werden die Bewertungen in einer Weise verwendet, die dem Bildungssystem nutzt?**
Marsh & Roche (1997)	Wie effektive Lehre sind die Bewertungen konzeptuell und empirisch multidimensional. Ihre Validität und besonders ihr Nutzen als Rückmeldung nehmen ab, wenn man diese Multidimensionalität nicht beachtet.	Verschiedene Dimensionen der studentischen Bewertungen weisen einen konsistenten Zusammenhang mit Kriterien effektiver Lehre auf, mit denen sie logischerweise in Verbindung stehen sollten. Das belegt ihre Validität.	Die Bewertungen sind verhältnismäßig unbeeinflusst von möglichen Verzerrungen. Die (Fehl-) Interpretation von Biasvariablen beachten typischerweise nicht die legitimen Einflüsse auf die Lehre (z.B. Kursstärke), welche die Bewertungen genau widerspiegeln.	Multidimensionale Ratings in Kombination mit Weiterbildung verbessern die Effektivität der Lehre (ihr wichtigster Zweck). Ihre Anwendung bei Personalentscheidungen sollte vorsichtig und systematisch geschehen.
d'Apollonia & Abrami (1997)	Obwohl Lehre multidimensional ist, enthalten die Bewertungen einen starken globalen Faktor. Dieser besteht aus mehreren hoch korrelierten Faktoren niedrigerer Ordnung.	Globale studentische Bewertungen oder ein gewichteter Mittelwert spezifischer Ratings korrelieren nur mäßig mit dem durch den Dozenten hervorgerufenen Lernen der Studenten.	Es gibt nur wenig Belege für Verzerrungen der Bewertungen; wenige Merkmale haben einen differenziellen Einfluss auf Bewertungen und vom Dozenten hervorgerufenen Lernen der Studenten.	Bewertungen liefern valide Informationen über die Effektivität eines Dozenten. Trotzdem sollten sie weder die einzige Informationsquelle darstellen, noch sollte man sie überbewerten.
Greenwald & Gillmore (1997)	Da studentische Bewertungen durch einen global Faktor dominiert werden, erfassen viele Items vor allem diesen globalen Faktor und nicht ihren charakteristischen Inhalt.	Bewertungsskalen korrelieren mäßig mit Leistung in multisektionalen Designs.	Derselbe Dozent erhält bessere Bewertungen, wenn er bessere Noten vergibt oder wenn die Kurse kleiner sind. Frühere Forschung zeigt auch, dass die Bewertungen bei einem enthusiastischen Lehrstil besser sind.	Das Streben nach besseren Bewertungen verursacht einen subtilen Mildeeffekt bei den Dozenten, der sowohl a) den akademischen Inhalt verringern und b) die Zahl guter Noten aufblähen kann.
McKeachie (1997)	Es gibt einen g-Faktor in den Bewertungen, aber trotzdem unterscheidbare Faktoren niedrigerer Ordnung.	Studentische Bewertungen liefern valide, wenn auch nicht perfekte, Maße effektiver Lehre.	Einflüsse von Bias-Variablen (statt effektiver Lehre) auf die Ratings sind ein Problem bei der bedauernswerten Praxis, Mittelwerte der Ratings mit Normen zu vergleichen.	Studentische Bewertungen tragen zu Urteilen über die Effektivität von Lehre bei, aber ihre Anwendung könnte verbessert werden.

2.1.4.1 Lernerfolg

Vor allem in den USA wird Lernerfolg als die entscheidende Erfolgsvariable effektiver Lehre verwendet (Cashin, 1995). Häufig wird Lernerfolg über die spätere Prüfungsleistung der Studenten operationalisiert, obwohl der Zusammenhang zwischen der Lehreffektivität und diesem Kriterium durch mehrere Variablen moderiert wird (ausführlich dazu Rindermann, 2001, S. 164). McKeachie (1997) weist zudem darauf hin, dass es noch weitere Zielvariablen guter Lehre gibt, etwa Problemlösefähigkeiten oder persönliche Entwicklung, die nicht unbedingt mit dem Prüfungsergebnis korrelieren.

2.1.4.2 Fremd- und Selbsteinschätzung des Dozenten

Gegen die Selbsteinschätzung des Dozenten gibt es zahlreiche Einwände, die von Verzerrungen durch soziale Erwünschtheit bis zu statistischen Argumenten reichen (geringere Reliabilität aufgrund einer einzelnen Einschätzung; Rindermann, 2001, S. 165). Dennoch gibt es Studien, die mittlere Korrelationen zwischen Studenten- und Selbstratings von Dozenten gefunden haben (vgl. Cashin, 1995). Zudem können Selbsteinschätzungen sehr sinnvoll für Interventionen zur Verbesserung der Lehre sein (Marsh & Roche, 1997).

Die Übereinstimmung zwischen den studentischen Urteilen und Fremdratings durch Kollegen oder geschulte Beobachter liegen ebenfalls in mittlerer Höhe (z. B. $r = .54$ bei Rindermann, 2001, S. 166). Cashin (1995, S. 3) kommt daher zu dem Schluss: “If one is willing to grant that the ratings of administrators, colleagues, alumni, and others have some validity [...] then student ratings share that validity”.

2.1.4.3 Die „Hexenjagd nach Biasvariablen“

Marsh (1987, S. 305) spricht von einer „Hexenjagd auf Biasvariablen“ (“witch hunt for potential biases in students' evaluations”), und auch wenn man sich dieser Metapher nicht anschließen möchte, gibt es außerordentlich viele empirische Arbeiten, die studentische Ratings auf Verzerrungen untersuchen. In der Regel versteht man unter *Biasvariablen* solche Faktoren, die unabhängig von der Lehreffektivität die Beurteilungen der Studenten kausal beeinflussen (Marsh, 1987, S. 263). Tabelle 4 stellt die wichtigsten Ergebnisse vor: Einige der untersuchten Faktoren korrelieren mit dem studentischen Urteil, sind allerdings nicht automatisch als Bias zu werten, da

die Art des Zusammenhanges umstritten ist; für andere wurden keine substanziellen Zusammenhänge festgestellt.

Als *Dr.-Fox-Effekt* (Marsh, 1987, S. 331ff.; Rindermann, 2001, S. 184) wird das Phänomen bezeichnet, dass ein engagierter und enthusiastischer Dozent unter Laborbedingungen bessere Beurteilungen erhält, auch wenn seine Vorträge inhaltlich schwach sind[4]. In realen Veranstaltungen tritt der Effekt jedoch nicht oder nur minimal auf. Zudem sind vor allem Skalen betroffen, die das Engagement des Dozenten erfassen. Einige Autoren (Cashin, 1995; Marsh, 1987) argumentieren deshalb, dass es sich um einen gültigen Effekt und nicht um einen Bias handele.

Der *Mildeeffekt* (*grading leniency*, Marsh & Roche, 2000, 1999, 1997; Gillmore & Greenwald, 1999; McKeachie, 1997) bezeichnet den Effekt, dass es eine moderate Korrelationen zwischen den erwarteten Noten und der Beurteilung des Kurses gibt. Das könnte man mit verschiedenen Hypothesen erklären: durch verfälschte Urteile der Studenten in Kursen mit besseren Noten (*Mildehypothese*), die Annahme, von vorneherein interessierte Studenten lernten mehr in einem Kurs (*Eigenschaftshypothese*) oder den Umstand, dass Studenten, die mehr lernen, auch die Veranstaltung besser beurteilen (*Validitätshypothese*). Marsh und Roche (2000) betonen, dass sich der Effekt nur auf einigen Skalen des SEEQ zeige und zudem gering sei (höchstens etwa r=.20, je nach Skala). Während die Autoren die Validitäts- und Eigenschaftshypothesen bestätigt sehen, lehnen sie die Mildehypothese für ihre eigenen Studien ab, auch wenn sie generell nicht ausgeschlossen werden kann.

Arbeitsaufwand und *Schwierigkeit* korrelieren zwar mit der Beurteilung der Veranstaltung, allerings erhalten schwerere Kurse eher bessere Beurteilungen, zu leichte dagegen schlechtere: “SETs were lower in ‘Mickey Mouse’ courses” (Marsh & Roche, 1997, S. 1191; Roche & Marsh, 1998).

Der von den Studenten angegebene *Besuchsgrund* einer Veranstaltung hängt ebenfalls mit der Bewertung der Veranstaltung zusammen. Veranstaltungen, die aus Interesse besucht werden, erhalten eine bessere Beurteilung als Pflichtveranstaltungen (Rindermann, 1994). Sobald man die Beurteilungen von zwei Veranstaltungen vergleichen möchte, sollte man das berücksichtigen. Das gilt im übrigen auch für die anderen in Tabelle 4 als kritisch aufgeführten Faktoren – solange man die studentische Lehrveranstaltungsevaluation aber nicht für Personalentscheidungen einsetzt, sondern Dozenten ihre Lehre durch Rückmeldung der Studenten verbessern wollen, stellen sie kaum ein Problem dar.

[4] In der ursprünglichen Studie verkörperte ein Schauspieler einen Dozenten names „Dr. Fox“.

Tabelle 4: Ergebnisse zu verschiedenen Klassen von Biasvariablen, Kategorisierung erweitert nach Cashin (1995), der die Ergebnisse verschiedener Metaanalysen zusammengestellt hat (z. B. Centra, 1993); vgl. auch Marsh & Roche (1997), die Literaturübersicht von Rindermann (2001, S. 181ff.) sowie el Hage (1996).

	Klassen von möglichen Biasvariablen			
	Dozentenvariablen	**Studentische Variablen**	**Veranstaltungs-variablen**	**Durchführungs-variablen**
Kontrolle nicht erforderlich	• *Alter* und *Lehrerfahrung* (Marsh & Hocevar, 1991) • *Geschlecht* d. Doz. • *Rasse* des Dozenten • *Persönlichkeit:* Einfluss von *positivem Selbstwert* u. *Enthusiasmus*, aber kein Bias (Cashin, 1995) • Produktivität in der *Forschung*	• *Alter* • *Geschlecht* • *Studiendauer* • *GPA-Note* • *Persönlichkeit*	• *Kursgröße*: Tendenziell erhalten Dozenten kleinerer Kurse bessere Noten, allerdings ist der Effekt minimal und erwartungskonform (Marsh & Roche, 1997) • *Tageszeit*, zu der die Veranstaltung stattfindet	• *Zeitpunkt* während des Semesters, zu dem die Befragung durchgeführt wird
Kontrolle vielleicht erforderlich	• *Beschäftigungsverhältnis:* Dozenten erhalten bessere Beurteilungen als Tutoren etc.; das scheint ein Effekt, kein Bias zu sein • *Ausdruckskraft*: Dr.-Fox-Effekt, s.o.; wirkt vor allem auf Enthusiasmus-Skalen, aber auch das ist ein Effekt, kein Bias • *Sympathie/Popularität:* mittlere bis hohe Korrelationen (v.a. mit Dozentenvariablen); wahrscheinlicher bei hoher Interaktionen mit Studierenden; Richtung des Zusammenhangs unklar	• *Motivation*: *Vorinteresse* für das Thema führt zu günstigeren Ratings, Pflichtveranstaltungen zu schlechteren (*Besuchsgrund*); muss kontrolliert werden! • *Erwartete Noten:* Geringe, positive Korrelationen zwischen Rating und Note müssen nicht auf Bias zurückgehen (Marsh & Roche, 1997; im deutschsprachigen Raum weniger beachtet)	• *Semester*, in dem die Veranstaltung angeboten wird: Im Hauptstudium geringfügig besser beurteilt als im Grundstudium • *Akademisches Feld*: Künstlerische Fächer werden besser als Sozialwissenschaften, diese wiederum besser als mathematische Fächer beurteilt • *Arbeitsaufwand / Schwierigkeit*: Entgegen häufiger Annahme positiver Zusammenhang, d.h. schwere Kurse werden besser beurteilt	• *nicht-anonyme* Beurteilungen: Unterschriebene Beurteilungen sind besser • *Anwesenheit* des Dozenten während der Befragung: Beurteilungen sind besser • *Zweck* der Beurteilung: Wenn die Ergebnisse für Personalentscheidungen verwendet werden, sind Ratings in der Regel milder

Wenig beachtet wurden bislang die momentane *Stimmung und Befindlichkeit* der Studierenden als Ursache von verzerrten Beurteilungen. Allerdings werden Urteile nicht automatisch verfälscht, wenn man Stimmung und Befindlichkeit als Informationen in den Urteilsprozess mit einbezieht (Clore, Wyer, Dienes, Gasper, Gohm & Isbell, 2001):

> According to the information principle, emotional feelings are a conscious representation of unconscious appraisal processes. Thus any subjective experience that arise while we consider a decision [...] may accurately index our evaluation. There is, therefore, no reason to assume as is so often done in the judgment

> and decision literature that decisions based on feelings must be biased. The only issue is whether [...] judges [...] focus on task-relevant goals and standards rather than purely personal goals. (S. 39)

Beim Einfluss von Stimmungen und Gefühlen auf Bewertungen spielt zudem die Attribution dieser Gefühle auf den Gegenstand der Bewertung eine wichtige Rolle. Wenn der Gegenstand, der bewertet werden soll, nicht als Ursache der Gefühle interpretiert wird, haben Stimmungen keinen Einfluss auf die Bewertung (ebd., S. 41). Die Stimmung der Studierenden sollte also nur dann einen Einfluss auf ihre Bewertung einer Lehrveranstaltung haben, wenn diese Veranstaltung als Ursache der Stimmung interpretiert wird. In diesem Fall dürfte man die Stimmung aber nicht als Biasvariable verstehen.

Zudem sollten Stimmung und Befindlichkeit nur in bestimmten Situationen überhaupt einen Einfluss auf den Urteilsprozess (Forgas, 2001) haben. Neben der Situation spielen weitere Variablen wie die persönliche Motivation und Valenz der Bewertung, die Vertrautheit der Aufgabe oder die verfügbare kognitive Kapazität eine Rolle für die Frage, welche Verarbeitungsstrategie angewendet wird.

2.1.5 Fazit

Die Lehrveranstaltungsevaluation durch Studenten hat zu hitzigen Debatten geführt (Kromrey, 1996a, 1996b, 1994; Diehl, 1996; Giesen, 1994), in deren Verlauf die Gütekriterien ausführlich beleuchtet wurden. Cashin (1995, S. 6) kommt dabei zu folgendem Fazit: “In general, student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control; probably more so than any other data used for evaluation”.

Trotzdem flammt die Diskussion über die studentische Lehrveranstaltungsevaluation immer wieder auf. Bei ihrer Durchführung im Internet könnte sich erneut die Frage nach der Validität stellen.

2.2 Experimente und Datenerhebung im Internet

Computer spielen seit mehreren Jahrzehnten eine große Rolle in der Diagnostik und der experimentellen Forschung (Hänsgen, 1999; Zentrum für Psychologische Information und Dokumentation [ZPID], 2003). Seit Mitte der neunziger Jahre wird darüber hinaus die Datenerhebung im Internet immer beliebter, wobei der kognitiven Psychologie eine Vorreiterrolle zukommt (Musch & Reips, 2000). Neben spezifischen Problemen (Abschnitt 2.2.1) bietet das neue Medium auch

viele Vorteile (2.2.2). Batinic (2001, S.12) weist auf eine Reihe von besonderen Merkmalen von Fragebogenuntersuchung im Internet hin. Die wichtigsten sind *Asynchronität* und *Alokalität,* d.h. die Probanden können den Zeitpunkt und Ort der Befragung selbst bestimmen.

2.2.1 Schwierigkeiten der Datenerhebung im Internet

Das Äquivalenzproblem (2.2.1.1) ergibt sich aus den Besonderheiten der computerbasierten Diagnostik (2.2.1.2). Die Datenerhebung im Internet hat viele dieser Besonderheiten geerbt, zeichnet sich aber aber auch durch spezifische methodische Fragestellungen aus (2.2.1.3).

2.2.1.1 Das Äquivalenzproblem

Kann man die Ergebnisse von computerbasierten und Papier-und-Bleistift-Testverfahren vergleichen, wenn es diese beiden Varianten eines Verfahrens gibt? Diese Frage bezeichnet man häufig als *Äquivalenzfrage* (Klinck, 1998) oder *Äquivalenzproblem*. Das Testkuratorium der Föderation Deutscher Psychologenvereinigungen forderte schon 1986:

> Um Beeinträchtigungen in der Aussagekraft von Ergebnissen zu vermeiden, ist zu beachten:
> 1. Die Normtabellen bestehender Tests dürfen nicht ungeprüft auf die Ergebnisse bei EDV-gesteuerter Vorgabe übertragen werden; im Regelfall wird eine neue Testeichung erforderlich sein.
> 2. Ergebnisse von Validitätsstudien dürfen nicht ungeprüft von der manuellen auf die EDV-gestützte Testung übertragen werden; auch hier ist eine neue empirische Untersuchung erforderlich.
> 3. Besonderes Gewicht muss bei allen Anwendungen auf die Prüfung der Frage gelegt werden, ob sich gegenüber konventionellen Formen der Testung die Testfairness verändert. (S. 164)

2.2.1.2 Unterschiede zwischen computerbasierten Leistungs- und Persönlichkeitstests

Das Testkuratorium (1986) legt also besonderen Wert auf die Validität, Normierung und Fairness computerbasierter Verfahren. Seit dieser Forderung zur generellen Neu-Normierung computerbasierter Instrumente ist die Entwicklung dieser Verfahren allerdings fortgeschritten. Hänsgen (1999, S. 19) schlägt vor, zwischen „fragebogenartige[n] Verfahren (allgemeine[n] Niveautests ohne Zeitbegrenzung)“ und Persönlichkeitstests einerseits und Leistungstests andererseits zu unterscheiden. Bei den Fragebogen-Verfahren sieht er den Trend, dass „‚[j]üngere‘ Vergleiche [...] eher eine Gleichheit der Normen, [...] ältere eher Unterschiede feststellten“ und begründet das mit der schlechteren Ergonomie älterer Geräte und mit der im Laufe der Jahre abnehmenden Angst vor Computern. Auch Sidiropoulou (1997, S. 406) kommt zu dem Schluss, dass die „psychometrische Äquivalenz bei Persönlichkeitstests [höher ist] als bei Leistungstests“.

Computer- und papierbasierte Versionen von Leistungstests sind in der Regel nicht äquivalent, und zwar umso weniger, je stärker das Testergebnis von der Reaktionszeit abhängt, wie z.B. bei Speedtests (Hänsgen 1999; Klinck, 1998; siehe auch Tabelle 5). Da Menschen am Bildschirm langsamer lesen, ist die Reaktionszeit am Bildschirm länger; meist werden auch weniger Items pro Bildschirmseite dargeboten als in der Papierform. Solche Bedingungsänderungen können nicht nur zu Mittelwertsunterschieden und unterschiedlichen Streuungen, sondern auch zu (meist problematischeren) Rangplatzänderungen der Versuchspersonen beim Einsatz verschiedener Medien führen, die nicht einfach durch neue Normen aufgefangen werden können.

Tabelle 5: Nicht-minderungskorrigierte Korrelationen zwischen Papier- und Computerversionen verschiedener Batterien von Leistungstests, getrennt nach Fähigkeits- und Speedtests (vereinfacht nach Mead & Drasgow, 1993). Papier- und Computerversionen von *Fähigkeitstests* liefern eindeutig ähnlichere Ergebnisse als *Speedtests*.

	Fähigkeitstests		**Speedtests**	
Testbatterie	***n***	***r***	***n***	***r***
Keine / verschiedene	16	.91	8	.66
Multidimensional Aptitude Battery (MAB)	5	.90		
Differential Aptitude Test (DAT)	42	.90	6	.34
Armed Services Vocational Aptitude Battery (ASVAB)	60	1.00	22	.79

2.2.1.3 Fragebogenerhebungen im Internet

Fragebogenartige Verfahren (Umfragen bzw. Persönlichkeitstests) spielen bei der Datenerhebung im Internet (als spezieller Variante der computerbasierten Verfahren) eine wesentlich größere Rolle als Leistungstests und Experimente im Internet (Polkehn & Wandke, 1999), wie auch ein Blick auf die Linksammlung der ZPID (2003) belegt. Bei den wenigen verfügbaren Leistungstests handelt es sich meist um Niveautests, da hier insbesondere das Problem der Latenzzeiten bei der Datenübertragung über das Internet (s.u.) weniger ins Gewicht fällt. Häufig stellt sich die Frage nach der Äquivalenz bei vielen Untersuchungen im Internet allerdings nicht, weil es gar kein papierbasiertes Pendant dazu gibt. Dennoch muss man im Vergleich zu klassischen Laborexperimenten bei öffentlich zugänglichen Experimenten im Internet einige Besonderheiten beachten (Reips, 2000):

- Die *experimentellen Kontrolle* kann sich verringern durch mehrfache Teilnahme, nicht steuerbare Situationseinflüße während der Teilnahme am Experiment und durch die Varianz der technischen Ausstattung. Dazu zählen auch Reaktionszeitdifferenzen durch unterschiedliche

Messverfahren, Prozessorgeschwindigkeiten oder Verzögerungen bei der Datenübertragung. Diese Varianz kann andererseits systematische Effekte der Laborgeräte durch stärkere Randomisierung verhindern.

- *Drop-Outs* werden besonders bei öffentlichen Web-Experimenten ohne (finanzielle) Anreize zum Problem. Bei Untersuchungsdesigns mit mehreren Gruppen, z. B. unterschiedlichen Schwierigkeitsgraden, kann man die Abbrecherquote auch als abhängige Variable auswerten (dazu ausführlicher Knapp & Heidingsfelder, 1999).

Ein weiterer wichtiger Aspekt ist die *Selbstselektion* der Teilnehmer (Bosnjak, 2003, S. 57): Welcher Personenkreis innerhalb der Gesamtpopulation interessiert sich für ein bestimmtes Thema[5]? Geht das Interesse so weit, an der Untersuchung teilzunehmen? Die wichtigste Gründe, an einer E-Mail-Befragung teilzunehmen, sind nach Bosnjak und Batinic (1999) *Neugier*, der Wunsch einen *Beitrag für die Forschung* zu leisten, *Selbsterkenntnis* und an letzter Stelle *materielle Anreize*.

Obwohl die Analyse von Beteiligungsquoten bei Internet-Befragungen in der Praxis häufig auf Fallstudien ohne einen psychologischen Hintergrund beruht (Theobald, 2003), kann man verschiedene psychologische Theorien darauf anwenden (Bosnjak, 2002, Kap. 3.2.2). Praktische Hinweise für die Gestaltung von Fragebögen liefert die *Tailored Design Method* von Dillman (2000), der eine Befragung als sozialen Austauschprozess auffasst: “People are seen as more likely to complete and return self-administered questionnaires if they trust that the rewards of doing so will, in the long run, outweigh the costs they expect to incur” (ebd., S. 26). Als Belohnung können z. B. Wertschätzung und Dank für das Ausfüllen des Fragebogens dienen. Obwohl der Ansatz von Dillman Anregungen für die Praxis liefern kann, ist er theoretisch bislang unzureichend belegt.

Couper (2000; vgl. auch Schäffer, 1996) unterscheidet detailliert verschiedene Arten von Fehlern, welche die Qualität einer Datenerhebung per Fragebogen im Internet beeinflussen können. Grundsätzlich ergeben sich diese Schwierigkeiten auch bei papierbasierten Befragungen, aufgrund der geringeren experimentellen Kontrolle und der Drop-Outs aber wesentlich stärker bei Online-Befragungen:

[5] Reips (2000) schlägt vor, zur Abschätzung der bei der Selektion wirksamen Faktoren die verweisenden Seiten auszuwerten (*Referrer*). Solange sich die Daten von Teilnehmern nicht unterscheiden, die über Links auf thematisch unterschiedlichen Seiten das Experiment gefunden haben, könne man den Effekt der Selbstselektion vernachlässigen.

1. Als *Abdeckungsfehler* (*coverage error*) bezeichnet er systematische Verzerrungen durch Unterschiede zwischen der *Zielpopulation*, über die man Aussagen treffen möchte, und der *Rahmenpopulation*, die man mit dem Fragebogen tatsächlich erreichen kann.

2. Dieses Problem kann noch durch den *Stichprobenfehler* (*sampling error*) verstärkt werden, denn in der Regel können nicht alle Mitglieder der Rahmenpopulation an der Datenerhebung tatsächlich teilnehmen.

3. Der *Nichtbeantwortungsfehler* (*nonresponse error*) kann dadurch entstehen, dass ein Teil der Rahmenpopulation trotz optimaler Voraussetzungen nicht an der Untersuchung teilnimmt. Die Größe der Verzerrung durch diese fehlende Beantwortung hängt davon ab,

   - wie groß die Zahl der Befragten ist, die nicht antwortet und
   - wie sehr sich antwortende und nicht antwortende Personen in den untersuchten Merkmalen unterscheiden.

   Falls die Rahmenpopulation unbekannt ist, kann man Abdeckungsfehler und Nichtbeantwortungsfehler kaum unterscheiden. Bei offenen Befragungen im Internet kann man die Zahl der Personen, die sich eine Seite anschauen ohne zu antworten, meist nur schätzen.

Schließlich führt auch der aus der Statistik bekannte *Messfehler* (*measurement error*) zur Abweichung der Antworten vom wahren Wert. Während für diese Art des Fehlers verschiedene Messmodelle existieren, gibt es sie für die zuvor genannten Fehlerarten nicht.

2.2.2 Vorteile und Chancen

Diesen Schwierigkeiten stehen einige Vorteile gegenüber. Die wichtigsten Argumente für die computergestützte Datenerhebung im Allgemeinen sind (z. B. Sidiropoulou, 1997; Hänsgen, 1999):

- *Leichte Datenerhebung* und deren *automatische Auswertung*. Das Eingeben der Daten von Hand entfällt, die Test- oder Untersuchungsergebnisse sind häufig unmittelbar nach Ende der Datenerhebung verfügbar. Dieser Punkt kann zu einer besseren *Urteilsgüte* im Sinne der statistischen Urteilsbildung führen.
- Durch *höhere Objektivität* (z. B. aufgrund der geringeren Versuchleitervarianz oder einer automatischen Fehlerprüfung) besteht die Hoffnung auf eine bessere Qualität der gewonnen Daten – diese bessere *Merkmalsgüte* muss aber im Einzelfall empirisch belegt werden.

- *Flexibilität*. Bei der computerbasierten Datenerhebung ist man nicht auf die Texte oder Bilder beschränkt, sondern kann auch auf Animationen, Filme und Geräusche zurückgreifen (Batinic, 2001, S. 13). Diese Vielfalt kann grundsätzlich ebenfalls die *Merkmalsgüte* steigern.
- *Dokumentierbarkeit*. Zusätzlich zu den eigentlichen Antworten kann der Untersuchende auch „Daten über den Untersuchungsprozess (Befragungszeitpunkt, Dauer, Unterbrechungen usw.)“ sammeln, was gesonderte Dokumentationen überflüssig macht (Batinic, 2001, S. 13).
- *Adaptives Testen*. Zusammen mit probabilistischen Messmodellen erlaubt die Testdurchführung am Computer ökonomische, adaptive Tests, die mit konventionellen Tests nur sehr aufwendig durchzuführen sind. Dem Einwand, dass durch das Fehlen subjektiv einfacher Items bei adaptiven Tests die Testängstlichkeit steige (Sidiropoulou, 1997), halten Pitkin und Vispoel (2001) in ihrem Review die Lösung entgegen, den Probanden das Schwierigkeitsniveau selbst wählen zu lassen. Dadurch sinke die Ängstlichkeit, ohne das Testergebnis bedeutsam zu verzerren.
- *Geringe soziale Erwünschtheit* bei Fragebogenuntersuchungen wird manchmal als Vorteil der computerbasierten Datenerhebung genannt. Dagegen folgerten Dwight und Feigelson (2000), dass beide Medien in diesem Punkt äquivalent sind:

  > Although a slight overall effect was found for [Impression Management][6], with IM being slightly slower under computer administration, the effect is likely to have little practical significance. In addition, the effect of computer administration on IM appears to have diminished over time [...]. Computerized testing does not appear to be the holy grail that some had hoped [...], a method for reducing or eliminating those nagging concerns about response distortion [...]. (S. 360)

Allerdings haben Dwight und Feigelson (2000) in ihrer Metaanalyse die Datenerhebung per Internet noch nicht berücksichtigt. Tatsächlich gibt es Hinweise, dass aufgrund der höheren Anonymität hier die soziale Erwünschtheit geringer ist (Schumacher, Hinz, Hessel & Brähler, 2002). Über diese allgemeinen Vorteile hinaus hat die Online-Datenerhebung weitere günstige Eigenschaften (z. B. Reips, 2000):

- *Große Stichproben*. Die offene Datenerhebung im Internet kann eine größere Stichprobe erreichen und dadurch neben einer höheren statistischen Aussagekraft auch eine größere externe Validität erreichen, sofern die in 2.2.1 geschilderten Probleme vermieden werden.

[6] Dwight und Feigelson (2000) schließen sich der Unterscheidung von *sozialer Erwünschtheit* in bewusstes *Impression Management* (IM) und nicht bewusstes *Self-Deceptive Enhancement* (SED) an.

- Hänsgen (1999, S. 49) schlägt vor, zur Lösung des Normenproblems die „Normgewinnung [durch] ein[en] netzgestützte[n] anonymisierte[n] Datenrücklauf von einzelnen Anwendern“ zu organisieren. Dadurch könnten „‚Generationen‘ von Normen“ entstehen, die Längs- und Querschnittsvergleiche erlaubten.

Noch mehr als die computerbasierte ist die internetbasierte Datenerhebung besonders „zeit- und kosten-ökonomisch“ (Batinic, 2001, S. 13). Dabei zeichnen sich Online-Panels[7] gegenüber anderen Verfahren wie offenen Befragungen durch s*chnelle Rücklaufzeiten* und *hohe Ausschöpfungsraten* aus (ebd., S. 76).

2.2.3 Fazit

Die Datenerhebung im Internet ist ein wertvolles Instrument für die experimentelle und Fragebogen-Forschung. In der Äquivalenzfrage sind ähnliche Effekte zu erwarten, wie sie auch in der allgemeinen computerbasierten Diagnostik auftreten: Anders als Leistungstests mit einer starken Geschwindigkeitskomponente liefern fragebogenartige Verfahren und die Leistungstests mit einer hohen Fähigkeitskomponente eher Ergebnisse, die mit der Papierversion vergleichbar sind. Bei offenen Online-Befragungen muss man mit Selektionseffekten rechnen.

Die Forderung des Testkuratoriums (1986) nach der Überprüfung von Testnormen und Validität bei der Übertragung von konventionellen Tests auf den Computer, und damit auch auf das Internet, gilt auch weiterhin. Eine Aufgabe bleibt es, die eigentlichen Einflussgrößen zu identifizieren, die zu Unterschieden zwischen den Medien führen können (Hänsgen, 1999):

> Nach strenger Auffassung gilt eine nachgewiesene Äquivalenz nur für die untersuchte Methode und die verwendete Hard- und Software [...]. Auch in der konventionellen Diagnostik könnte man allerdings über die Beschaffenheit des Ausfüllplatzes, die Art des verwendeten Stiftes, die Tisch- und Stuhlhöhe, der Beleuchtung als Varianzquelle nachdenken und vergleichbar illusorische Forderungen für Äquivalenznachweise aufstellen. (S. 20)

Die Datenerhebungen im Internet kann dieses Problem vermeiden: Die Vielzahl der eingesetzten Systeme kann zur Vermeidung systematischer Effekte einzelner Komponenten des Computersystems führen (Reips, 2000).

[7] Göritz, Reinhold & Batinic (2002) definieren ein Online-Panel als einen “pool of registered persons who have agreed to take part in online studies on a regular basis”.

2.3 Hypothesen

Aus den dargestellten theoretischen Hintergründen lassen sich einige konkrete Hypothesen zum Thema der studentischen Evaluation im Internet ableiten (These 1 bis These 5). Darüber hinaus bietet sich auch die Gelegenheit, eine explorative Fragen zu klären, für die sich nicht direkt Hypothesen ableiten lassen (Frage 6)[8].

Sowohl allgemeine computergestützte und Internetumfragen haben sich als weitgehend äquivalent zu papierbasierten Verfahren erwiesen (2.2.1.2). Durch den Account im Uninetz haben alle Studenten Zugang zur Online-Befragung; systematische Selektionseffekte (vgl. 2.2.1.3) sind daher unwahrscheinlich.

These 1: Es gibt keinen systematischen Effekt des Mediums auf die studentische Beurteilung der Lehrveranstaltung, d.h., die papierbasierte und internetbasierte Datenerhebung unterscheiden sich nicht.

Als Ergänzung dieser globalen Haupt-Hypothese ergeben sich weitere Hypothesen zu möglichen Biasvariablen. Diese Thesen werden im Folgenden erläutert.

Die Rücklaufquote bei offenen Befragungen ist im Internet geringer. Für die studentische Evaluation heißt das:

These 2: Die Teilnehmerzahlen liegen beim Internetfragebogen unter denen der Befragung mit dem Papierfragebogen.

Im Unterschied zu offenen Internet-Befragungen werden bei (Online-) Panels häufig Anreize verwendet; dort ist die Ausschöpfungsrate höher als bei den offenen Befragungen. Einen analogen Effekt darf man bei der studentischen Lehrveranstaltungsevaluation ebenfalls erwarten.

These 3: Die Rücklaufquote ist bei Verwendung eines Anreizes höher.

Der Einfluss der momentanen Befindlichkeit von Probanden auf die Einschätzung der Lehre wurde bislang kaum untersucht (2.1.4.3). Da aber die experimentelle Kontrolle bei der Erhebung im Internet geringer ist als bei der Befragungen am Ende der Vorlesung (im Sinne eines gemeinsamen Treatments unmittelbar vor der Datenerhebung), lautet These 4:

[8] Um Missverständnisse und Uneindeutigkeiten bei der Nummerierung der Hypothesen und der explorativen Frage zu vermeiden, sind beide gemeinsam nummeriert. Damit folgt nach These 5 direkt Frage 6.

These 4: Die Variabilität der Befindlichkeit ist bei der Befragung im Internet höher.

Die studentische Lehrveranstaltungsevaluation hat sich als robust gegenüber verschiedenen Arten von Verzerrungen gezeigt (2.1.4). Obwohl die Befindlichkeit der Studierenden als Biasvariable bislang kaum untersucht wurde, sollte sie ebenso wie die Vorbereitung der Studierenden auf die Veranstaltung keinen Einfluss auf die studentischen Ratings haben (2.1.1.2). Die Sympathie für den Dozenten sollte schwach mit den studentischen Ratings korrelieren, allerdings zeigt sich dieser Effekt stärker in Seminaren als in Vorlesungen (2.1.4.3). Bei der Datenerhebung im Internet könnten zudem zusätzliche Biasvariablen (2.2.1) eine Rolle spielen: Insbesondere eine geringe Vertrautheit mit Computern und dem Internet sollte zu Selektionseffekten führen. Diese Überlegungen führen zu folgender These:

These 5: Die Befindlichkeit und die eigene Vorbereitung hängen nicht mit den studentischen Beurteilungen der Lehre zusammen, die Sympathie für den Dozenten weist nur einen geringen Zusammenhang mit den Beurteilungen auf.

Je weniger vertraut die Studierenden mit Computern sind und je weniger sie das Internet nutzen, desto seltener nehmen sie an Internetbefragungen teil.

Korrelationen studentischer Ratings aus verschiedenen Kursen mit Prüfungsleistungen in diesen Kursen sind häufig ein Versuch, ihre Kriteriumsvalidität nachzuweisen (2.1.4.1). Studenten in besseren Kursen sollten bessere Noten erhalten, man vergleicht hier also *zwischen* verschiedenen Kursen. Zusammenhänge zwischen Prüfungsnote und Beurteilungen *innerhalb* eines Kurses wurden dagegen bislang kaum oder höchstens als Verzerrung betrachtet (*grading leniency*, vgl. 2.1.4.3); eine mögliche Ursache dafür könnten z.B. Unterschiede in der Motivation oder dem Vorwissen der Studierenden sein. Selektionseffekte spielen in der Diskussion um die Validität der studentischen Evaluation von Lehrveranstaltungen nur eine untergeordnete Rolle[9]: Die Frage, welche Personengruppe an einer Untersuchung teilnimmt und inwieweit das Ergebnis der Befragung repräsentativ für die ganze Population ist, beschäftigt viel mehr die Marktforschung. Als Kennzeichen der Repräsentativität könnte aber auch die Prüfungsleistung dienen:

Frage 6: Unterscheiden sich die Teilnehmer an der Evaluation in ihrer Prüfungsnote von denen, die überhaupt nicht oder nicht regelmäßig teilnehmen?

[9]Allenfalls indirekt bei der Störvariablen *Besuchsgrund* (vgl. 2.1.4.3): Unterschiedliches Vorinteresse wird als Ursache für Selektionseffekte gesehen. Kromrey (2001, S. 12) spricht in diesem Zusammenhang von „einer selbstselektiven Teilmenge von Studierenden“.

3 Methoden

Der Klärung dieser Hypothesen diente die Evaluation der beiden Veranstaltungen *Quantitative Methoden B* und *Forschungsmethoden II* während des Sommersemesters 2002 im Fach Psychologie an der Universität Trier. Im Folgenden möchte ich nach einer Beschreibung der Stichprobe (3.1) auf den Versuchsplan (3.2), die verwendeten Fragebögen (3.3) und den Ablauf (3.4) der Untersuchung eingehen. Überlegungen zu Störeinflüssen (3.6), eine Darstellung der Datenaufbereitung (3.5) und Planung zur Auswertung der Daten (3.7) schließen den Methodenteil ab.

3.1 Versuchspersonen

Die Versuchspersonen waren Psychologiestudenten an der Universität Trier, die eine der beiden Veranstaltungen *Quantitative Methoden B* im Grundstudium oder *Forschungsmethoden II* im Hauptstudium besuchten. Insgesamt nahmen 215 Studierende an der Untersuchung teil, davon 169 Frauen und 46 Männer. Tabelle 6 schlüsselt die Beschreibung der Stichprobe nach den besuchten Veranstaltungen auf.

Tabelle 6: Deskriptive Statistiken der Versuchspersonen, aufgeteilt nach der besuchten Veranstaltung.

		Geschlecht		**Alter in Jahren**			**Fachsemester**		
Veranstaltung	***N***	**w**	**m**	***M***	***SD***	**Spanne**	***M***	***SD***	**Spanne**
Quantitative Methoden B	124	102	22	21.8	3.2	19-41	2.0	0	keine
Forschungsmethoden II	91	67	24	24.4	3.0	21-37	7.3	1.5	6-14

Die Versuchspersonen wurden vor Beginn der Untersuchung darüber aufgeklärt, dass sie diese jederzeit ohne Angabe von Gründen abbrechen könnten und nicht in jeder Woche an der Evaluation teilnehmen müssten, obwohl das sehr wünschenswert sei. Zudem wurde eine schriftliche Einverständniserklärung zur Teilnahme eingeholt. Studenten im zweiten Semester erhielten bei regelmäßiger Teilnahme an der Evaluation als Anreiz eine Versuchspersonenstunde am Semesterende.

3.2 Versuchsplan

Man kann zwei Strategien verwenden, um die Hauptfrage der Untersuchung (These 1) zu prüfen: die Äquivalenz der papier- und computerbasierten Lehrveranstaltungsevaluation.

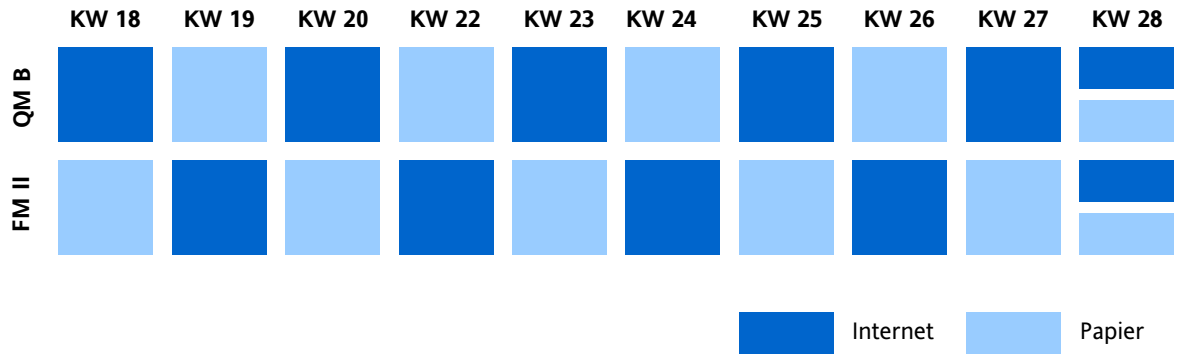


Abbildung 1. Versuchsplan nach Kalenderwochen der beiden Veranstaltungen *Quantitative Methoden* (QM B), *Forschungsmethoden II* (FM II). Hellblaue Felder stellen die Datenerhebung per konventionellem Papier-und-Bleistift-Fragebogen dar, dunkelblaue Felder die Datenerhebung im Internet. In der Kalenderwoche 21 fanden wegen der Pfingstferien keine Veranstaltungen statt.

1. Man bildet aus den Besuchern einer Vorlesung zwei Gruppen mit randomisierter Gruppenzugehörigkeit und läßt die Studierenden in diesen Gruppen entweder einen klassischen Papierfragebogen oder einen Internetfragebogen ausfüllen. Aus versuchsplanerischer Sicht ist diese Strategie die Methode der Wahl, da es sich um eine Zufallszuweisung zu den experimentellen Bedingungen handelt. Das zufällige Aufteilen der Stichprobe und die nachträgliche Kontrolle, ob diese Aufteilung auch tatsächlich eingehalten wurde, sind jedoch sehr aufwendig und fehleranfällig.

2. Jede Woche wird abwechselnd der Internet- oder Papierfragebogen ausgefüllt. Diese Strategie führt zu einem quasi-experimentellen Feldstudien-Design (Hager, 1987), da die Zuweisung der Versuchsperson zu den experimentellen Bedingungen nicht randomisiert stattfindet. Der Nachteil dieser Vorgehensweise besteht darin, dass in den Bewertungen zwei Varianzquellen nicht zu trennen sind, nämlich die Varianz aufgrund der unterschiedlichen Qualität der einzelnen Vorlesungen und die Varianz zu Lasten des Faktors *Medium*. Rindermann und Amelang (1994; vgl. Abschnitt 2.1.3) errechneten allerdings mittlere[10] Retestreliabilitäten für die *Veranstaltungsmittel* (meist über .85). Die Reliabilität der *individuellen Rohwerte* war jedoch nur gering (maximal .75)[11].

   Sofern es einen Effekt zwischen den Vorlesungen gibt, kann man ihn nicht eindeutig auf eine der beiden Ursachen zurückführen. Falls es aber keinen bedeutsamen Unterschied gibt, kann man auch nicht von einem systematischen Einfluss des Mediums sprechen.

[10] Fisseni (1997, S. 124) beurteilt Reliabilitätskoeffizienten von $r_{tt} > .80$ als mittel und $r_{tt} > .90$ als hoch.

[11] Die Autoren verwendeten eine Vorform des HILVE-I. Rindermann (2001, S. 142) konnte dieses Ergebnis durch eigene Studien mit dem HILVE-I und eine umfangreiche Literaturübersicht jedoch bestätigen.

In dieser Studie wurde auf beide Erhebungsstragien zurückgegriffen. Strategie 1 kam beim letzten Erhebungszeitpunkt zum Einsatz, d.h. in der letzten Vorlesungswoche wurde die Stichprobe zufällig geteilt. Dieser Termin bot sich an, da erfahrungsgemäß die Teilnehmerzahlen unmittelbar vor den Semesterferien (und damit der bevorstehenden Klausur) steigen. Somit war die Chance höher, eine ausreichende Teststärke zu erreichen.

Die zweite Strategie wurde über das Semester hinweg verfolgt. Von Vorlesung zu Vorlesung wechselte das Medium zwischen Papier und Internet, wobei in einer Kalenderwoche jeweils unterschiedliche Medien in den beiden betrachteten Veranstaltungen im Haupt- und Grundstudium verwendet wurden. Abbildung 1 zeigt den vollständigen Versuchsplan. Der folgende Abschnitt erläutert, welche Fragebögen bei den einzelnen Terminen zum Einsatz kamen.

3.3 Erhebungsinstrument

Zwei praktische Argumente sprachen dagegen, eines der in Abschnitt 2.1.1 dargestellten Inventare für eine semesterbegleitende Evaluation am Ende der einzelnen Veranstaltungen einzusetzen (vgl. Abschnitt 3.2). Zum einen wurden die Items für eine Evaluation am Semesterende formuliert, zum anderen ist die Itemanzahl vergleichsweise hoch:

- Ein zu langer Fragebogen könnte die Teilnehmer abschrecken und sie von einer wöchentlichen Beantwortung des Fragebogens abhalten. Dadurch könnte sich die experimentelle Mortalität erhöhen (vgl. el Hage, 1996, S. 115).
- Da der papierbasierte Fragebogen wenigstens zum Teil während der Vorlesungszeit ausgefüllt wird, sind die Teilnehmer während dieser Zeit abgelenkt. Ein langer Fragebogen könnte darum ebenfalls die Akzeptanz des Fragebogens bei Studenten und Dozenten verringern.

Aus diesen Gründen kam eine eigene, überarbeite Auswahl von Items zum Einsatz, die Gegenstand von Abschnitt 3.3.1 ist. Im Anschluss daran möchte ich den Abschlussfragebogen vorstellen, der u.a. einige mögliche Biasvariablen erfragen sollte (3.3.2).

Am Ende des Semesters sollten die Ergebnisse des wöchentlichen Fragebogens zur Validierung mit den Ergebnissen des TRIL (vgl. 2.1.1.3) verglichen werden. Dieses bezog sich auf das Semester insgesamt und wurde ebenfalls im Internet erhoben. Da an dieser Befragung allerdings nur ein Bruchteil der Studenten teilnahm, werde ich diese Daten im Folgenden nicht weiter berücksichtigen.

3.3.1 Der wöchentliche Fragebogen

Bei dem Fragebogen, der wöchentlich nach den Vorlesungen eingesetzt werden sollte, wurde besonderer Wert auf die Kürze und Prägnanz des Fragebogens gelegt. Das entspricht der Empfehlung von Cashin (1990, S. 2): “For evaluation, use a few global or summary items or scores. This recommendation is more a personal opinion but such summary, or global, student rating items tend to correlate more highly with student learning than do more specific items“. Rindermann (2001, S. 132.) konnte außerdem zeigen, dass allgemeine Items übereinstimmender und reliabler beurteilt werden als gut beobachtbare Merkmale (z. B. konkrete Verhaltensweisen). Er führt diesen Effekt darauf zurück, dass Personen die „intraindividuelle Variabilität der Wahrnehmung verschiedener Situationen“ zusammenfassen (ebd., S. 128) und so zu einem verlässlicheren Gesamturteil kommen. Das widerspricht der manchmal geäußerten Vermutung, konkrete äußerlich beobachtbare Merkmale würden besser und reliabler eingeschätzt als globale Items (z. B. el Hage, 1996) – unabhängig davon sind solche Items allerdings als Rückmeldung an den Dozenten sehr nützlich und hilfreich (vgl. auch Abschnitt 3.3.2).

Obwohl kein vorhandenes Inventar direkt verwendet werden konnte, waren die gut belegte Multidimensionalität der verschiedenen Inventare und deren Einzelskalen (vgl. Anhang A bis Anhang D) die Ausgangspunkte für die Auswahl der Items. Die Selektion der Items fand in Absprache mit den Dozenten der Lehrveranstaltungen statt. Folgende Items wurden ausgewählt:

- *Wie verständlich waren für Sie die vorgetragenen Inhalte insgesamt?* Dieses Item lehnt sich an Item 9 des SEEQ an, das dort dem Faktor *Organisation* zugeordnet wird.
- *Wie verständlich waren die verwendeten Lernmittel (z.B. Folien, Präsentationen, Beispiele)?* Diese Frage ähnelt Item 5 des TRIL (Faktor *Struktur und Organisation*) und Item 10 aus dem SEEQ (ebenfalls Faktor *Organisation*).
- *Wie schätzen Sie das Tempo ein, das der Vermittlung zugrunde lag?* Diese Frage verbindet die Items 19 aus dem HILVE-I (Skala *Überforderung*) und Item 25 aus dem überarbeiteten HILVE-II (Skala *Anforderungen*).
- *Alles in allem hat sich der Besuch der Lehrveranstaltung für mich gelohnt*. Diese Aussage orientiert sich an einem sehr ähnlichen Item des HILVE-I (Nr. 40), das dort dem Faktor *Allgemeinbeurteilung* zugeordnet wird. Dasselbe Item fällt im TRIL in den Bereich *Anwendungsbezug* (dort Nr. 26).

Tabelle 7: Vorbilder und Quellen der einzelnen Items.

Fragebogen	Item	Vorbild	Nr.	Skala / Faktor des Vorbilditems
Wöchentlicher Fragebogen	Wie verständlich waren für Sie die vorgetragenen Inhalte insgesamt?	SEEQ	9	*Organisation*
	Wie verständlich waren die verwendeten Lernmittel (z.B. Folien, Präsentationen, Beispiele)?	TRIL	5	*Struktur und Organisation*
		SEEQ	10	*Organisation*
	Wie schätzen Sie das Tempo ein, das der Vermittlung zugrunde lag?	HILVE-I	19	*Überforderung*
		HILVE-II	25	*Anforderungen*
	Alles in allem hat sich der Besuch der Lehrveranstaltung für mich gelohnt.	HILVE-I	40	*Allgemeinbeurteilung*
	Meiner Meinung nach war die Lehrveranstaltung gut strukturiert.	HILVE-I	1	*Struktur*
	Ich habe mich auf die Stunde vorbereitet.	HILVE-I	29	*Fleiß*
Abschlussfragebogen	An wievielen Tagen der Woche haben Sie am Computer gearbeitet (außer Internetnutzung)?			
	Wieviel Zeit haben Sie insgesamt am Computer gearbeitet (außer Internetnutzung)?			
	An wievielen Tagen der Woche haben Sie das Internet genutzt?			
	Wieviel Zeit haben Sie insgesamt im Internet verbracht?			
	Spontan finde ich den Dozenten eher... (sympathisch=1, nicht sympathisch=5)			
	Was hat Ihnen konkret an der Vorlesung gefallen?	TRIL, HILVE		*Offene Fragen*
	Was hat Ihnen konkret an der Vorlesung nicht gefallen oder Sie gestört?	TRIL, HILVE		*Offene Fragen*

- *Meiner Meinung nach war die Lehrveranstaltung gut strukturiert*. Die Skala *Struktur* des HILVE-I enthält ein sehr ähnliches Item (Nr. 1).
- *Ich habe mich auf die Stunde vorbereitet*. Diese Item lehnt sich ebenfalls an das HILVE-I an (Nr. 29), wo es dem Faktor *Fleiß* zugerechnet wird. Damit bezieht es sich, anders als die vorherigen Items, nicht auf Variablen des Dozenten oder der Veranstaltung, sondern auf eine Studentenvariable (vgl. Tabelle 1 auf Seite 10).
- *Wie sehr wünschen Sie, dass die Lehrveranstaltung anders gestaltet werden sollte?* Im Unterschied zu den vorherigen Items stammt diese Frage nicht aus einem der Inventare, die in Abschnitt 2.1.1 vorgestellt wurden. Stattdessen soll es ein Maß für die Gesamtbeurteilung der Vorlesung sein, ohne aber auf ein Schulnoten-Schema zurückzugreifen, wie es das HILVE-I verwendet (Item 42).

Tabelle 7 in Abschnitt 3.3.1 führt u.a. die einzelnen Items und ihre Quellen noch einmal auf. Die ausgewählten Items wurden in Absprache mit den Dozenten zur Verwendung für wöchentliche Vorlesungen formuliert. Einige gegensätzlich formulierte Adjektivpaare sollten zudem die Befindlichkeit zum Zeitpunkt des Ausfüllens erfassen (*glücklich-unglücklich*, *zufrieden-unzufrieden*, *wohl-unwohl*, *gut-schlecht*). Anhang E enthält die Papierversion des kompletten Fragebogens, Anhang F Bildschirmfotos der Internet-Variante.

3.3.2 Abschlussfragebogen

Am Ende des Semesters erhielten die Teilnehmer einen Fragebogen, der vor allem die wöchentliche Häufigkeit und Dauer der Computernutzung im Allgemeinen und des Internets im Besonderen erfragen sollte (vgl. These 5 in Abschnitt 2.3). Darüber hinaus wurden die Studierenden gebeten anzugeben, wie sympathisch sie den Dozenten fanden, um diese mögliche Biasvariable bei Bedarf kontrollieren zu können (vgl. 2.1.4.3 und These 5).

Neben den globalen Beurteilungen des wöchentlichen Fragebogens sollten die Dozenten auch konkrete, handlungsnahe Rückmeldung erhalten, die Verhaltensänderungen und Interventionen ermöglichen (Rindermann, 2001, S. 250). Darum wurden am Semesterende auch offene Fragen zu guten und schlechten Aspekten der Vorlesungen über das gesamte vergangene Semester gestellt. Diese Fragen orientieren sich an ähnlichen Items in den verschiedenen Inventaren (HILVE, TRIL). Anhang G enthält den vollständigen Abschlussfragebogen, in Tabelle 7 sind zusammenfassend alle verwendeten Items und ihre Vorbilder aus den vorgestellten Inventaren aufgeführt.

3.4 Ablauf der Untersuchung

Zum Auftakt in der ersten Vorlesungswoche wurde die Untersuchung den Studierenden vorgestellt. Sie wurden gebeten, eine Einverständniserklärung zur Teilnahme auszufüllen, und erhielten eine kurze Einführung über die Verwendung der Internetfragebogens.

Beim Einsatz der Papierversion des Fragebogens fanden die Befragungen gegen Ende der Vorlesungen statt. In der Veranstaltung im Hauptstudium wurden die Fragebögen etwa zehn Minuten vor Ende der Veranstaltung im Hörsaal verteilt, in der Veranstaltung im Grundstudium diente dazu eine fünfminütige Pause nach der Hälfte der Vorlesung. In beiden Veranstaltungen wurden die Fragebögen nach dem Ende der Lehrveranstaltung eingesammelt. Für die Teilnehmer, die während der Vorlesung keine Fragebögen mehr erhielten, lagen vor den Büros der Dozenten

zusätzliche Fragebögen aus. Dort konnten auch solche Fragebögen abgegeben werden, die die Teilnehmer versehentlich nach der Vorlesung eingepackt hatten anstatt sie einsammeln zu lassen. Diese Möglichkeit wurde aber kaum genutzt.

Die Befragung im Internet wurde automatisch nach dem Ende der Vorlesung freigeschaltet. Die Studierenden konnten dreieinhalb Tage lang daran teilnehmen: Der Termin der Veranstaltung im Grundstudium lag jeweils am Montagmittag, die Internetbefragung lief entsprechend bis zum Donnerstagabend. Die Veranstaltung im Hauptstudium fand am Dienstagmittag statt, so dass der Befragungszeitraum folglich bis zum Freitagabend lief. Der Befragungszeitraum wurde so begrenzt, dass verzerrende Erinnerungseffekte nicht zu stark werden sollten, insbesondere wurde die Befragung deshalb vor dem Wochenende abgeschlossen. Jeweils nach eineinhalb Tagen wurde eine Erinnerungsmail an diejenigen Teilnehmer verschickt, die den Fragebogen noch nicht ausgefüllt und einer Erinnerung per E-Mail zugestimmt hatten (Anhang I.1).

Die eingegebenen Daten wurden durch das Fragebogensystem auf dem Webserver (Anhang F und Anhang H) auf verschiedene Eingabefehler überprüft. Es kontrollierte, ob von dem jeweiligen Teilnehmer eine Einverständniserklärung vorlag, er in der aktuellen Woche bereits einen Fragebogen ausgefüllt hatte und ob der Fragebogen vollständig ausgefüllt wurde. Am Schluss des Fragebogens erhielten die Teilnehmer eine Übersicht über ihre bisherigen Antworten und die Möglichkeit, diese zu ändern.

Die Dozenten erhielten nach dem Ende des Erhebungszeitraumes (im Sinne einer formativen Evaluation) eine Rückmeldung zur vergangenen Vorlesung, die aus den wesentlichen deskriptiven Kennwerten der Ergebnisse bestand (Mittelwert und Streuung pro Item). Ein Beispiel zeigt das Bildschirmfoto in Abbildung F.17 im Anhang.

Jeweils 10 bis 15 Minuten vor Ende der einzelnen Veranstaltungen wurden die Teilnehmer gezählt, um später den tatsächlichen Rücklauf der Fragebögen ermitteln zu können. Die Studierenden im Grundstudium erhielten in der vorletzten Semesterwoche einen Serienbrief per E-Mail, in der je nach der Anzahl ausgefüllter Fragebögen entweder eine volle (mehr als achtmal teilgenommen), oder eine halbe (mehr als viermal teilgenommen) Versuchspersonenstunde vergeben wurde (Anhang I.2). Dieser Zeitpunkt war deshalb sinnvoll, weil die Studierenden diese Bescheinigung in der darauf folgenden Woche benötigten, um sich für ein Empiriepraktikum im folgenden Semester anmelden zu können.

Zu Schwierigkeiten kam es durch zwei unerwartete Ereignisse:

1. Durch einen Schaden an der Hardware des Webservers, auf dem die Datenerhebung lief, kam es in der Kalenderwoche 20 zu einem Datenverlust von unbekanntem Ausmaß. Betroffen waren die Teilnehmer, die den Internetfragebogen noch am Tag der Vorlesung ausgefüllt hatten. Um den Schaden so gering wie möglich zu halten, wurden die Studierenden per E-Mail gebeten, den Fragebogen noch einmal auszufüllen.
2. In der 23. Kalenderwoche nahmen aufgrund eines Streikes der Studentenschaft gegen die Sparmaßnahmen der Landesregierung erheblich weniger Studierende an der Vorlesung teil. Da in dieser Woche das Risiko bestand, dass nur eine selegierte Stichprobe an der Evaluation teilnahm, wurden die Daten dieses Messzeitpunktes nicht ausgewertet.

Der Abschlussfragebogen (3.3.2) wurde in am letzten Messzeitpunkt verteilt. Eine Internetversion dieses Fragebogens gab es nicht.

3.5 Datenaufbereitung

Die per Internet erhobenen Daten wurden von dem Fragebogensystem in eine Datenbank auf dem Webserver gespeichert (vgl. Anhang H), die täglich gesichert wurde. Die Antworten der Papierfragebögen wurden von Hand über ein spezielles nicht-öffentliches Formular ebenfalls in diese Datenbank eingegeben. Dadurch konnten sowohl für den Papier- als auch den Internetfragebogen dieselben Skripte des Fragebogensystems zur Auswertung verwendet werden. Nach dem Ende der Datenerhebung wurden die Daten zur weiteren Verarbeitung in ein SPSS-lesbares Format umgewandelt.[12]

3.6 Überlegungen zu Störeinflüssen bei der Datenerhebung

In Abschnitt 2.2.1.3 wurden einige Probleme erläutert, die bei Befragungen im Internet auftreten können. Manche davon waren in dieser Studie entweder nicht zu erwarten oder vernachlässigbar, andere sollten durch versuchsplanerische Maßnahmen vermieden oder abgeschätzt werden:

- Der Abdeckungsfehler stellt generell im Falle der studentischen Lehrveranstaltungsevaluation kein Problem dar, da die Zielpopulation genau bekannt ist, nämlich alle Teilnehmer der Vorlesung. Jeder Besucher der Vorlesung kann grundsätzlich den Fragebogen ausfüllen, da Stu-

[12] Eine Schwierigkeit war die völlig gegensätzliche Organisation der Daten der SQL-Datenbank (jedes einzelne Datum erhält eine eigene Zeile) und der Datenorganisation von SPSS (alle Daten einer Versuchsperson müssen in einer einzigen Zeile stehen).

dierende vom Rechenzentrum der Universität Trier Zugang zum Campusnetz und Internet erhalten; zudem sind auch private Internetzugänge bei Studierenden weit verbreitet (Hochschul-Informations-System GmbH [HIS], 2002, S. 182). Die Rahmenpopulation, die durch den Fragebogen erreicht werden konnte, entsprach also der Zielpopulation.

- Auch der Stichprobenfehler (im Sinne einer Verzerrung aufgrund der Vorauswahl der Teilnehmer durch den Versuchleiter; vgl. Couper, 2000) sollte aufgrund der breit angelegten Datenerhebung nicht auftreten: Jeder Besucher der Vorlesung hatte nicht nur die Möglichkeit, diese zu beurteilen, er wurde sogar dazu ermutigt.

- Um die Gefahr des Nichtbeantwortungsfehlers zu minimieren, wurden die Papierfragebögen jeweils rechtzeitig vor dem Ende der Veranstaltung verteilt. Auch das Versenden der Erinnerungsmails bei der Datenerhebung im Internet diente diesem Zweck.

  Das Zählen der Besucher der Vorlesung sollte der Ermittlung der Rücklaufquote dienen. Die Gruppe der Nicht-Teilnehmer setzte sich allerdings zum Einen aus Studierenden zusammen, die von Beginn an nicht an der Studie teilnahmen, und zum Anderen aus Studenten, die zwar grundsätzlich an der Untersuchung teilnahmen, aber trotzdem die Veranstaltung nicht beurteilten. Daher ist die Bezeichnung *Beteiligungsquote* für diesen Kennwert genauer.

  Erfahrungen mit anderen Instrumenten berichten eine ausreichend hohe Interraterreliabilität schon ab einer Stichprobe von $n=15$ Studierenden (vgl. 2.1.3.3); sogar bei einer geringen Rücklaufquote sollte die Stichprobe also hinsichtlich der Reliabilität groß genug sein. Frage 6 (Abschnitt 2.3 auf S. 28) bezieht sich darauf, ob diese Ergebnisse auch valide sind.

Tabelle 4 in Abschnitt 2.1.4.3 führt einige Variablen auf, die mit den studentischen Ratings korrelieren und daher kontrolliert werden sollten; viele dieser Einflussgrößen können aber nur beim Vergleich unterschiedlicher Vorlesungen ein Problem werden. Die wichtigsten dieser Variablen wurden in dieser Untersuchung konstant gehalten: So waren beide Vorlesungen Pflichtveranstaltung (gleicher *Besuchsgrund*) aus dem Bereich Methodenlehre (gleiches *Fach*).

Unterschiede bestanden darin, dass eine Veranstaltung im Grundstudium, die andere im Hauptstudium stattfand; in der ersten wurden zudem Anreize in Form von Versuchspersonenstunden vergeben, in der zweiten nicht. Diesen Unterschieden wurde durch eine entsprechende Auswertungsstrategie Rechnung getragen.

3.7 Operationale Hypothesen und Auswertungsverfahren

Die Daten, die in den beiden Vorlesungen erhoben wurden, ließen sich nur getrennt auswerten, da es sich um unterschiedliche Dozenten und Themen handelte. Deshalb führten die beiden in Abschnitt 3.2 erläuterten Strategien zur Datenerhebung zu insgesamt vier einzelnen Auswertungsdesigns, die in Abbildung 2 dargestellt werden. Zunächst konnte man die messwiederholten

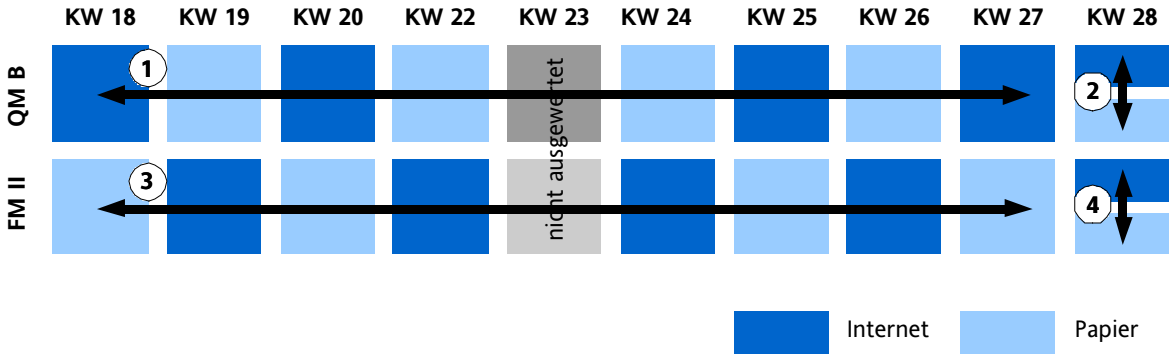


Abbildung 2: Auswertungsdesigns. Kalenderwoche 23 wird aufgrund der geringen Teilnehmerzahl (vgl. 3.4) nicht ausgewertet.

Daten über den Semesterverlauf analysieren: Für die Veranstaltung *Quantitative Methoden B* (QM B) ergab sich damit Auswertungsdesign 1, für die Veranstaltung *Forschungsmethoden II* (FM II) das Auswertungsdesign 3. Die Aufteilung der Stichprobe am letzten Erhebungstermin erlaubte zwei weitere unabhängige Auswertungen (Auswertungsdesign 2 beziehungsweise 4).

In den Auswertungsdesigns 1 und 3 wurden zum einen die Daten aller Messzeitpunkte mit Datenerhebung im Internet und zum anderen die Daten mit Datenerhebung per Papierfragebogen zusammengefasst. Dafür sprachen mehrere Gründe:

- Bei Varianzanalysen mit mehreren messwiederholten Faktoren lägen nur von wenigen Studierenden vollständige Daten vor (siehe auch Abschnitt 4.1.2). Das würde die Teststärke und die Repräsentativität der Daten gefährden.
- Komplexere Auswertungsverfahren für Zeitreihen erfordern ein Vielfaches der gegebenen Messzeitpunkte[13].
- Rindermann und Amelang (1994) argumentieren, dass studentische Rohwerte wesentlich weniger stabil als Veranstaltungsmittelwerte seien, weil sich individuelle Fehleranteile nicht

[13] So sind z.B. 50 Messzeitpunkte für ARIMA-Zeitreihen nach dem Box-Jenkins-Modell erforderlich (Bortz & Döring, 2002, S. 568). Nach Hapuarachchi, March & Wronski (1997) sollten beim ARIMA-Modell idealerweise 100 unabhängige Beobachtungen vorliegen.

ausmitteln könnten. Sie empfehlen deshalb, wenn möglich Veranstaltungsmittel zu verwenden. Zudem erweisen sich Veranstaltungsmittelwerte von Beurteilungen einzelner Veranstaltungen über ein Semester als sehr stabil (Rindermann, 1996, Kap. 16.1). Folgt man dieser Logik (die sich aus dem zweiten Axiom der klassischen Testtheorie ableitet; vgl. z. B. Fisseni, 1997, S. 71), sind auch mehrere über verschiedene Vorlesungen zusammengefasste studentische Beurteilungen reliabler als die Einzelratings.

Wenn man die in Abschnitt 2.3 formulierten Hypothesen vor dem Hintergrund des Versuchsplans dieser Studie betrachtet, kommt man zu folgenden operationalen Hypothesen:

These 1: Die Beurteilungen, die per Internet erhoben wurden, unterscheiden sich nicht von denjenigen, die per Papierfragebogen erhoben wurden. Das gilt sowohl für beide betrachteten Vorlesungen als auch für die beiden unterschiedlichen Auswertungsdesigns (1 und 2 bzw. 3 und 4).

Da substanzielle Korrelationen zwischen den einzelnen Items des Fragebogens theoretisch zu erwarten waren (vgl. Spalte 1 von Tabelle 3 auf S. 16), und sie auch empirisch bestanden (vgl. Anhang J), waren multivariate Testverfahren angemessen, da diese bei hoher Interkorrelation über eine höhere Teststärke verfügen als unabhängige Einzeltests (Stevens, 1996, S. 152). Zudem eigneten sich diese Verfahren besonders für die vorliegende Fragestellung, da sie alle Variablen des Fragebogens simultan testeten und keine Einzelhypothesen für bestimmte Variablen vorlagen.

Für die Auswertungsdesigns 1 und 3 wurde Hotellings T^2-Test für abhängige Stichproben verwendet (vgl. O'Brian & Kaiser, 1985), für die Auswertungsdesigns 2 und 4 Hotellings T^2-Test für unabhängige Stichproben. Die statistische Hypothese war jeweils die globale Nullhypothese. Als abhängigen Variablen wurden die sechs veranstaltungsbezogenen Items des wöchentlichen Fragebogens verwendet.

These 2: Die Teilnehmerzahlen der Messzeitpunkte mit Datenerhebung im Internet in den Auswertungsdesigns 1 und 3 liegen unter denen derjenigen Messzeitpunkte, bei denen der Papierfragebogen eingesetzt wurde.

Angemessen war ein χ^2-Test der summierten Teilnehmerzahlen der Internet- und Papierfragebögen in den Auswertungsdesigns 1 und 3. Die statistische Hypothese war die Alternativhypothese zur Hypothese der stochastischen Unabhängigkeit als Nullhypothese.

These 3: Die Teilnehmerzahlen der Veranstaltung im Grundstudium, in der die Studierenden einen Anreiz für ihre Teilnahme erhalten, sind in den Auswertungsdesigns 1 und 3 höher als die der Veranstaltung im Hauptstudium.

Zum Test dieser operationalen Hypothese wurden die prozentualen Teilnehmerzahlen in den einzelnen Vorlesungen als angemessener Kennwert verglichen; Internet- und Papierfragebögen wurden dabei zusammengefasst. Als statistischer Test diente ein *t*-Test, die Voraussetzungen dafür waren trotz der kleinen Stichprobe erfüllt (Anhang K.2). Die experimentelle Hypothese war eine gerichtete Alternativhypothese.

These 4: Die Variabilität der Befindlichkeit ist bei den Messzeitpunkten mit Datenerhebung im Internet in den Auswertungsdesigns 1 und 3 höher als bei den Messzeitpunkten, zu denen der Papierfragebogen eingesetzt wird.

Zur Prüfung dieser Hypothese wurden die acht Befindlichkeitsitems zu einem Kennwert zusammengefasst. Als statistische Tests waren grundsätzlich der Bartlett- oder Levene-Test auf Varianzhomogenität denkbar, aufgrund der deskriptiven Statistik war in diesem Fall allerdings der Levene-Test passender (Näheres dazu in Abschnitt 4.1.4). Die experimentelle Hypothese war die Alternativhypothese.

These 5: Die Variablen *Befindlichkeit* und *eigene Vorbereitung* korrelieren nicht mit den jeweils gleichzeitig erhobenen studentischen Ratings in den Auswertungsdesigns 1 bis 4.
Die *Sympathie für den Dozenten* korreliert schwach mit den Beurteilungen in den Auswertungsdesigns 1 und 3.
Die *Computer-* und *Internet-Nutzung* beeinflusst die Teilnahmehäufigkeit in den Auswertungsdesigns 1 und 3 signifikant.

In den Auswertungsdesigns 1 und 3 wurden die Korrelationen der *Befindlichkeit* und *eigenen Vorbereitung* mit den Beurteilungen in den jeweiligen Wochen korreliert. Diese Korrelationen wurden anschließend gemittelt und deskriptiv mit den Korrelationen in den Auswertungsdesigns

2 und 4 verglichen. Die experimentelle Hypothese war die Nullhypothese, dass keine signifikanten Korrelationen vorlagen.

Die Variable *Sympathie* wurde nur mit den gleichzeitig erhobenen Beurteilungen in Kalenderwoche 28 korreliert (d.h. in den Auswertungsdesigns 2 und 4). Die *Computer-* und *Internetnutzung* dienten als Prädiktoren einer multiplen Regression zur Vorhersage der Teilnahmehäufigkeit an der Internetbefragung. Die experimentelle Hypothese war die Alternativhypothese, dass die Prädiktoren einen substanziellen Beitrag zur Varianzaufklärung leisten sollten.

Frage 6: Erzielen (regelmäßige) Teilnehmer an der Evaluation der Vorlesung im Grundstudium andere Ergebnisse in der Klausur am Ende des Semesters als Nicht-Teilnehmer?

Aufgrund der hohen Teilnahmerate wurde die Gruppe der Teilnehmer anhand der kumulierten Prozente der Teilnahmehäufigkeit in zwei Untergruppen geteilt (*bis sechsmal teilgenommen* und *mehr als sechsmal teilgenommen*). Diese Aufteilung diente einer ausgewogeneren Besetzung der Zellen bei der nachfolgenden Berechnung der ANOVA. Zusammen mit der Gruppe der Nicht-Teilnehmer ergaben sich folglich drei Stufen des Faktors *Teilnahme*. Hier konnten – wegen des experimentellen Charakters der Frage – keine experimentellen oder statistischen Hypothesen formuliert werden.

4 Ergebnisse

Abschnitt 4.1 stellt die Ergebnisse der einzelnen Hypothesentests vor und erläutert jeweils ihre Bedeutungen. Diese Ergebnisse werden in Abschnitt 4.2 zusammenfassend diskutiert.

4.1 Ergebnisse der Einzelhypothesen

Die statistischen Prüfungen der einzelnen Hypothesen (4.1.1 bis 4.1.6) wurden mit SPSS für Windows durchgeführt (Version 11.0; SPSS Inc., 2001). Bei den Berechnungen zur Teststärke und Effektgröße kam jeweils das Programm GPOWER (Faul & Erdfelder, 1992) zum Einsatz. Da es sich bei der Untersuchung um eine Feldstudie handelte, bei der der Stichprobenumfang nicht planbar und nur begrenzt zu beeinflussen war, wurde die Teststärke jeweils nach Abschluss der Untersuchung (*ex post*) berechnet.

4.1.1 These 1: Äquivalenz der Internet- und Papierversion

Abschnitt 4.1.1.1 geht zunächst auf die Auswertungsdesigns 2 und 4 ein, da diese experimentellen Designs eindeutiger zu interpretieren sind. Die Nullhypothese wurde in beiden Designs durch einen Hotellings T^2-Test für unabhängige Stichproben geprüft. In den quasi-experimentellen Auswertungsdesigns 1 und 3 wurden die Nullhypothese jeweils durch einen Hotellings T^2-Test für abhängige Stichproben getestet (Abschnitt 4.1.1.2). Die Ergebnisse im Auswertungsdesign 3 erfordern eine ausführlichere Erläuterung (Abschnitt 4.1.1.3).

Hotellings T^2-Test beruht auf drei Annahmen (Stevens, 1996, S. 238): Unabhängigen Beobachtungen, der multivariaten Normalverteiltheit der abhängigen Variablen in beiden Gruppen und schließlich homogenen Kovarianzmatrizen der abhängigen Variablen. Diese Voraussetzungen waren erfüllt (Anhang K.1).

4.1.1.1 Auswertungsdesigns 2 und 4

Im Auswertungsdesign 2, d.h. Aufteilung der Besucher der Veranstaltung im Grundstudium in Kalenderwoche 28 in zwei Gruppen, waren beide Gruppe mit jeweils 48 Teilnehmern besetzt. Der entsprechende T^2-Test war nicht signifikant ($F(6, 89) = 1{,}53$; $p = .18 > .05$). Da aus der Literatur keine Effektgrößen abgeleitet werden konnten, wurde bei der anschließenden Berechnung der

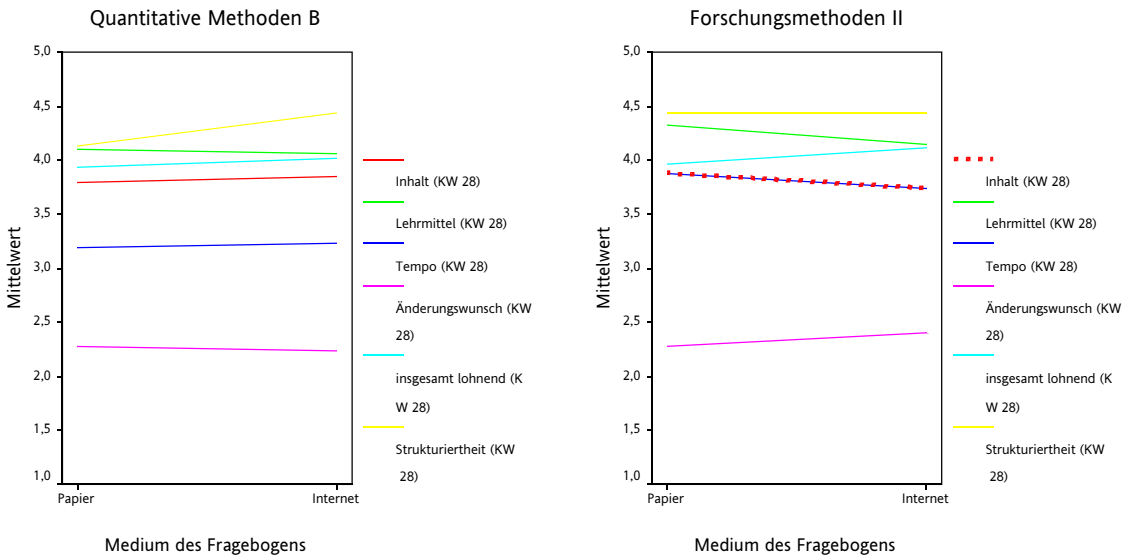


Abbildung 3: Mittelwerte der abhängigen Variablen in den Auswertungsdesigns 2 und 4 in Abhängigkeit vom eingesetzen Medium. Die Unterschiede innerhalb einer Vorlesung sind nicht signifikant.

Teststärke für den interessierenden Effekt ein mittlerer Wert von $\phi^2 = .15$ angenommen (Cohen, 1988, S. 478). Damit ergab sich eine ausreichende Teststärke von 1-ß=.79 (Cohen, 1992).

Das bedeutet: Aufgrund der hohen Teststärke kann die Nullhypothese für das Auswertungsdesign 2 beibehalten werden. Das Medium hat somit keinen bedeutsamen Effekt auf die studentische Beurteilung der Vorlesung. Abbildung 3 stellt zur Verdeutlichung die Mittelwerte der einzelnen abhängigen Variablen dar.

Im Auswertungsdesign 4 (Veranstaltung im Hauptstudium in der Kalenderwoche 28) füllten 27 Versuchspersonen den Papierfragebogen aus, 25 Studierende den Internetfragebogen. Der T^2-Test zum Vergleich der Ratings beider Gruppen ergab kein signifikantes Ergebnis ($F(6,45) = 0.60$; $p = .73 > .05$). Die Teststärke (ebenfalls mit $\phi^2 = .15$) lag hier bei 1-ß=.77. Damit erreichte sie beinahe das Niveau von .80, ab dem man konventionell die Nullhypothese interpretieren kann (Cohen, 1992). Die geringen Unterschiede der Mittelwerte der abhängigen Variablen (Abbildung 3) waren also insgesamt ebenfalls nicht bedeutsam.

4.1.1.2 Auswertungsdesigns 1 und 3

Im Auswertungsdesign 1 (Vergleich der Mittelwerte der per Papier- und Internetfragebogen gewonnen Beurteilungen über den Semesterverlauf im Grundstudium) lagen die Daten von insgesamt 117 Versuchspersonen vor. Der messwiederholte T^2-Test wurde nicht signifikant ($F(6,111) = 1.08$; $p = .38 > .05$). Für die Teststärke ergab sich bei einem mittleren Effekt (ϕ^2=0.15)

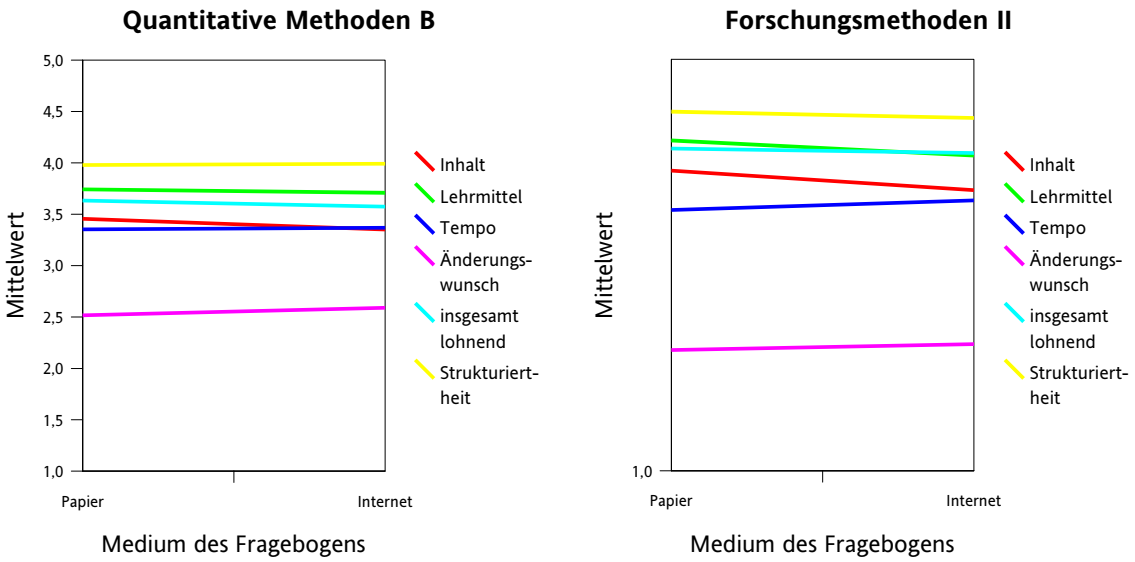


Abbildung 4: Mittelwerte der abhängigen Variablen in den Auswertungsdesigns 1 und 3.

ein hoher Wert von 1-ß=.88. Auch in diesem Fall kann folglich die Nullhypothese beibehalten werden. Abbildung 4 stellt die Mittelwerte der abhängigen Variablen in den beiden Stufen des Faktors *Medium* dar.

Im Auswertungsdesign 3 (Veranstaltung im Hauptstudium) gingen die Daten von 52 Studierenden in die Auswertung ein. Im Unterschied zu den bisher geschilderten Ergebnissen wurde der T^2-Test allerdings signifikant ($F(6,79)=2.18$; $p=.053\approx.05$; 1-ß=.80). Die Nullhypothese muss für Auswertungsdesign 3 also verworfen werden: Die Antwortmuster zwischen Internet- und Papierfragebögen unterschieden sich hier (Abbildung 4). Zudem macht dieses Ergebnis Folgetests erforderlich, um die Struktur des globalen Effektes im Einzelnen aufzudecken.

4.1.1.3 Folgetests im Auswertungsdesign 3

Als Folgetests wurden jeweils einzelne *t*-Tests für gepaarte Stichproben für alle abhängigen Variablen gerechnet (Tabelle 8). Zur Korrektur der α-Fehler-Kumulierung aufgrund der mehrfachen Vergleiche wurde das Signifikanzniveau mit Hilfe der Bonferroni-Korrektur angepasst. Um die Folgetests trotzdem mit einer ausreichenden Teststärke rechnen zu können, wurde das allgemeine α-Fehler-Niveau auf .15 festgesetzt, so dass sich für die einzelnen Tests ein Signifikanzniveau von $\alpha=.15/6=.025$ ergab.

Die Folgetests ergaben ein signifikantes Ergebnis bei den Items *Inhalt* und *Lehrmittel*, wobei die im Internet gesammelten Beurteilungen bei beiden Items niedriger sind. Beide Effekte haben eine mittlere Größe ($0.01<f^2<0.09$; Faul & Erdfelder, 1992). Man muss bei der Interpretation

Tabelle 8: Einzelne *t*-Tests für abhängige Stichproben der abhängigen Variablen in Auswertungsdesign 3. Aufgrund der Bonferroni-Korrektur liegt das Signifikanzniveau bei .025, signifikante Effekte sind hervorgehoben.

	Gepaarte Differenzen (Papierfragebogen - Internetfragebogen)					*t*	*df*	Sig. (2-seitig)	f^2
	M	*s*	*SE(M)*	95% Konfidenzintervall der Differenz					
Items				**Untere**	**Obere**				
Inhalt	0.19	0.59	.06	0.06	0.32	2.99	84	**.004**	0,077
Lehrmittel	0.15	0.55	.06	0.03	0.26	2.49	84	**.015**	0,050
Tempo	-0.09	0.66	.07	-0.23	0.05	-1.3	84	.194	
Änderungswunsch	-0.06	0.47	.05	-0.16	0.04	-1.13	84	.263	
insgesamt lohnend	0.04	0.69	.07	-0.11	0.19	0.58	84	.564	
Strukturiertheit	0.06	0.52	.06	-0.05	0.18	1.09	84	.281	

dieser Effekte allerdings berücksichtigen, dass die betrachteten Variablen die individuellen Differenzwerte der Mittelwerte von jeweils vier Vorlesungen mit Datenerhebung per Papier- und Internetfragebogen sind (vgl. Abschnitt 3.7).

Um den Effekt genauer zu analysieren, wurden deshalb die ursprünglichen Variablen *Inhalt* und *Lehrmittel* über den Semesterverlauf betrachtet. In dieser Situation lag es nahe, eine messwiederholte ANOVA über den Semesterverlauf der Items *Inhalt* und *Lehrmittel* zu rechnen. Das erwies sich jedoch als wenig sinnvoll, da nur von elf Studierenden die für dieses Verfahren erforderlichen vollständigen Daten über alle Messzeitpunkte hinweg vorlagen und die Repräsentativi-

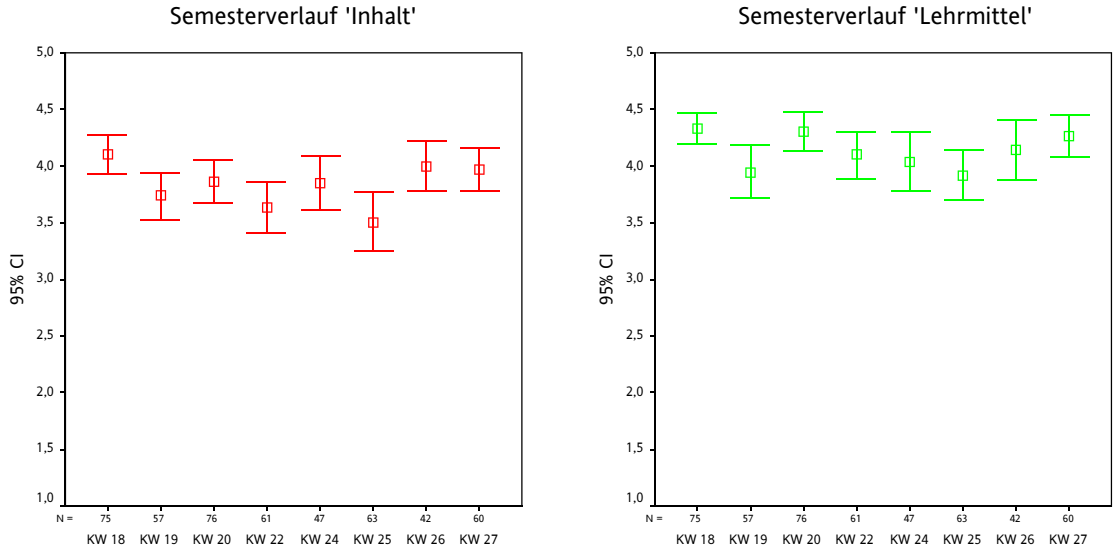


Abbildung 5: Verlauf der Mittelwerte und Konfidenzintervalle der Items *Inhalt* und *Lehrmittel* der Veranstaltung Forschungsmethoden II. In den Kalenderwochen 19, 22, 24 und 26 wurde die Befragung per Internetfragebogen, an den übrigen Terminen (18, 20, 25, 27) per Papierfragebogen durchgeführt (vgl. auch Abbildung 2).

tät dieser Untergruppe zweifelhaft ist (vgl. Abschnitt 3.7). Abbildung 5 zeigt stattdessen grafisch den Verlauf der Mittelwerte und die dazugehörigen Konfidenzintervalle dieser Items. Ein systematischer Effekt einer bestimmten Vorlesung oder eines Mediums zeigt sich in dieser grafischen Darstellung nicht.

4.1.1.4 Fazit und weitere Überlegungen

Was besagen die Ergebnissen der vier einzelnen statistischen Tests über die Bewertung von These 1 insgesamt? Am aussagekräftigsten sind die Ergebnisse in den experimentellen Auswertungsdesigns 2 und 4, in denen die Nullhypothesen mit ausreichenden Teststärken interpretiert werden können.

Wie in Abschnitt 2.3 bereits angesprochen, waren in den quasi-experimentellen Auswertungsdesigns 1 und 3 mehrere Varianzquellen konfundiert, nämlich die verschiedenen Vorlesungen und die unterschiedlichen Medien. Trotz der Mittelung der Daten über mehrere Vorlesungen bedeutet das:

- Ein signifikanter Unterschied zwischen den Ergebnissen des Papier- und Internetfragebogens lässt sich nicht eindeutig auf eine dieser Varianzquellen oder ihre Wechselwirkung zurückführen. Falls es einen systematischen Einfluss einer Varianzquelle gibt, sollte dieser sich bei den Folgetests zeigen.
- Bei einem nicht-signifikanten Ergebnis kann man allerdings nicht von einem systematischen Haupteffekt einer der Varianzquellen sprechen, denn dieser sollte sich auch in den gemittelten Daten zeigen. Allerdings kann man auch eine systematische Wechselwirkung der beiden Varianzquellen nicht ausschließen, weil sie einen möglichen Haupteffekt einer Varianzquelle verdecken könnte.

Damit stützt das bei ausreichender Teststärke nicht-signifikante Ergebnis von Auswertungsdesign 1 die Nullhypothese, dass Papier- und Internetfragebogen äquivalent sind. Das Ergebnis in Auswertungsdesign 3 stützt diese These zwar nicht, widerspricht ihr aber aufgrund der Konfundierung der Varianzquellen, der nur auf zwei Variablen sichtbaren Effekte und vor allem der uneindeutigen Ergebnisse der Folgetests auch nicht.

4.1.2 These 2: Teilnehmerzahlen

Beim Vergleich der Teilnehmerzahlen an den Befragungen per Papier- und Internetfragebogen wurden nur die Auswertungsdesigns 1 und 3 berücksichtigt, da die Zahl der Teilnehmer an der Evaluation hier verlässlicher zu bestimmen war als in den Auswertungsdesigns 2 und 4. Dafür gab es drei Gründe:

1. In den Auswertungsdesigns 2 und 4 wurden die Teilnehmer anhand des Merkmals, ob ihre Matrikelnummer gerade oder ungerade war, den experimentellen Bedingungen zugeteilt (gerade Matrikelnummer: Papierfragebogen, ungerade Matrikelnummer: Internetfragebogen). Trotz ausführlicher Demonstration dieser Unterscheidung kam es in der Vorlesung zu einer Reihe von ungültigen Fragebögen: Einige Studierende füllten trotz ungerader Matrikelnummer den Papier- statt des Internetfragebogens aus (im Grundstudium zehn, im Hauptstudium vier Studenten).
2. Vergleichbare Fehlerzahlen lagen für die Befragung im Internet nicht vor, da hier noch vor Auswertung der Fehlerquoten ungültige Antworten gelöscht wurden (d.h. solche von Studierenden mit geraden Matrikelnummern).
3. Die Studierenden, die zwar die Vorlesung in Kalenderwoche 28 besuchten, aber nicht an der Evaluation teilnahmen, teilten sich vermutlich ebenfalls anhand des Randomiserungskriteriums (gerade oder ungerade Matrikelnummer) in zwei etwa gleichgroße Gruppen. Genaue Zahlen über die Zugehörigkeit dieser *Nicht-Teilnehmer* zu einer der experimentellen Gruppen kann man aber nicht angeben.

Diese Schwierigkeiten führten dazu, dass die genaue Beteiligungsquote in den Auswertungsdesigns 2 und 4 nicht zu berechnen war. Allerdings bedeutete das nur den Verzicht auf die Verwendung der Daten eines Messzeitpunktes, während in den Auswertungsdesigns 1 und 3 acht Messzeitpunkte vorlagen. Abbildung 6 zeigt den Verlauf der Teilnehmerzahlen und die daraus berechneten Beteiligungsquoten für beide Veranstaltungen über das Semester.

Zur Berechnung der Teilnehmerzahlen an den Befragungen per Papier- und Internetfragebogen wurden zum einen die Zahl der Teilnehmer an der Evaluation und zum anderen die Zahl der Nicht-Teilnehmer über alle Vorlesungen mit Einsatz des Papierfragebogens aufsummiert. Anschließend wurden diese Werte mit den Summen der Teilnehmer- und Nicht-Teilnehmer-Zahlen in den Vorlesungen mit Internetbefragung verglichen.

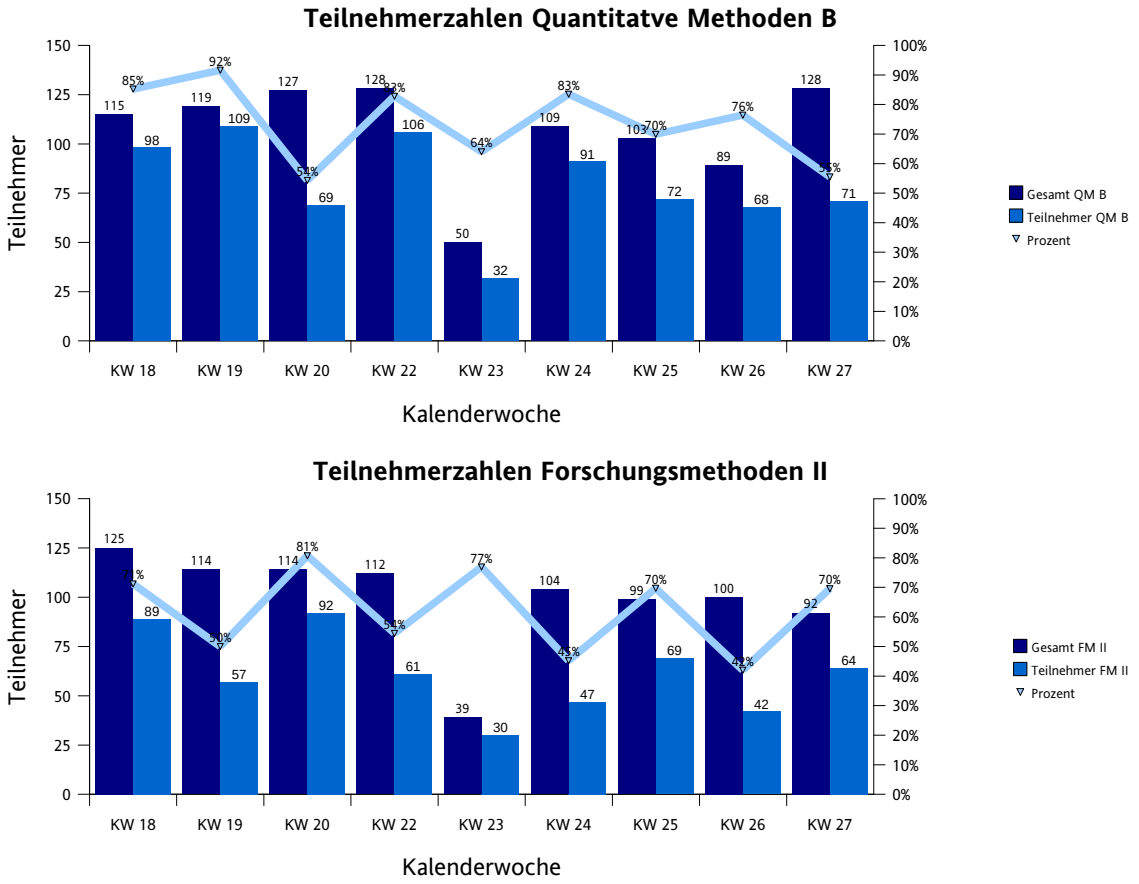


Abbildung 6. Verlauf der Teilnehmerzahlen über das ganze Semester. Deutlich werden die niedrigeren Teilnehmerzahlen in Kalenderwoche 23, in der ein Studentenstreik stattfand. Die Teilnehmerzahlen dieser Woche wurden nicht ausgewertet.

Zum Einsatz kam dabei ein Vier-Felder-χ^2-Test[14] mit den zwei Faktoren *Medium* (Papier und Internet) sowie *Teilnahme* (Teilnehmer und Nicht-Teilnehmer). Die Nullhypothese dieses Tests war die *Unabhängigkeitshypothese*, die die stochastische Unabhängigkeit der beiden Faktoren annimmt. Dagegen entsprach die experimentelle Hypothese einer gerichteten Alternativhypothese:

Tabelle 9: Zwei unabhängige Vier-Felder-χ^2-Tests zum Zusammenhang von Medium und Antwortverhalten für beide Vorlesungen.

	Quantitative Methoden B				**Forschungsmethoden II**			
Medium	**Teilnahme**				**Teilnahme**			
	Teilnehmer		Nicht-Teilnehmer		Teilnehmer		Nicht-Teilnehmer	
Papier	374	(a)	71	(b)	314	(a)	116	(b)
Internet	310	(c)	163	(d)	207	(c)	223	(c)
Pearson χ^2	41.34 (p<0.001)				55.75 (p<0.001)			
Cohens $w = \phi$	.21				.25			
w^2	.045				.065			

[14]Die Voraussetzungen dieses Tests waren erfüllt (unabhängige Beobachtungen, disjunkte und exhaustive Kategorien, erwartete Häufigkeiten >5; Bortz, 1993, S. 163).

Die Teilnehmerzahl sollte bei den Befragungen per Papierfragebogen über der Teilnehmerzahl bei Verwendung des Internetfragebogens liegen. Tabelle 9 enthält die Zellhäufigkeiten und die Ergebnisse der beiden unabhängigen χ^2-Tests. Bei der Veranstaltung im Grundstudium war $\chi^2_{df=1}=41.34$ ($p<.001$), der entsprechende Wert bei der Veranstaltung im Hauptstudium betrug $\chi^2_{df=1}=55.75$ (ebenfalls $p<.001$). Die Richtung der Alternativhypothese wurde bestätigt: In beiden Vorlesungen war die Zahl der Teilnehmer höher, wenn der Papierfragebogen verwendet wurde. Diese Interpretation bestätigt auch das positive Vorzeichen von ϕ, das als Effektgrößenmaß berechnet wurde. Ein positives Vorzeichen dieses Maßes bedeutet, dass die Mehrzahl der Fälle auf der Diagonalen (ad) liegen (Benninghaus, 2002, S. 112). Schließlich zeigt der Vergleich des Effektstärkenmaßes *w*, dass der Effekt im Hauptstudium numerisch größer als im Grundstudium war, in beiden Vorlesungen allerdings eine geringe bis mittlere Stärke aufwies ($.10<w<.30$; Cohen, 1988, S. 225).

4.1.3 These 3: Beteiligungsquote

Um die Hypothese zu prüfen, dass die Beteiligungsquote im Grundstudium insgesamt höher ist als im Hauptstudium, wurden die Beteiligungsquoten in diesen beiden Veranstaltungen über den Semesterverlauf verglichen (Auswertungsdesigns 1 und 3). Die prozentuale Beteiligungsquote wurde für jede Vorlesung aus dem Verhältnis der Zahl der Studierenden, die an der Evaluation teilnahmen, und der Anzahl der Besucher dieser Vorlesung insgesamt berechnet. Abbildung 6 in Abschnitt 4.1.2 stellt neben den absoluten Beteiligungs- und Besucherzahlen auch die relativen Beteiligungsquoten dar.

Aufgrund der geringen Stichprobe ($n=16$) wurden die Voraussetzungen zur Berechnung des *t*-Tests ausführlich geprüft. Diese Voraussetzungen waren erfüllt (Anhang K.2). Tabelle 10 enthält Mittelwerte und Streuungen der Beteiligungsquoten in den beiden Veranstaltungen. Das Ergebnis des einseitigen *t*-Testes war signifikant ($t_{df=14}=2.07$, $p=0.03<0.05$). Für die Effektgröße ergab sich mit $d=1.03$ ein großer Effekt. Damit bestätigte der Test nachdrücklich die Hypothese, dass die Beteiligungsquote in der Veranstaltung im Grundstudium höher als in der Veranstaltung im Hauptstudium liegt.

Tabelle 10: Deskriptivstatistik der Rücklaufquoten in den beiden Vorlesungen.

Veranstaltung	*n*	*M*	*s*	*SE*
Quantitative Methoden B	8	74.90	13.89	4.91
Forschungsmethoden II	8	60.35	14.21	5.02

Bei der inhaltlichen Interpretation dieses Ergebnisses muss man berücksichtigen, dass auch hier aufgrund des quasi-experimentellen Versuchsplans mehrere Varianzquellen konfundiert waren. Neben der Vergabe von Versuchspersonenstunden im Grundstudium als Anreiz, an der Evaluation teilzunehmen, unterschieden sich die beiden Gruppen in weiteren Variablen: Die Teilnehmer im Hauptstudium waren älter als die Studenten im Grundstudium und hatten eine längere Studienerfahrung. Welcher dieser Faktoren für diesen Effekt verantwortlich war, lässt sich mit dem vorliegenden Versuchsplan nicht eindeutig sagen.

4.1.4 These 4: Variabilität der Befindlichkeit

Die Hypothese, dass die Variabilität der Befindlichkeit bei der Befragung im Internet höher als bei der Befragung per Papierfragebogen liegt, wurde durch den Levene-Test auf Varianzhomogenität geprüft. Dieser Test zeichnet sich dadurch aus, dass er robuster gegenüber der Verteilungsform der Daten ist als der Bartlett-Test, allerdings auch etwas weniger sensibel (National Institute of Standards and Technology [NIST], 2003). Insbesondere die Verwendung des getrimmten Mittels als Basis des Levene-Tests stellt einen guten Kompromiss zwischen Teststärke und Robustheit dar (ebd.). Obwohl die Werte für Schiefe und Kurtosis nach den Kriterien von West, Finch und Curran (1995) unproblematisch sind (Tabelle 11), war angesichts der (in einigen Gruppen signifikanten) Ergebnisse des Shapiro-Wilk-Testes auf Normalverteilung (Tabelle K.3.1 im Anhang) der Levene-Test eher angemessen.

Für die Veranstaltung im Grundstudium ergab der Levene-Test ein Ergebnis von $F(1,94)=6.57$ und war damit signifikant ($p=.01<.05$). Als Effektgröße zeigte sich mit $f=0.24$ ein mittlerer Effekt (Cohen, 1988, S. 286). Das Ergebnis für die Veranstaltung im Hauptstudium war allerdings bei ausreichender Teststärke nicht signifikant ($F(1,50)=1.98$, $p>.05$, 1-ß=.78 für einen mittleren Effekt). Abbildung 7 stellt Mittelwerte und die (doppelte) Standardabweichungen der allgemeinen Befindlichkeit in den beiden Vorlesungen dar.

Diese Ergebnisse sind uneindeutig und bestätigen die experimentelle Hypothese nur zum Teil. Während bei der Veranstaltung im Grundstudium die Variabilität der Befindlichkeit bei Befra-

Tabelle 11: Deskriptivstatistik der allgemeinen Befindlichkeit in Kalenderwoche 28.

Veranstaltung	Medium	*n*	*M*	Minimum	Maximum	*s*	s^2	Schiefe	Kurtosis
Quantitative Methoden B	Papier	48	3.89	1.88	4.63	0.58	0.339	-1.47	2.50
	Internet	48	3.78	2.00	5.00	0.80	0.646	-0.68	-0.32
Forschungsmethoden II	Papier	25	3.96	2.50	5.00	0.66	0.433	-0.73	0.38
	Internet	27	3.55	1.50	4.63	0.80	0.636	-0.63	-0.02

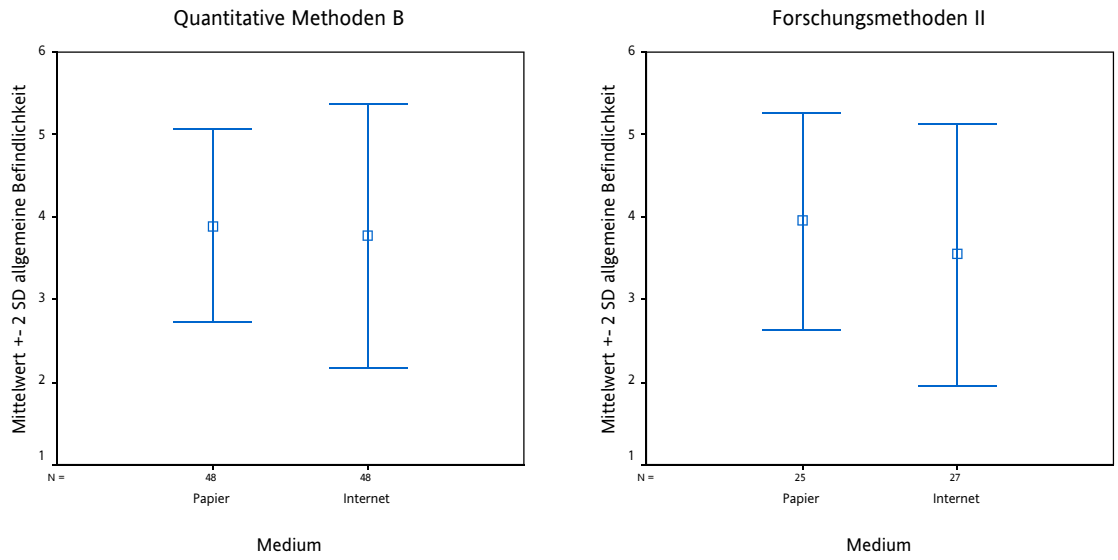


Abbildung 7: Mittelwerte und (doppelte) Standardabweichungen der allgemeinen Befindlichkeit in den beiden Vorlesungen am letzten Messzeitpunkt.

gung im Internet höher war als bei Verwendung des Papierfragebogens, zeigte sich dieser Zusammenhang im Hauptstudium nicht (zumindest nicht in der angenommenen Höhe). Stattdessen muss man dort aufgrund der Teststärke von .80 die Nullhypothese beibehalten, denn die Unterschiede in der Variabilität waren nur gering. Offenbar ist der Zusammenhang zwischen dem verwendeten Medium und der Variabilität der Befindlichkeit komplexer als angenommen.

4.1.5 These 5: Einfluss von Kovariaten

Die Biasvariable *Sympathie* für den Dozenten wurde mit den jeweils gleichzeitig erhobenen studentischen Beurteilungen der Vorlesung korreliert. Auch die *eigene Vorbereitung* auf die Vorlesung und die *Befindlichkeit* wurden als Kontrollvariablen mit berücksichtigt. Tabelle J.1 und Tabelle J.2 im Anhang enthalten die vollständigen Korrelationsmatrizen für die beiden Vorlesungen im Grund- und Hauptstudium am letzten Messzeitpunkt[15], Abschnitt 4.1.5.1 erläutert die Zusammenhänge.

Die deskriptiven Kennwerte der Vertrautheit mit Computern der Studenten in beiden Vorlesungen finden sich in Tabelle K.4.1 im Anhang. Die Rohwerte wurden um Extremwerte bereinigt, *z*-standardisiert und zu zwei Kennwerten zusammengefasst (*allgemeine Computernutzung* und *allgemeine Internetnutzung*). Die Vorhersage der Teilnahmehäufigkeit durch diese Variablen wird in Abschnitt 4.1.5.2 dargestellt.

[15] Die Tabellen J.3 bis J.6 im Anhang enthalten zusätzlich jeweils die einzelnen Korrelationsmatrizen bei Verwendung des Papier- und Internetfragebogens.

Auch dem Signifikanztest bei Korrelationen liegen die Annahmen von Normalverteilung und homogenen Varianzen zugrunde. Da sich aber sowohl der Test selbst als auch Power-Schätzungen als robust gegenüber der Verletzung der Annahmen erwiesen haben (Cohen, 1988, S. 75), werden diese Annahmen im Folgenden nicht geprüft.

4.1.5.1 Zusammenhänge von Kontrollvariablen mit den studentischen Beurteilungen

Wenn man danach fragt, wie die studentischen Beurteilungen mit den Störvariablen zusammenhängen, muss man nur einen Teil der Korrelationsmatrizen betrachten: Die Interkorrelationen der einzelnen abhängigen Variablen sind in diesem Fall nicht wichtig. Stattdessen sind nur die 18 bivariaten Korrelationen zwischen der Biasvariable *Sympathie* und den Kontrollvariablen *eigene Vorbereitung* und *Befindlichkeit* einerseits mit den sechs abhängigen Variablen andererseits von Bedeutung[16].

Die Korrelationsmatrizen der beiden Vorlesungen im Grund- und Hauptstudium zum letzten Messzeitpunkt (Kalenderwoche 28) wurden getrennt voneinander ausgewertet. Beim Test der Korrelationskoeffizienten auf statistische Signifikanz wurde jeweils das allgemeine α-Fehler-Niveau auf $\alpha = .20$ gesetzt. Für die einzelnen *t*-Tests ergab sich damit bei Anwendung der Bonferroni-Korrektur ein Wert von $\alpha = .0\overline{1}$. Auf diesem Niveau wurden folgende Korrelationen statistisch signifikant:

- Die *Sympathie für den Dozenten* korrelierte im Grundstudium positiv mit der Beurteilung des *Inhalts* ($r = .27$) und der *Strukturiertheit* ($r = .30$). Dagegen korrelierte der *Änderungswunsch* negativ ($r = -.30$) mit der Sympathie. Die Effekte haben also eine mittlere Größe (Cohen, 1988, S. 77). Die übrigen Korrelationen wurden statistisch nicht signifikant (bei $1-ß = .72$ für einen mittleren Effekt von $r = .30$ und $n = 89$).

  Im Hauptstudium zeigten die Vorzeichen der Korrelationskoeffizienten ein ähnliches Muster, allerdings waren diese hier nicht signifikant. Dabei muss man allerdings berücksichtigen, dass die Teststärke hier äußerst gering war ($1-ß = .42$ für einen mittleren Effekt und $n = 47$).

- Die *eigene Vorbereitung* korrelierte im Grundstudium (im Auswertungsdesign 2, also Kalenderwoche 28) signifikant mit der Bewertung des *Inhaltes* ($r = .40$) und der Einschätzung der Veranstaltung als *insgesamt lohnend* ($r = .40$). Diese Korrelationskoeffizienten hatten damit ebenfalls eine mittlere Höhe, die Teststärke lag bei $1-ß = .77$ (für einen mittleren Effekt und

[16] Die für diese Fragestellung wichtigen Ausschnitte sind in Tabelle J.1und Tabelle J.2 im Anhang hervorgehoben.

$n=96$). Auch für diese Variable lag die Teststärke für die Veranstaltung im Hauptstudium wesentlich niedriger ($1-ß=.47$, $n=52$), es gab dort keine signifikanten Effekte.

Insgesamt ähnelten diese Effekte in Kalenderwoche 28 in ihrer Struktur den Ergebnissen, die man bei der Zusammenfassung der Korrelationsmatrizen über den Semesterverlauf erhielt (Tabelle 12). Allerdings waren die über den Verlauf des Semesters gemittelten Korrelationskoeffizienten insgesamt kleiner.

- Der Zusammenhang der Variable *Befindlichkeit* mit den studentischen Beurteilungen zeigte insgesamt ein uneinheitliches Bild (siehe auch Tabelle J.8 im Anhang). Im Auswertungsdesign 4 (Kalenderwoche 28 im Hauptstudium) korrelierte die Befindlichkeit signifikant positiv und in mittlerer Höhe mit der Variablen *Inhalt* ($r=.37$; $p=.01$); andere Korrelationen waren auf dem α-Fehler-Niveau von $\alpha=.0\overline{1}$ nicht signifikant und von der Höhe gering ($r<.30$, vgl. Cohen, 1988, S. 79f.) oder zu vernachlässigen ($r<.10$).

  Die über den Semesterverlauf (Auswertungsdesigns 1 und 3) gemittelten Korrelationen waren im allgemeinen numerisch größer (Tabelle 12). Ihr Muster deckte sich aber kaum mit den Ergebnissen aus den Auswertungsdesigns 2 und 4.

Tabelle 12: Gemittelte Korrelationen der Kontrollvariablen *eigene Vorbereitung* und *Befindlichkeit* mit den studentischen Bewertungen über den Semesterverlauf (Auswertungsdesigns 1 und 3).

Kontroll-variable	Vorlesung	Inhalt	Lehrmittel	Tempo	Änderungs-wunsch	insgesamt lohnend	Struktur-iertheit
eigene Vorbereitung	QM B	.23	.13	-.03	-.01	.20	.11
	FM II	.13	.14	.02	.05	.10	.14
Befindlichkeit	QM B	.29	.29	-.16	-.14	.26	.14
	FM II	.26	.21	-.06	-.21	.31	.23

Anmerkung: Zur Berechnung wurden die Korrelationen Fisher-Z-transformiert. Die Rohwerte sind in Tabelle J.7 und Tabelle J.8 im Anhang aufgeführt.

Die mittleren Korrelationen der Variable *Sympathie* mit einigen Variablen der studentischen Ratings im Grundstudium entsprechen grundsätzlich der Hypothese eines schwachen Zusammenhangs. Für einige Variablen (*Inhalt*, *Strukturiertheit* und *Änderungswunsch*) zeigte sich ein stärkerer Zusammenhang, andere wurden dagegen nicht signifikant. Aufgrund der geringen Teststärke waren keine Aussagen über die Korrelationen im Hauptstudium möglich.

Die Ergebnisse widersprechen allerdings der Nullhypothese, dass die *Befindlichkeit* nicht mit der Beurteilung der Vorlesung zusammenhängt. Allerdings kann aufgrund der Korrelation nicht auf die Art des Zusammenhangs geschlossen werden; neben einer kausalen Beziehung in der einen

oder anderen Richtung sind auch gegenseitige Beeinflussungen oder Effekte von Drittvariablen möglich (vgl. dazu z.B. Jaccard, Turrisi & Wan, 1990, S. 9). Das gilt auch für die unerwarteten Zusammenhänge der abhängigen Variablen mit der Kontrollvariable *eigene Vorbereitung*.

4.1.5.2 Vertrautheit mit Computern und Teilnahmehäufigkeit im Internet

Im Unterschied zu den bislang erläuterten Korrelationen wurde der Zusammenhang von Computer- und Internetnutzung mit der Teilnahmehäufigkeit bei Verwendung des Papier- und Internetfragebogens per multipler Regression geprüft. Dabei gingen *allgemeine Computernutzung* und *allgemeine Internetnutzung* als Prädiktoren für die Teilnahmehäufigkeit in die Gleichung ein.

Weder die Computernutzung noch die Internetnutzung stellten signifikante Prädiktoren der Teilnahmehäufigkeit dar (vgl. Tabelle L.1 und Tabelle L.2 im Anhang). Die Teststärke betrug für beide Fälle 1-ß=.97 (für einen mittleren interessierenden Effekt von ϕ^2=.15; Cohen, 1988, S. 413; Cohen, Cohen, West & Aiken, 2003, S. 93).

Zur Kontrolle wurde analog eine Regression der Teilnahmehäufigkeit bei Verwendung des Papierfragebogens auf dieselben Prädiktoren gerechnet. Auch hier wurde bei gleicher Teststärke kein Prädiktor signifikant (Tabelle L.3 und Tabelle L.4 im Anhang).

Abbildung 8 zeigt die Regressionsgeraden von Teilnahmehäufigkeit und Internetnutzung für die beiden Medien des Fragebogens (Internet und Papier). Neben den geringen Steigungen der Ge-

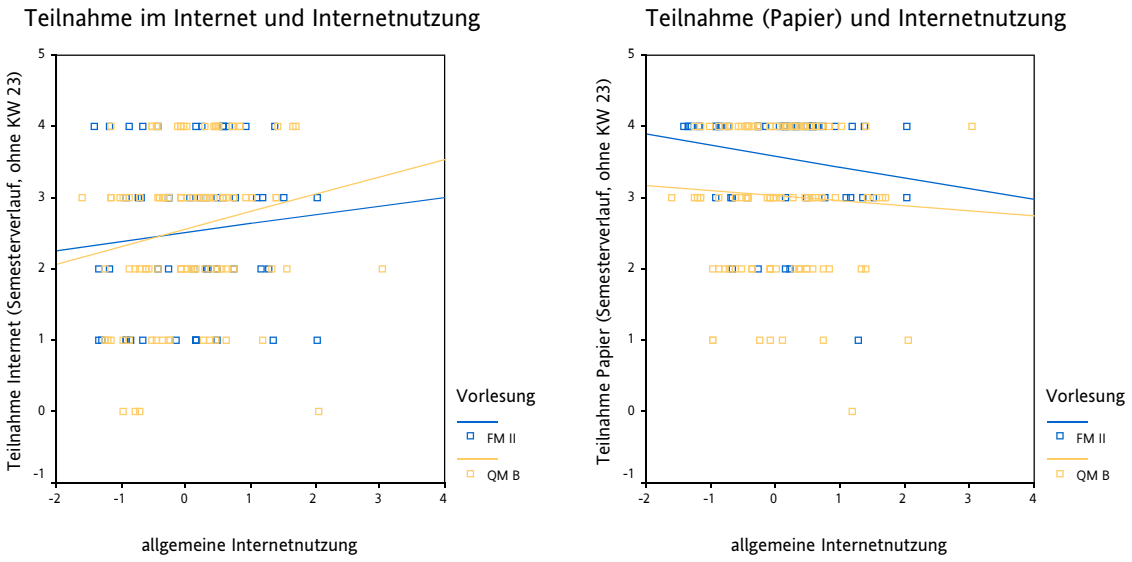


Abbildung 8: Streudiagramme und Darstellung der bivariaten Regression der Teilnahmehäufigkeit auf die *allgemeine Internetnutzung*. Aufgrund der geringen Regressionskoeffizienten der Variable *Computernutzung* wurde diese nicht dargestellt. Das linke Diagramm stellt den Zusammenhang für den Internetfragebogen dar, das rechte Diagramm den Zusammenhang bei Verwendung des Papierfragebogens: Beide Effekte waren nicht signifikant. Um die Ergebnisse einfacher interpretierbar zu halten, wurden die Teilnahmehäufigkeiten nicht *z*-standardisiert.

raden, die den schwachen Zusammenhang zwischen diesen Variablen verdeutlichen, fällt auch die hohe vertikale Lage der Geraden auf, besonders im Vergleich der beiden Diagramme. Diese Lage verdeutlicht den generellen Einfluss des Mediums auf die Teilnahmehäufigkeit, die bei Verwendung des Papierfragebogens höher lag (vgl. dazu Abschnitt 4.1.2). Die Berücksichtigung der individuelle Nutzung von Computern und Internet als Prädiktoren kann die Vorhersage der individuellen Teilnahmehäufigkeiten also nicht weiter verbessern.

4.1.6 Frage 6: Prüfungsnoten

Die Teilnehmer und Nicht-Teilnehmer an der Evaluation im Grundstudium wurden hinsichtlich ihrer Prüfungsleistung verglichen (erreichte Punktzahl in der Klausur am Semesterende). Die Studierenden wurden dazu jeweils in eine der drei Stufen *nicht teilgenommen* (47 Studenten), *bis sechsmal teilgenommen* (50 Studenten) und *mehr als sechsmal teilgenommen* (74 Studenten) eingeordnet (Tabelle 13).

Tabelle 13: Mittlere Ränge der abhängigen Variable *Prüfungsleistung* in den einzelnen Gruppen.

	Teilnahmehäufigkeit	***n***	**Mittlerer Rang**
Erreichte Punktzahl in der Klausur	nicht teilgenommen	47	59.57
	bis sechsmal	50	88.81
	mehr als sechsmal	74	100.89
	Gesamt	$N = 171$	

Beim Vergleich dieser Gruppen kam eine Kruskal-Wallis-Rangvarianzanalyse zum Einsatz (Bortz, Lienert & Boehnke, 2000, S. 222), da die Voraussetzungen zur Berechnung einer ANOVA verletzt waren (Anhang K.5). Tabelle 13 zeigt die mittleren Ränge in den einzelnen Gruppen. Das Ergebnis dieser Rangvarianzanalyse war signifikant ($\chi^2_{df=2} = 20.26$, $p < 0.001$). Um eine Vorstellung von der Größe des Effektes zu erhalten, wurde die rangbiseriale Korrelation berechnet (vgl. Glass, 1966), die einen Wert von Kendalls $\tau_b = .26$ ergab. Aufgrund der Höhe dieses Korrelationskoeffizienten kann man von einem geringen bis mittleren Effekt sprechen (Cohen, 1988, S. 80).

Um schließlich die Struktur dieses Effektes aufzudecken, wurden als Folgetests insgesamt drei Mann-Whitney-*U*-Tests zum Vergleich zwischen den Gruppen berechnet. Das allgemeine α-Fehler-Niveau lag bei .05, das per Bonferroni-Korrektur ermittelte Signifikanzniveau für die einzelnen Tests ergab jeweils einen Wert von $.05/3 = .017$. Es zeigte sich, dass die Unterschiede zwischen der Gruppe der Nicht-Teilnehmer und den zwei anderen Gruppen signifikant waren (jeweils $p < .01$), die beiden Gruppe der Teilnehmer untereinander sich aber nicht statistisch signifi-

kant unterschieden (Tabelle 14). Abbildung 9 stellt darüber hinaus die mittlere erreichte Punktzahl in den verschiedenen Gruppen grafisch dar.

Tabelle 14: Ergebnisse der als Folgetests berechneten *U*-Tests. Die Nicht-Teilnehmer an der Evaluation erzielten im Schnitt ein schlechteres Ergebnis als die Teilnehmer. Da sich die *U*-Verteilung bei großen Stichproben asymptotisch der Normalverteilung annähert, sind auch die *z*-Werte aufgeführt.

Vergleich zwischen...					
Gruppe 1	**Gruppe 2**	**Mann-Whitney *U***	***z*-Wert**	***p***	**Kendalls τ_b**
nicht teilgenommen	bis sechsmal	786.0	-2.80	.005	.24
nicht teilgenommen	mehr als sechsmal	886.0	-4.54	< .001	.34
bis sechsmal	mehr als sechsmal	1601.5	-1.27	.21	.10
Anmerkung:	Kendalls τ_b beschreibt den Zusammenhang zwischen Gruppenzugehörigkeit und erzieltem Rangwert jeweils für die beiden verglichenen Gruppen.				

Diese Ergebnisse zeigen, dass sich die Teilnehmer an der Evaluation hinsichtlich ihrer Prüfungsnoten deutlich von den Nicht-Teilnehmern unterschieden. Wie oft sie an der Evaluation teilnahmen, spielt dagegen eine geringere Rolle. Man sollte bei der Interpretation der Ergebnisse wiederum bedenken, dass sich die Gruppe der Nicht-Teilnehmer aus zwei Untergruppen zusammensetzte, die in dieser Studie nicht zu unterscheiden waren:

1. Studierende, die zwar die Vorlesung besuchten, aber nicht an der Evaluation teilnahmen (Untergruppe A) und
2. Studenten, die nie die Vorlesung besuchten (Untergruppe B).

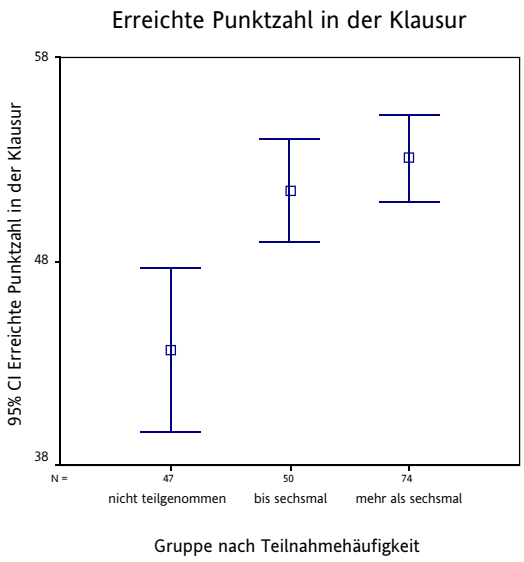


Abbildung 9: Mittlere Punktzahl in der Klausur am Semesterende. Ab 38 Punkten gilt die Prüfung als nicht bestanden.

Unterschiede in den Prüfungsleistungen zwischen den regelmäßigen Teilnehmern und der Untergruppe A der Nicht-Teilnehmer könnten auf Selektionseffekte bei der Teilnahme hindeuten, die die Validität der Evaluationsergebnisse gefährden könnten (etwa wenn nur bessere Studierende an der Evaluation teilgenommen hätten). Unterschiede zwischen den regelmäßigen Teilnehmern und der Untergruppe B sind im Sinne der Veranstaltungsevaluation allerdings sogar günstig, da sie eine wirksame Wissensvermittlung in der Vorlesung nahe legen.

4.2 Zusammenfassung der Ergebnisse

Die Hauptthese dieser Untersuchung, die Äquivalenz zwischen der Lehrveranstaltungsevaluation per Papier- und Internetfragebogen, konnte deutlich belegt werden. Besonders aussagekräftig sind dabei die Ergebnisse des letzten Messzeitpunktes, an dem die Stichprobe aufgeteilt wurde.

Besonderes Augenmerk wurde auf den Einfluss von systematischen Selektionseffekten gelegt. Obwohl die Teilnehmerzahlen bei Verwendung des Internetfragebogens geringer waren als bei Verwendung des Papierfragebogens und obwohl die Verwendung eines Anreizes die Beteiligungsquote erhöhte, verzerrten diese Effekte nicht systematisch das Ergebnis der Evaluation. Die Fragebögen in beiden Medien erwiesen sich trotzdem als äquivalent.

Die individuelle Teilnahmehäufigkeit an der Internetbefragung war zwar geringer als die Teilnahmehäufigkeit bei papierbasierten Befragungen; über diesen allgemeinen Effekt des Mediums hinaus hatte die individuelle Vertrautheit mit Computern und dem Internet jedoch keinen Einfluss auf die Teilnahmebereitschaft. Auch diese Variablen verursachten also keine systematische Selektion, die zu einer Verzerrung der Ergebnisse hätten führen können.

Eine Schwierigkeit der Datenerhebung im Internet ist die geringere experimentelle Kontrolle. Diese hätte sich darin zeigen können, dass die Stimmung, in der die Versuchspersonen den Internet-Fragebogen ausfüllten, stärker variierte als bei der Befragung mit dem Papierfragebogen. Dieser Effekt zeigte sich allerdings nur zum Teil, d.h. nur im Grundstudium, und dort auch nur in mittlerer Stärke.

Die *Sympathie* der Studierenden für den Dozenten korrelierte in mittlerer Höhe mit der Beurteilung der Vorlesung. Über die Richtung des Zusammenhangs kann bei dem vorliegenden Versuchsplan allerdings keine Aussage getroffen werden.

Die Kontrollvariable *eigene Vorbereitung* wies im Grundstudium einen stabilen, mittleren Zusammenhang mit den Beurteilungsmaßen *Inhalt* und der Einschätzung der jeweiligen Veranstaltung als *insgesamt lohnend* auf; die Korrelationen mit den Variablen *Lehrmittel*, *Strukturiertheit*, *Tempo* und *Änderungswunsch* waren dagegen nur gering oder zu vernachlässigen. Marsh (1987; vgl. auch Marsh & Roche, 1997) weist auf Parallelen zum Vorgehen bei der Konstruktvalidierung hin (Multitrait-Multimethod-Validierung; Fisseni, 1997, S. 108):

> [This] approach [...] is based upon the assumption that specific variables (for example, background characteristics, validity criteria, experimental manipulations, etc.) should logically and theoretically be related to some specific components of student's evaluations, and less related to others. According to this ap-

> proach, if a variable is most highly correlated with the dimensions to which it is most logically connected, then the validity of the ratings is supported. (S. 312)

Nach dieser Argumentation kann man trotzdem das relativ stabile Muster der Korrelation im Grundstudium als Beleg für die Validität der Beurteilungen interpretieren; schließlich bestand nur ein schwacher Zusammenhang zwischen der studentischen Variable *eigene Vorbereitung* und Dozentenvariablen wie *Lehrmittel* und *Strukturiertheit*. Vorsicht bei weiter reichenden Interpretationen ist allerdings deshalb angebracht, weil dieses Korrelationsmuster im Grundstudium zwar relativ stabil, aber nicht auf das Hauptstudium übertragbar war.

Auch der Zusammenhang der studentischen Ratings mit der *Befindlichkeit* der Studierenden scheint insgesamt nur gering zu sein. Es gab zwar eine signifikante Korrelation mittlerer Stärke (mit der Bewertung des *Inhaltes*) in Auswertungsdesign 4, andere Korrelationen waren aber meist geringer und statistisch nicht signifikant. Zudem waren die Zusammenhänge über den Semesterverlauf geringer; zwischen den einzelnen Auswertungsdesigns zeichnete sich kein klares Korrelationsmuster ab.

Insgesamt erzielten Studierende, die regelmäßig an der Vorlesungsevaluation im Grundstudium teilnahmen (und damit auch die Vorlesung besuchten), am Ende des Semesters bessere Prüfungsleistungen als Studierende, die nicht an der Evaluation teilnahmen (und möglicherweise auch nicht die Vorlesung besuchten).

5 Diskussion und Ausblick

Abrami und d'Apollonia (1999) kommen zu dem Schluss, dass die Randbedingungen für den Einsatz der studentischen Lehrveranstaltungsevaluation zu einem großen Teil erforscht sind, die Einwände dagegen aber trotzdem ähnlich bleiben ("current concerns are past concerns"). Wenn man der studentischen Lehrevaluation durch Papierfragebögen nun ihre Validität zugesteht – und das sollte man aufgrund der Fülle der vorliegenden empirischen Ergebnisse tun – dann muss diese Einschätzung auch für die studentische Lehrevaluation im Internet gelten. Die vorliegende Studie konnte die Äquivalenz der beiden Erhebungsformen zeigen, einen systematischen Selektionseffekt gab es trotz der geringeren Beteiligungsquote im Internet nicht. Offenbar sind fehlende Computer- oder Internet-Kenntnisse kein Hinderungsgrund, an der Evaluation im Internet teilzunehmen. Der in der Fragebogenforschung gefürchtete Abdeckungsfehler liegt beim Einsatz von Internetfragebögen zur Lehrveranstaltungsevaluation nicht vor, zumindest nicht stärker als bei beim Einsatz von Papierfragebögen.

Im Verlauf der Durchführung und Auswertung dieser Untersuchung ergaben sich jedoch einige Probleme und Überlegungen, wie das Untersuchungsdesign bei zukünftigen Studien verbessert werden könnte (Abschnitt 5.1). Zum Teil münden diese Überlegungen in praktischen Vorschläge zur Durchführung der Evaluation im Internet (5.2).

5.1 Probleme und Verbesserungsvorschläge

Die wichtigsten Verbesserungsvorschläge betreffen den verwendeten Fragebogen (5.1.1) und den Versuchsplan (5.1.2). Das Feedback an die Dozenten könnte vergleichsweise einfach im Detail verbessert werden (5.1.4), die weitere Analyse der Zusammenhänge mit den Kontrollvariablen (5.1.3) ist dagegen aufwendiger. Die Diskussion zu fehlenden Werten und komplexeren Auswertungsstrategien (5.1.5) leitet schließlich zu einigen praktischen Überlegungen zur Durchführung der studentischer Lehrveranstaltungsevaluationen im Internet über.

5.1.1 Fragebogen

Ein großes Problem dieser Untersuchung war die Verwendung eines selbst entworfenen Fragebogens mit nur wenigen Items. Dadurch sind die Korrelationen zwischen den studentischen Ur-

teilen und Kontrollvariablen mit den Ergebnissen anderer Studien nur schwer zu vergleichen. Marsh (1987, S. 260) sieht ein Problem der Forschung zur studentischen Lehrveranstaltungsevaluation “in the plethora of *ad hoc* instruments based upon varied item content and untested psychometric properties”. Man sollte in zukünftigen Studien versuchen, diesem Vorwurf durch Verwendung eines bewährten, validierten Fragebogens zu entgehen. Trotzdem trifft dieser Vorwurf nur zum Teil zu, da die abhängigen Variablen nicht ungeprüft zu Skalen zusammengefasst wurden, sondern die Antworten auf der Ebene der Einzelitems ausgewertet wurden. Ihre Interkorrelation wurde vor Beginn der statistischen Auswertung überprüft (Anhang J). Zudem standen bei der Konstruktion des Fragebogens bewährte Verfahren im Hintergrund.

Die Verwendung eines eigenen Fragebogens führt zu einem weiteren Problem: Streng genommen ist die Äquivalenz des Internet- und Papierfragebogens nämlich nur für den verwendeten Fragebogen nachgewiesen und müsste den Forderungen des Testkuratoriums (1986) entsprechend für andere Verfahren wie das HILVE oder das TRIL zusätzlich belegt werden. Allerdings erweisen sich die Forderungen des Testkuratoriums als sehr starr. Hänsgen (1999) schlägt vor, nicht die Äquivalenz für jeden einzelnen Test nachzuweisen, sondern inhaltliche und gestalterische Einflussfaktoren zu finden und zu untersuchen, wie diese auf die Äquivalenz wirken können. In dieser Studie wurden die Fragebögen in beiden Medien möglichst ähnlich gestaltet (Anhang E und Anhang F); da sich Fragebogenverfahren im Allgemeinen aber als äquivalent erwiesen haben (Abschnitt 2.2.1.2), sollte die Gestaltung keinen großen Einfluss auf das Antwortverhalten haben. Im Vergleich zur Frage, ob ein Student überhaupt an der Befragung teilnimmt (Selektionseffekte, die in dieser Studie nicht gefunden wurden), ist der Effekt der Gestaltung des Fragebogens vermutlich gering.

Der selbst entworfene, kurze Fragebogen wurde auch deshalb verwendet, weil die Befragung wöchentlich durchgeführt wurde und das Ausfüllen eines längeren Fragebogens die Lehrveranstaltungen gestört hätte. Da sich die Ergebnisse der Evaluation über die unterschiedlichen Wochen als relativ stabil erwiesen (Anhang M), kann man für die praktische Durchführung der Evaluation die Schlussfolgerung ziehen, dass man die Studierenden seltener befragen kann (z. B. zweimal pro Semester), dafür aber einen ausführlicheren Fragebogen verwenden sollte (siehe dazu auch Abschnitt 5.2). Die Wahl und Gestaltung des Fragebogens hängt folglich auch vom Versuchsplan ab.

5.1.2 Versuchsplan

Der in dieser Studie verwendete Versuchsplan (Abschnitt 3.2) hat sich nur teilweise bewährt. Während die Ergebnisse in den Auswertungsdesigns 2 und 4 (d.h. bei Aufteilung der Stichprobe) eindeutig zu interpretieren waren, waren in den Auswertungsdesigns 1 und 3 (Semesterverlauf) mindestens zwei Varianzquellen konfundiert: die unterschiedliche Qualität der Vorlesungen in den einzelnen Wochen und das Medium des Fragebogens. Dieses Problem führte zu erheblichen Schwierigkeiten bei der Interpretation des Ergebnisses im Hauptstudium, das einen signifikanten Unterschied zwischen den Antworten ergab, die per Internet- und Papierfragebogen erhoben wurden (Abschnitt 4.1.1.2). Diese Unterschiede blieben allerdings ohne klare Struktur (4.1.1.3).

Trotz der beschriebenen Probleme bei der Analyse der studentischen Ratings in den Auswertungsdesigns 1 und 3 erwies sich der verwendete Versuchsplan bei der Untersuchung der Teilnehmerzahlen in denselben Designs als günstig. Da die Vorlesungen hier abwechselnd im Internet oder per Papierfragebogen beurteilt wurden, kam es bei den einzelnen Terminen zu keinen Verwechslungen oder ungültigen Fragebögen aufgrund des falschen Mediums. Die Teilnehmerzahlen (und damit auch die Beteiligungsquoten) konnten so genauer bestimmt werden als in den Auswertungsdesigns 2 und 4 am Ende des Semesters (vgl. 4.1.2).

5.1.3 Einfluss von Kontrollvariablen

Die Korrelationen der Beurteilungen der Grundstudiumsvorlesung mit der *Sympathie* der Studenten für den Dozenten waren im Mittel nur gering; im Hauptstudium zeigte sich dieser Effekt überhaupt nicht (Abschnitt 4.1.5.1). Insgesamt waren die Korrelationen kleiner als die Zusammenhänge, die Rindermann (2001, S. 179ff.; vgl. auch 2.1.1.2) zwischen den studentischen Ratings und dem Faktor *Popularität* fand. Ein Teil der Unterschiede entstand vermutlich durch die unterschiedliche Operationalisierung von *Sympathie für den Dozenten* und *Popularität*, so dass die Ergebnisse nicht vollständig vergleichbar sind (wie in Abschnitt 5.1.1 schon angesprochen). Außer der Beobachtung, dass sich der Zusammenhang nur im Grundstudium, aber nicht im Hauptstudium zeigte, trägt die vorliegende Studie in diesem Punkt nichts Neues zur Forschung über die Validität der studentischen Urteile bei. Grundsätzlich könnten komplexere Untersuchungsdesigns Aufschluss über die Rolle von Kontrollvariablen bei der Beurteilung der Vorlesung geben. Marsh (1987, S. 312) schlägt mehrere Strategien vor:

1. Den Vergleich von zwei Vorlesungen desselben Dozenten in verschiedenen Semestern und

2. die experimentelle Variation der betreffenden Variable.

Im ersten Fall sind in den Differenzwerten zwischen den Vorlesungen allerdings weiterhin mehrere Varianzquellen konfundiert (z.B. andere Studenten und unterschiedliche Rahmenbedingungen) und die Ergebnisse wären damit schwierig zu interpretieren. Die zweite Möglichkeit kommt nicht für alle Kontrollvariablen in Betracht: So läßt sich u.a. die Sympathie für den Dozenten nur schwer experimentell manipulieren. Diese Schwierigkeiten könnten erklären, warum die Richtung des Zusammenhangs zwischen der Sympathie für den Dozenten und den studentischen Ratings weiterhin ungeklärt ist.

Das Korrelationsmuster der *eigenen Vorbereitung* der Studierenden auf die Vorlesungen mit ihrer Bewertung der Veranstaltung unterstreicht allerdings die Validität der studentischen Urteile: Im Grundstudium hingen sie ausschließlich mit der Einschätzung des persönlichen Gewinns und des Inhalts der Vorlesung zusammen. Diese Korrelationen waren im Mittel gering, im Hauptstudium zeigte sich kein Zusammenhang. Offenbar bewerteten die Studierenden die Qualität einer Vorlesung weitgehend unabhängig von ihrem eigenen Lerneifer. Wenn sie es doch taten, dann in solchen Bereichen, in denen Zusammenhänge theoretisch plausibel sind (5.1.1).

Einen neuen Beitrag in der Diskussion über die Validität der studentischen Beurteilungen stellen die Korrelationen mit der momentanen *Befindlichkeit* der Studierenden dar. Obwohl die momentane Stimmung und Befindlichkeit durchaus kognitive Bewertungsprozesse beeinflussen und auch verzerren können (vgl. z.B. Clore et al., 2001), hing die Stimmung der Studierenden beim Ausfüllen des Fragebogens im Allgemeinen nicht mit ihrem Urteil über die Veranstaltung zusammen. Forgas (2001) betont die große Bedeutung der Situation bei der Urteilsfindung auf die Art der Verarbeitungsstrategie. Diese bestimmt das Ausmaß des Einflusses, den die Befindlichkeit auf das Urteil ausüben kann. Obwohl die Ergebnisse dieser Studie einen sehr geringen Einfluss der Befindlichkeit auf die Urteile nahe legen, sollten zur Bestimmung der tatsächlichen kognitiven Verarbeitungsstrategie in zukünftigen Studien weitere Variablen erhoben werden, z.B. die persönliche Wichtigkeit der Evaluation, Motivationsvariablen und die wahrgenommene Situation.

Insgesamt waren die Korrelationen der Kontrollvariablen mit den studentischen Ratings gering. Rindermann (2001, S. 180) beurteilt Zusammenhänge dieser Größenordnung als unkritisch: „Bei nicht vorhandenen Zusammenhängen (Korrelationen zwischen r=-.20 und r=-.20) kann ein Verzerrungseffekt [...] ausgeschlossen werden“. Die Ergebnisse dieser Studie bestätigen also im Ganzen die Validität studentischer Urteile.

5.1.4 Feedback an die Dozenten

Ein weiterer Vorschlag zur besseren Durchführung der Evaluation bezieht sich auf die Art des Feedbacks an die Dozenten: Neben Mittelwert und Streuung sollte die Häufigkeitsverteilung der Items mit angegeben werden (wie z.B. bei Marsh, 1982). Diese erleichert die Interpretation der deskriptiven Kennwerte, wenn die Ergebnisse stark variieren – was sie in dieser Studie allerdings nicht taten (vgl. auch Anhang M).

5.1.5 Fehlende Werte und komplexere Auswertungsstrategien

Ein sehr großes Problem war die hohe Zahl fehlender Werte: Nur von wenigen Studierenden lagen vollständige Daten über alle Messzeitpunkte vor. Damit waren trotz des großen Datensatzes komplexere und aussagekräftigere Auswertungen nicht möglich, z.B. mit Hilfe linearer Strukturgleichungsmodelle oder die Analyse der individuellen Variation der Beurteilungen unterschiedlicher Vorlesungen[17].

5.2 Praktische Vorschläge zur Lehrveranstaltungsevaluation (im Internet)

Trotz der Probleme durch fehlende Werte in dieser Studie könnte die Durchführung der Lehrveranstaltungsevaluation im Internet langfristig helfen, komplexere Auswertungen zu ermöglichen. Sobald einmal ein Fragebogensystem im Internet zur Lehrveranstaltungsevaluation (wie es z.B. in dieser Untersuchung verwendet wurde) an einem Hochschulinstitut aufgebaut ist, wird der Einsatz dieses Werkzeugs für die Dozenten der Lehrveranstaltungen sehr einfach, weil die manuelle Eingabe der Fragebögen entfällt und das Ergebnis sofort verfügbar ist (Abschnitt 2.2.2). Das erleichtert folgende Auswertungen:

- Der Einsatz des Fragebogensystems über mehrere Semester wird erleichtert, so dass Vergleichswerte für gleiche Vorlesungen desselben Dozenten einfacher gewonnen werden können. Die studentischen Bewertungen einer Vorlesung lassen sich durch diese Längsschnittdaten besser einschätzen.
- Mit Hilfe des Internets ist es möglich, die Bewertungen verschiedener Vorlesungen von unterschiedlichen Fächern, Instituten und Hochschulen zentral zu erfassen, z.B. als Dienstleis-

[17] Eine allgemeine Analyse verschiedener mit dem HILVE gewonnenen studentischen Beurteilungen von Vorlesungen und Seminaren durch lineare Strukturgleichungsmodelle findet sich bei Rindermann (2001, Kap. 10.3).

tung für die Universitäten. Dadurch stünden Entwicklern und Anwendern der Fragebögen breitere Normen (Querschnittsdaten) sehr einfach zur Verfügung, so dass sich die Ergebnisse leichter interpretieren ließen (vgl. Hänsgen, 1999).

Allerdings stellen diese Auswertungen erhebliche Ansprüche an den Datenschutz und die Sicherheit des Fragebogensystems. Es muss insbesondere den Schutz der Privatsphäre der Dozenten und die Integrität der Daten gewährleisten (ausführlich dazu Schneier, 2000, Kap. 5). Die Integrität oder Unverfälschtheit der Daten wird umso wichtiger, je stärker Personalentscheidungen (z. B. Dozentenauswahl, Berufungen oder Beförderungen) von diesen Daten abhängen. Durch Sicherungsmaßnahmen, etwa die Verschlüsselung der Daten bei der Übermittlung und in der Datenbank, steigen jedoch häufig die Kosten der Evaluation. Darum sollte man sie bereits bei der Planung der Evaluation von Beginn an berücksichtigen.

Neben der Rückmeldung, die Dozenten durch den Einsatz der studentischen Lehrveranstaltungsevaluation über die Effektivität ihrer Lehre erhalten können, sollte den Dozenten die Möglichkeit zur Weiterbildung angeboten werden. Gerade bei ungünstig beurteilten Veranstaltungen reicht das bloße Feedback zu einer Verbesserung der Lehre nicht aus (Rindermann, 2001; vgl. auch ebd., S. 249):

> Studien, in denen nur Evaluationen durchgeführt (ohne Feedback) oder Ergebnisse ohne spezifische Diskussionsmethoden und Beratungs- und Weiterbildungsangebote rückgemeldet [...] wurden, zeigen, dass entweder kein [...] oder nur ein vergleichsweise geringer Verbesserungseffekt durch zusätzliches Feedback zu erreichen ist. (S. 230)

Die studentischen Beurteilungen sollten also generell nicht nur für eine summative Bewertung am Semesterende, sondern ebenso für eine formative Evaluation der Lehre genutzt werden (in dieser Untersuchung bot die kontinuierliche wöchentliche Rückmeldung an die Dozenten die Chance dafür). Die möglicherweise erforderlichen Weiterbildungsangebote sollten von der Universität oder externen Dienstleistern angeboten werden, damit die Evaluation nachhaltigen Nutzen haben kann (vgl. z. B. Willems, Gijselaers & de Bie, 1994, Kap. 4). Diesem Zweck könnte auch die „Berücksichtigung der Lehrqualifikation bei Berufungen und Bleibeverhandlungen“ (ebd., S. 264) dienen.

Sofern man aber bei Personalentscheidungen die Ergebnisse von studentischen Vorlesungsbeurteilungen heranziehen möchte, muss man auch die Rahmenbedingungen der Veranstaltung, des Instituts und der Universität berücksichtigen (vgl. Abschnitt 2.1.3.3 und die Veranstaltungsvariablen in Tabelle 4 auf S. 19). Das Ziel einer Evaluation bestimmt ihr Design und die Wahl der Be-

urteilungskriterien (ausführlich dazu Patton, 1997). Die studentische Beurteilung von Lehrveranstaltungen können für diesen Zweck nur einen, wenn auch wichtigen, Teil der Informationen liefern.

Zusammenfassung

In dieser Studie wurde die Äquivalenz zwischen der studentischen Lehrveranstaltungsevaluation per Internet und Papier- und Bleistiftverfahren geprüft. Gleichzeitig wurden weitere mögliche Einflussgrößen berücksichtigt, die diese Äquivalenz gefährden könnten, nämlich die Vertrautheit mit Computern und dem Internet sowie mögliche systematische Selektionseffekte durch unterschiedliche Beteiligungsquoten in den beiden Medien. Darüber hinaus wurden generelle, vom Medium unabhängige Selektionseffekte beim Einsatz der studentischen Lehrveranstaltungsevaluation untersucht, indem Teilnehmer und Nicht-Teilnehmer an der Evaluation hinsichtlich ihrer Prüfungsnoten verglichen wurden. Die Literatur zur studentischen Lehrveranstaltungsevaluation (per Papierfragebogen) zeigt die generelle Robustheit der studentischen Beurteilungen der Lehre gegenüber einer Reihe von denkbaren Störvariablen. Neben der klassischen Kontrollvariable der Sympathie für den Dozenten wurden insbesondere die eigene Vorbereitung der Studierenden und die momentane Befindlichkeit beim Ausfüllen der Fragebögen berücksichtigt.

Untersucht wurden zwei Methodenvorlesungen im Fach Psychologie an der Universität Trier über das Sommersemester 2002 hinweg. Die Ergebnisse der beiden Medien erwiesen sich als äquivalent, obwohl weniger Studierende an den Befragungen im Internet teilnahmen. Die individuelle Vertrautheit mit Computern und dem Internet hatte darüber hinaus keinen Einfluss auf die Teilnahmehäufigkeit. Die Sympathie für den Dozenten wies eine geringe Korrelation mit verschiedenen Beurteilungsmaßen auf, wobei über die Richtung des Zusammenhangs keine Aussage möglich ist; dagegen zeigte die momentane Befindlichkeit keinen Zusammhang mit den Beurteilungen. Das Muster der Zusammenhänge zwischen eigener Vorbereitung auf die Vorlesung und den Urteilen über ihre Qualität sprach für die Validität dieser Urteile. Studierende, die nicht an der Evaluation (und vermutlich auch nicht an der Vorlesung) teilnahmen, erzielten im Mittel schlechtere Prüfungsnoten.

Die Durchführung der studentischen Lehrveranstaltungsevaluation im Internet gefährdet nicht ihre Validität. Stattdessen erleichtert sie die Erfassung, Auswertung und Interpretation der Beurteilungen. Diesen Vorteile wiegen umso schwerer, je mehr Vorlesungen beurteilt werden.

Literatur

Hochschulrahmengesetz, Deutscher Bundestag (1998).

Abrami, P. C. & d'Apollonia, S. (1999). Current concerns are past concerns. *American Psychologist, 54*(7), 519-520.

Batinic, B. (2001). *Fragebogenuntersuchungen im Internet*. Aachen: Shaker Verlag.

Benninghaus, H. (2002). *Deskriptive Statistik. Eine Einführung für Sozialwissenschaftler.* (9., überarb. Aufl.). Wiesbaden: Westdeutscher Verlag.

Bortz, J. (1993). *Statistik für Sozialwissenschaftler* (4., vollst. überarb. Aufl.). Berlin: Springer.

Bortz, J. & Döring, N. (2002). *Forschungsmethoden und Evaluation* (3. Aufl.). Berlin: Springer.

Bortz, J., Lienert, G. A. & Boehnke, K. (2000). *Verteilungsfreie Methoden in der Biostatistik* (2., korr. u. aktual. Aufl.). Berlin: Springer.

Bosnjak, M. (2002). *(Non)Response bei Web-Befragungen - Auswahl, Erweiterung und empirische Prüfung eines handlungstheoretischen Modells zur Vorhersage und Erklärung des Partizipationsverhaltens bei Web-basierten Fragebogenuntersuchungen.* Universität Mannheim, Mannheim.

Bosnjak, M. (2003). Teilnahmeverhalten bei Web-Befragungen - Nonresponse und Selbstselektion. In A. Theobald, M. Dreyer & T. Starsetzki (Hrsg.), *Online-Marktforschung* (2., vollst. überarb. u. erw. Aufl.). Wiesbaden: Gabler.

Bosnjak, M. & Batinic, B. (1999). Determinanten der Teilnahmebereitschaft an internet-basierten Fragebogenuntersuchungen am Beispiel E-Mail. In B. Batinic, A. Werner, L. Gräf & W. Bandilla (Hrsg.), *Online Research* (Bd. 1). Göttingen: Hogrefe.

Brandenburg, R. T. & Remmers, H. H. (1927). A rating scale for instructors. *Educational Administration and Supervision, 13*, 399-406.

Cashin, W. E. (1990). *Student ratings of teaching: Recommendations for use* (IDEA paper No. 22). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E. (1995). *Student ratings of teaching: The Research revisited* (IDEA paper No. 32). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Centra, J. A. (1993). *Reflective faculty evaluation: enhancing teaching and determining faculty effectiveness*. San Francisco: Josey-Bass.

Clore, G. L., Wyer, R. S., Dienes, B., Gasper, K., Gohm, C. & Isbell, L. (2001). Affective Feelings as Feedback: Some Cognitive Consequences. In L. L. Martin & G. L. Clore (Hrsg.), *Theories of Mood and Cognition: A User's Handbook*. Mahwah: Erlbaum.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2. Aufl.). Hillsdale: Erlbaum.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin, 112*(1), 155-159.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the social sciences* (3. Aufl.). Mahwah: Erlbaum.

Couper, M. P. (2000). Web Surveys. A Review of Issues and Approaches. *Public Opinion Quarterly, 64*, 464-494.

d'Apollonia, S. & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*(11), 1198-1208.

Dillman, D. A. (2000). *Mail and Internet Surveys: The Tailored Design Method* (2. Aufl.). New York: Wiley.

Dwight, S. A. & Feigelson, M. E. (2000). A quantitative review of the effect of computerized testing on the measurement of social desirability. *Educational and Psychological Measurement, 60*(3), 340-360.

el Hage, N. (1996). *Lehrevaluation und studentische Veranstaltungskritik*. Bonn: Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie.

Faul, F. & Erdfelder, E. (1992). GPOWER: A priori, post-hoc and compromise power analyses for for MS-DOS (Version 2.0) [Computerprogramm]. Bonn: Psychologisches Institut der Universität Bonn.

Fisseni, H.-J. (1997). *Lehrbuch der psychologischen Diagnostik* (2., überarb. u. erw. Aufl.). Göttingen: Hogrefe.

Forgas, J. P. (2001). The Affect Infusion Model (AIM): An Integrative Theory of Mood Effects on Cognition and Judgments. In L. L. Martin & G. L. Clore (Hrsg.), *Theories of Mood and Cognition: A User's Handbook*. Mahwah: Erlbaum.

Giesen, H. (1994). Anmerkungen zu den Beiträgen in der Diskussionsgruppe „Evaluation der Lehre als Aufgabe der Psychologie“. *Empirische Pädagogik, 8*(2), 235-240.

Gillmore, G. M. & Greenwald, A. G. (1999). Using Statistical Adjustment to Reduce Biases in Student Ratings. *American Psychologist, 54*, 517-520.

Glass, G. V. (1966). Note on rank biserial correlation. *Educational and Psychological Measurement, 26*, 623-631.

Gollwitzer, M. & Schlotz, W. (2003). Das „Trierer Inventar zur Lehrveranstaltungsevaluation“ (TRIL): Entwicklung und erste testtheoretische Erprobungen. In G. Krampen & H. Zayer (Hrsg.), *Psychologiedidaktik und Evaluation IV* (S. 114-128). Bonn: Deutscher Psychologen Verlag.

Göritz, A. S., Reinhold, N. & Batinic, B. (2002). Online Panels. In B. Batinic, U.-D. Reips & M. Bosnjak (Hrsg.), *Online Social Sciences*. Seattle: Hogrefe & Huber.

Greenwald, A. G. (1997). Validity Concerns and Usefulness of Student Ratings of Instruction. *American Psychologist, 52*(11), 1182-1186.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen in der Psychologie. In G. Lüer (Hrsg.), *Allgemeine experimentelle Psychologie*. Stuttgart: Fischer.

Hänsgen, K.-D. (1999). *Computereinsatz in der Psychodiagnostik - Stand und mögliche Perspektiven* (Forschungsbericht Nr. 41). Freiburg (Schweiz): Psychologisches Institut der Universität Freiburg.

Hapuarachchi, P., March, M. & Wronski, A. (1997). Using Statistical Methods to Autocorrelated Processes to Analyze Survey Process Quality Data. In L. Lyberg, P. P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz & D. Trewin (Hrsg.), *Survey Measurement and Process Quality*. New York: Wiley.

HIS Hochschul-Informations-System GmbH (2002). *HIS Ergebnisspiegel*. Hannover: HIS.

Jaccard, J., Turrisi, R. & Wan, C. K. (1990). *Interaction Effects in Multiple Regression*. Thousand Oaks: Sage.

Joint Commitee on Standards for Educational Evaluation. (1994). *The program evaluation standards: how to assess evaluations of educational programs*. Thousand Oaks: Sage.

Klinck, D. (1998). Papier-Bleistift- versus computer-unterstützte Administration kognitiver Fähigkeitstests: Eine Studie zur Äquivalenzfrage. *Diagnostica, 44*(2), 61-70.

Knapp, F. & Heidingsfelder, M. (1999). Drop-Out-Analyse: Wirkung des Untersuchungsdesigns. In U.-D. Reips, B. Batinic, W. Bandilla, M. Bosnjak, L. Gräf, K. Moser & A. Werner (Hrsg.), *Aktuelle Online Forschung - Trends, Techniken, Ergebnisse*. Zürich: Online Press.

Krampen, G. & Zayer, H. (2000). Psychologiedidaktik und Evaluation in Deutschland gegen Ende des 20. Jahrhunderts: Heterogen, bunt, ein Patchwork - aber präsent! In G. Krampen & H. Zayer (Hrsg.), *Psychologiedidaktik und Evaluation II. Neue Medien, Psychologiedidaktik und Evaluation in der psychologischen Haupt- und Nebenfachausbildung*. Bonn: Deutscher Psychologen Verlag.

Kromrey, H. (1994). Wie erkennt man „gute Lehre“? Was studentische Vorlesungsbefragungen (nicht) aussagen. *Empirische Pädagogik, 8*(2), 153-168.

Kromrey, H. (1996a). Qualitätsverbesserung in Lehre und Studium statt sogenannter Lehrevaluation. Ein Plädoyer für gute Lehre und gegen schlechte Sozialforschung. *Zeitschrift für Pädagogische Psychologie, 10*(3-4), 153-166.

Kromrey, H. (1996b). „Gute“ oder „schlechte“ Sozialforschung? Einige notwendig scheinende Ergänzungen. *Zeitschrift für Pädagogische Psychologie, 10*(3-4), 171-173.

Kromrey, H. (2001). Studierendenbefragung als Evaluation der Lehre? Anforderungen an Methodik und Design. In U. Engel (Hrsg.), *Hochschul-Ranking*. Frankfurt: Campus.

Marsh, H. W. (1982). SEEQ: A reliable, valid and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52*, 77-95.

Marsh, H. W. (1987). Students' evaluation of university teaching: research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253-388.

Marsh, H. W., Hau, K. T., Chung, C. M. & Siu, T. L. P. (1997). Students' evaluations of university teaching: Chinese version of the Students' Evaluations of Educational Quality Instrument. *Journal of Educational Psychology, 89*(3), 568-572.

Marsh, H. W. & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching And Teacher Education, 7*(1), 9-18.

Marsh, H. W. & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187-1197.

Marsh, H. W. & Roche, L. A. (1999). Rely upon SET research. *American Psychologist, 54*(7), 517-518.

Marsh, H. W. & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology, 92*(1), 202-228.

McKeachie, W. J. (1997). Student Ratings: The Validity of Use. *American Psychologist, 52*(11), 1218-1225.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449-145.

Musch, J. & Reips, U.-D. (2000). A Brief History of Web Experimenting. In M. H. Birnbaum (Hrsg.), *Psychological Experiments on the Internet*. San Diego: Academic Press.

National Institute of Standards and Technology (2003). *NIST/SEMATECH e-Handbook of Statistical Methods*. [WWW]. Verfügbar unter: http://www.itl.nist.gov/div898/handbook [11. Juli 2003].

Object Management Group (2001). *OMG Unified Modeling Language Specification*. [WWW]. Verfügbar unter: http://www.omg.org/technology/documents/formal/uml.htm [2. Februar 2003].

O'Brian, R. G. & Kaiser, M. K. (1985). MANOVA Method for Analyzing Repeated Measure Designs: An Extensive Primer. *Psychological Bulletin, 2*, 316-333.

Patton, M. Q. (1997). *Utilization-focused evaluation: the new century text* (3. Aufl.). Thousand Oaks: Sage.

Pitkin, A. K. & Vispoel, W. P. (2001). Differences Between Self-Adapted and Computerized Adaptive Tests: A Meta-Analysis. *Journal of Educational Measurement, 38*(3), 235-247.

Polkehn, K. & Wandke, H. (1999). Web-unterstütztes Experimentieren: Das Netz im Labor? In U.-D. Reips, B. Batinic, W. Bandilla, M. Bosnjak, L. Gräf, K. Moser & A. Werner (Hrsg.), *Aktuelle Online Forschung - Trends, Techniken, Ergebnisse*. Zürich: Online Press.

Reips, U.-D. (2000). The Web Experiment Method: Advantages, Disadvantages, and Solutions. In M. H. Birnbaum (Hrsg.), *Psychological Experiments on the Internet*. San Diego: Academic Press.

Rindermann, H. (1996a). Zur Qualität studentischer Lehrveranstaltungsevaluationen: Eine Antwort auf Kritik an der Lehrevaluation. *Zeitschrift für Pädagogische Psychologie, 10*(3-4), 129-145.

Rindermann, H. (1996b). *Untersuchungen zur Brauchbarkeit studentischer Lehrevaluationen* (Psychologie Bd. 6). Landau: Empirische Pädagogik.

Rindermann, H. (1998). Übereinstimmung und Divergenz bei der studentischen Beurteilung von Lehrveranstaltungen: Methoden zu ihrer Berechnung und Konsequenzen für die Lehrevaluation. *Zeitschrift für Differentielle und Diagnostische Psychologie, 19*(2), 73-92.

Rindermann, H. (2001). *Lehrevaluation. Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts* (Psychologie Bd. 42). Landau: Empirische Pädagogik.

Rindermann, H. & Amelang, M. (1994). Entwicklung und Erprobung eines Fragebogens zur studentischen Veranstaltungsevaluation. *Empirische Pädagogik, 8*(2), 131-151.

Roche, L. A. & Marsh, H. W. (1998). Workload, grades, and students' evaluations of teaching: Clear understanding sometimes requires more patient explanations. *American Psychologist, 53*(11), 1230-1231.

Schneier, B. (2000). *Secrets and lies: digital security in a networked world*. New York: Wiley.

Schäffer, K.-A. (1996). Planung von Stichprobenerhebungen. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Hrsg.), *Handbuch Quantitative Methoden*. Weinheim: Psychologie Verlags Union.

Schumacher, J., Hinz, A., Hessel, A. & Brähler, E. (2002). Zur Vergleichbarkeit von internetbasierten und herkömmlichen Fragebogenerhebungen: Eine Untersuchung mit dem Fragebogen zum erinnerten elterlichen Erziehungsverhalten. *Diagnostica, 48*(4), 172-180.

Shapiro, S. S., Wilk, H. B. & Chen, H. J. (1965). A comparative study of various tests of normality. *Journal of the American Statistical Association, 63*, 1343-1372.

Sidiropoulou, E. (1997). Computerdiagnostik. In H.-J. Fisseni (Hrsg.), *Lehrbuch der psychologischen Diagnostik*. Göttingen: Hogrefe.

SPSS Inc. (2001). SPSS (Version 11.0) [Computerprogramm]. Chicago: SPSS Inc.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3. Aufl.). Mahwah: Erlbaum.

Testkuratorium der Föderation Deutscher Psychologenvereinigungen (1986). Richtlinien für den Einsatz elektronischer Datenverarbeitung in der psychologischen Diagnostik. *Psychologische Rundschau, 37*, 163-165.

Theall, M. & Franklin, J. (1991). Using student ratings for teaching improvement. In M. Theall (Hrsg.), *Effective practices for improving teaching*. San Francisco: Jossey-Bass.

Theobald, A. (2003). Rücklaufquoten bei Online-Befragungen. In A. Theobald, M. Dreyer & T. Starsetzki (Hrsg.), *Online-Marktforschung: Theoretische Grundlagen und praktische Erfahrungen* (2., vollst. überarb. u. erw. Aufl.). Wiesbaden: Gabler.

West, S. G., Finch, J. F. & Curran, P. J. (1995). Structural equitation models with nonnormal variables: problems and remedies. In R. H. Hoyle (Hrsg.), *Structural equitation modeling: concepts, issues, and applications.* Thousand Oaks: Sage.

Willems, J., Gijselaers, W. & de Bie, D. (1994). *Qualitätssorge in der Lehre: Leitfaden für die studentische Lehrevaluation* (R. Richter, Übers.). Neuwied: Luchterhand.

Wissenschaftsrat (1996). Empfehlungen zur Stärkung der Lehre in den Hochschulen durch Evaluation. *Bericht des Wissenschaftsrates*, 55-87.

Wottawa, H. & Thierau, H. (1998). *Lehrbuch Evaluation* (2., vollst. überarb. Aufl.). Bern: Huber.

Zentrum für Psychologische Information und Dokumentation (3. April 2003). *Online-Untersuchungen und Online-Tests*. [WWW]. Verfügbar unter: http://www.zpid.de/index.php ?wahl=products&uwahl=psycho&uuwahl=ptonlinetests [13. Mai 2003].

Anhang A: Dimensionen und Items des SEEQ

Aus: Marsh (1982). Antwortskala von 1=*very poor* bis 5=*very good*.

Learning

1. You found the course intellectually challenging and stimulating.
2. You have learned something which you consider valuable.
3. Your interest in the subject has increased as a consequence of this course.
4. You have learned and understood the subject materials in the course.

Enthusiasm

5. Instructor was enthusiastic about teaching the course.
6. Instructor was dynamic and energetic in conducting the course.
7. Instructor enhanced presentations with the use of humour.
8. Instructor's style of presentation held your interest during class.

Organisation

9. Instructor's explanations were clear.
10. Course materials were well prepared and carefully explained.
11. Proposed objectives agreed with those actually taught so you knew where course was going.
12. Instructor gave lectures that facilitated taking notes.

Group Interaction

13. Students were encouraged to participate in class discussion.
14. Students were invited to share their ideas and knowledge.
15. Students were encouraged to ask questions and were given meaningful answers.
16. Students were encouraged to express their own ideas and/or question the instructor.

Individual Rapport

17. Instructor was friendly towards individual students.
18. Instructor made students feel welcome in seeking help/advice in or outside class
19. Instructor had genuine interest in individual students.
20. Instructor was adequately accessible to students during office hours or after class.

Breadth

21. Instructor contrasted the implications of various theories.
22. Instructor presented the background or origin of ideas/concepts developed in class.
23. Instructor presented points of view other than his/her own when appropriate.
24. Instructor adequately discussed current developments in the field.

Examinations

25. Feedback on examinations/graded material was valuable.
26. Methods of evaluation student work were fair and appropriate.

27. Examinations/graded materials tested course content as emphasised by the instructor.

Assignments

28. Required readings/texts were valuable.
29. Readings, homework, etc. contributed to appreciation and understanding of subject.
30. How does this course compare with other courses you have hat at the University of Southern California?
31. How does this instructor compare with other instructors you have had at USC?

Student and Course Characteristics (Leave blank if no response applies)

32. Course difficulty, relative to other courses, was:	1=very easy/3=medium/5=very hard
33. Course workload, relative to other courses, was:	1=very light/3=medium/5=very heavy
34. Course pace was:	1=too slow/3=about right/ 5=too fast
35. Hours/weeks required outside of class: 1=0 to 5 2=2 to 5	3=5 to 7 4=8 to 12 5=over 12
36. Level of interest in the subject prior to this course:	1=very low/ 3=medium/ 5= very high
37. Overall Grade Point Average at USC 1=below 2,5 2=2,5 to 3,0	3=3,0 to 3,4 4=3,5 to 3,7 5=above 3,7
38. Expected grade in this course	1=F/ 2=D/ 3=C/ 2=B/ 1=A
39. Reason for taking this course: 1=major require 2=major elective	3=general ed require 4=minor/related field 5=general interest only
40. Year in school: 1=Freshman 2=Sophomore	3=Junior 4=Senior 5=Postgraduate
41. Major department: 1=SocSci/Comm 2=NatSci/Math. 3=Humanities 4=Business 5=Education	6=Engineering 7=Perf Arts 8=Pub Affairs 9=Other 10=Undeclared/undecided

Anhang B: Dimensionen und Items des HILVE-I

Aus: Rindermann (1996a, 2001). Antwortskala von 1=*trifft zu* bis 7=*trifft völlig zu*.

Struktur

24. Der inhaltliche Aufbau der Veranstaltung ist gut organisiert.
25. Die Veranstaltung ist gut organisiert.

Auseinandersetzung mit dem Thema

3. Der Stoff wird anhand von Beispielen aus dem Alltag oder der Praxis veranschaulicht.
4. Die Relevanz/Bedeutung/Nutzen der behandelten Themen wird nahe gelegt.
5. Ein Bezug zwischen Theorie und Praxis wird hergestellt.
6. Ich werde zum Mitdenken motiviert.
7. Zur kritischen Auseinandersetzung mit den behandelten Themen wird angeregt.

Lehrkompetenz

8. Der Dozent kann komplizierte Sachverhalte verständlich machen.
9. Die Dozentin/Der Dozent wirkt gut vorbereitet.
10. Die Dozentin/Der Dozent spricht frei und anregend.

Dozentenengagement

16. Der Dozent zeigt Engagement in seiner Lehrtätigkeit und versucht Begeisterung zu vermitteln.
17. Die Dozentin/Der Dozent nimmt die Lehre wichtig.
18. Dem Dozenten ist es wichtig, dass die Studierenden etwas in der Veranstaltung lernen können.

Klima

14. Der Dozent ist im Umgang mit den Studierenden freundlich und aufgeschlossen.
15. Die Dozentin/Der Dozent ist kooperativ.

Interessantheit

16. Die Veranstaltung ist interessant.
17. Die Veranstaltung zieht sich schleppend dahin.

Überforderung

18. Die Stoffmenge kann ich noch verkraften.
19. Das Tempo der Veranstaltung ist zu schnell.
20. Ich verstehe alles. *(wird umkodiert)*
21. Höhe der Anforderungen
    (1= zu niedrig, 4=angemessen, 7=zu hoch)

Lernen

22. Ich lerne viel in der Veranstaltung
23. Ich lerne etwas Sinnvolles und Wichtiges.

Thema

24. Das Thema der Veranstaltung interessiert mich.

Referate

25. Die Referate der Studierenden sind interessant.
26. Die Referate sind strukturiert und verständlich.
27. Die Referate sind nützlich und wertvoll.
28. Die Referenten werden durch den/die Dozent/in adäquat ergänzt.

Fleiß

29. Ich bereite mich auf die Veranstaltung vor oder bereite sie nach (z.B. durch Lesen der angegebene Literatur).
30. Mein Arbeitsaufwand für die Veranstaltung ist verglichen mit anderen Veranstaltungen sehr hoch.

Beteiligung

31. An der Veranstaltung beteilige ich mich durch Wortbeiträge.
32. Beim Einbringen eigener Beiträge fühle ich mich frei und äußerungsfähig.
    (Falls keine Beiträge bitte frei lassen.)

Diskussion

33. Es finden ausreichend Diskussionen statt.
34. Die Diskussionen in der Veranstaltung sind produktiv.
    (Falls keine Diskussionen bitte frei lassen)

Items 35 bis 39 sind für frei hinzufügbare veranstaltungsspezifische Items reserviert.

Allgemeineinschätzung

40. Der Besuch der Veranstaltung lohnt sich.
41. Die Veranstaltung fördert mein Interesse am Studienfach.
42. Wenn man alles in einer Note zusammen fassen könnte, würde ich der Veranstaltung die folgende Note geben:
    (Notenskala von sehr gut=1 bis ungenügend=6, es sind auch Bruchnoten wie 3,5 oder 2+ möglich).

Zusatzfragen

43. Grund für den Besuch der Veranstaltung (Mehrfachantworten möglich):
    A) Pflichtveranstaltung, Schein, Prüfungsrelevanz
    B) Wegen der Dozentin/dem Dozenten
    C) Aus Interesse, Thema
    D) und/oder: _____

Offene Fragen:

- Was ist besonders gut an der Veranstaltung?
- Was ist schlecht?
- Verbesserungsvorschläge:
- Kommentar zum Fragebogen und zur Erhebung:

Anhang C: Dimensionen und Items des HILVE-II

Aus: Rindermann (2001). Antwortskala von 1=*trifft zu* bis 7=*trifft völlig zu*.

Struktur beschreibt Aufbau und Organisation der Veranstaltung.

1. Der inhaltliche Aufbau der Veranstaltung ist logisch/nachvollziehbar.
2. Die Veranstaltung ist gut organisiert.

Auseinandersetzung thematisiert die erläuternde Behandlung des Stoffes.

3. Der Stoff wird anhand von Beispielen veranschaulicht.
4. Die Bedeutung/Nutzen der behandelten Themen wird vermittelt.
5. Ein Bezug zwischen Theorie und Praxis/Anwendung wird hergestellt.

Verarbeitung: Der Stoff wird nicht nur wissensmäßig vermittelt, sondern Studierende werden zum Mitdenken motiviert, Problemstellungen werden kritisch herausgearbeitet.

6. Zum Mitdenken und Durchdenken des Stoffes wird angeregt.
7. Die behandelten Themen werden kritisch/von verschiedenen Seiten beleuchtet.

Lehrkompetenz fragt danach, ob der Dozent didaktisch überzeugt.

8. Die Dozentin/der Dozent spricht verständlich und anregend.
9. Die Dozentin/der Dozent kann Kompliziertes verständlich machen.
10. Die Dozentin/der Dozent fasst regelmäßig den Stoff zusammen.
11. Die Dozentin/der Dozent wirkt gut vorbereitet.

Dozentenengagement erhebt Motivationsvariablen.

12. Die Dozentin/der Dozent engagiert sich bei der Lehrtätigkeit und versucht Begeisterung zu vermitteln.
13. Dem/Der Dozent/in ist es wichtig, dass die Teilnehmer etwas lernen.
14. Die Dozentin/der Dozent motiviert die Teilnehmer.

Mit *Klima* wird die Atmosphäre in der Veranstaltung erhoben.

15. Die Dozentin/der Dozent ist im Umgang mit den Studierenden freundlich.
16. Die Dozentin/der Dozent ist kooperativ und aufgeschlossen.

Interessantheit erhebt im Gegensatz zur Interessantheit des Themas die der Veranstaltung.

17. Die Veranstaltung wird in interessanter Form gehalten.
18. Die Veranstaltung zieht sich schleppend dahin.

Unter *Thema* sind zwei eher heterogene Inhalte zusammengefasst.

19. Ich habe mich schon *vor* dem Kurs sehr für die Themen interessiert.
20. Das Thema des Kurses als solches ist relevant (Beruf/Praxis/Gesellschaft).

Auch *Redundanz* erhebt zwei verschiedene Aspekte.

21. Mein Vorwissen: 1=zu wenig, um dem Kurs folgen zu können, 4=genau richtig, 7=alles schon bekannt gewesen, Besuch überflüssig
22. Es treten oft unnötige inhaltliche Überschneidungen mit anderen Kursen auf.

Überforderung oder *Anforderungen* soll messen, ob die Teilnehmer stark in den in der Stoffschwierigkeit, der Stoffmenge oder der Geschwindigkeit gefordert werden.

23. Schwere des Stoffes als solches: viel zu leicht=1/ genau richtig=4/ viel zu schwer=7
24. Umfang des Stoffes: viel zu wenig=1/ genau richtig=4/ viel zu viel=7
25. Das Tempo des Kurses: viel zu langsam=1/genau richtig=4/ viel zu schnell=7
26. Die Anforderungen sind viel zu niedrig=1/ genau richtig=4/ viel zu hoch=7

Lernen-quantitativ erfasst in der Selbsteinschätzung, ob die Studenten etwas in der Veranstaltung lernen können.

27. Ich lerne viel in der Veranstaltung.
28. Mein Wissensstand ist nach der Veranstaltung wesentlich höher als vorher.

Lernen-quantitativ erfasst den Inhaltsaspekt. Die beiden Lerndimensionen fragen insgesamt nach der Effektivität der Lehre.

29. Ich verfüge über ein grundlegenderes Verständnis als vor dem Kurs.
30. Ich lerne etwas Sinnvolles und Wichtiges.

Mit den folgenden Items wird Feedback auf Studentenbeiträge und *Betreuung* durch den Dozenten erhoben. In Vorlesungen fallen diese in den Rohwerten geringer aus.

31. Die Lehrkraft gibt auf Beiträge der Teilnehmer hilfreiches Feedback.
32. Außerhalb der Veranstaltung findet eine gute Betreuung statt.

Referate beurteilt die Qualität der Studenten-Referate hinsichtlich Didaktik, Inhalte und Lerneffektivität. Item 33 fragt nach der Moderation der Referate durch den Dozenten und wird nicht zur Dimensionsbildung herangezogen.

33. Die Lehrkraft ergänzt Referate viel zu wenig=1/ genau richtig=4/ viel zu viel=7
34. Die Vorgehensweise und Darbietung der Inhalte ist ansprechend/gut.
35. Die fachlich-inhaltliche Qualität der Referate ist hoch.
36. Ich lerne viel durch die Referate anderer Teilnehmer.

Fleiß/Arbeitshaltung soll die Mitarbeit der Studenten erheben.

37. Ich bereite den Kurs (z. B. Texte lesen) vor oder nach (1=sehr wenig, 7=sehr viel)
38. Mein Arbeitsaufwand ist verglichen mit den anderen Veranstaltungen hoch.
39. Mein üblicher Arbeitsaufwand für den Kurs pro Woche (nicht Kursdauer): Stunden, Minuten.

Interaktionsmanagement beurteilt die Förderung und Moderation von Teilnehmer-Interaktionen.

40. Der Dozent/die Dozentin fördert Fragen und aktive Mitarbeit.
41. Diskussionen werden gut geleitet (Anregung von Beiträgen, Eingehen auf Beiträge, Zeiteinteilung, Bremsen von Vielrednern).

Beteiligung

42. Ich beteilige mich mit Wortbeiträgen/bei Diskussionen (1=nie, 4=manchmal, 7=oft)
43. Beim Einbringen eigener Beiträge fühle ich mich frei und äußerungsfähig. (Falls nie Beiträge bitte frei lassen.)

Mit *Kommunikativen Unterrichtsformen* wird der Einsatz interaktiver Lehrformen (z. B. Diskussion oder Gruppenarbeit) erhoben.

44. Es finden ausreichend Diskussionen statt.
45. Es werden kommunikative Lehrformen eingesetzt (z. B. Gruppenarbeit).

Anomie erhebt Aspekte formaler Disziplin (Unruhe, Fehlzeiten) im Seminar.

46. Unruhe, Reden oder Störungen durch Teilnehmer beeinträchtigen den Kurs.
47. Ich habe an mehreren Sitzungen gefehlt (0=keinmal, 1=1x, usw.). Gründe in Worten:

Interessenförderung

48. Die Veranstaltung fördert mein Interesse am Studium.
49. Der Kurs motiviert dazu, sich selbst mit den Inhalten zu beschäftigen.

Mit *Allgemeineinschätzung* wird schließlich die allgemeine Bewertung einer Veranstaltung erhoben. Die Notengebung für die Veranstaltung spricht sie direkt an.

50. Der Besuch der Veranstaltung lohnt sich.
51. Wenn man alles in einer Note zusammen fassen könnte, würde ich der Veranstaltung die folgende Note geben: (Notenskala von sehr gut=1,0 bis ungenügend=6,0, inkl. Zwischennoten).

Zusatzfragen:

- Abschlusszeugnisnotenschnitt (Abitur/Matura)
- Ich werde am Ende zu den schwächeren (1), mittl. (4) oder sehr guten Studenten (7) gehören.

*Offene Fragen***:**

- Was ist besonders gut an der Veranstaltung?
- Was ist schlecht?
- Verbesserungsvorschläge: Kommentar zum Fragebogen / Erhebung:

Anhang D: Dimensionen und Items des TRIL

Antwortskala von 0=*trifft nicht zu* bis 3=*trifft zu*.

Struktur und Didaktik

1. Die Lehrziele waren klar und nachvollziehbar.
2. Der Inhaltliche Aufbau der Veranstaltung war den Zielen angemessen.
3. Die gesetzten Lehrziele sind erreicht worden.
4. Der Dozent war stets gut vorbereitet.
5. Er hat didaktische Hilfsmittel (z. B. Folien, Tafelbilder) sinnvoll eingesetzt.
6. Er hat komplizierte Dinge strukturiert erklärt.

Anregung und Motivation

7. Der Dozent wirkte in der Veranstaltung engagiert.
8. Er hat anregende und akustisch verständlich gesprochen.
9. Er hat die Veranstaltung interessant gestaltet.
10. Er hat mich motiviert, konzentriert bei der Sache zu bleiben.
11. Die Veranstaltung zog sich schleppend dahin.

Interaktion und Kommunikation

12. In der Veranstaltung herrschte ein offenes Klima für eigene Beiträge.
13. Es fanden ausreichend Diskussionen statt.
14. Die Diskussionen der Studierenden waren produktiv.
15. Fragen und Beiträge waren stets willkommen.

Persönlicher Gewinn durch die Veranstaltung

16. Das Thema der Veranstaltung hat mich interessiert.
17. Die behandelten Themen waren für mich relevant.
18. Ich habe in dieser Veranstaltung etwas Sinnvolles und Wichtiges gelernt.
19. Mein Verständnis für das Studienfach hat sich durch die Veranstaltung weiterentwickelt.

Anwendungsbezug

20. Es wurden Bezüge zwischen Theorie und Praxis aufgezeigt.
21. Der Dozent hat den Stoff an lebensnahen Beispielen veranschaulicht.
22. Er hat zur kritischen Auseinandersetzung mit den behandelten Themen angeregt.
23. Die behandelten Inhalte waren lebensfern.
24. Ich kann die Veranstaltung weiterempfehlen.
25. Ich kann den Veranstaltungsleiter weiterempfehlen.
26. Der Besuch der Veranstaltung lohnt sich.

Referate

27. Referate waren grundsätzlich ein nützlicher Bestandteil dieser Veranstaltung.
28. Die gehaltenen Referate waren strukturiert und verständlich.
29. Die Referate waren interessant.
30. Die Referenten/innen wurden durch den Dozenten adäquat ergänzt.

Weitere Fragen

31. Ich habe mich auf die Veranstaltung vorbereitet (z.B. durch Lesen der Literatur).
32. Ich habe die einzelnen Sitzungen nachbereitet (z.B. durch Diskussion mit Kommilitoninnen/Kommilitonen bzw. Lesen der Literatur).
33. Bei Fragen u.ä. war der Dozent auch außerhalb der Veranstaltung ansprechbar.

Arbeitsanforderungen

Die gestellten Anforderungen waren... -2=zu niedrig 0=genau richtig +2=zu hoch

Ergänzende Angaben

- Ich habe in der Veranstaltung ... gefehlt (zutreffendes bitte ankreuzen):
  1-2 mal 3-4 mal 5-6 mal mehr als 6 mal
- Grund für den Besuch der Veranstaltung (Mehrfachantworten möglich):
  A) Pflichtveranstaltung
  B) Schein
  C) Prüfungsrelevanz
  D) Dozent
  E) Interesse am Thema
  F) und/oder: _____

Abschließende offene Fragen

- Was ist besonders gut an der Veranstaltung?
- Was ist schlecht? Verbesserungsvorschläge?
- Kommentar zum Fragebogen und zur Erhebung:

Anhang E: Papierversion des Fragebogens

Vorlesungsevaluation *Forschungsmethoden II*

Vielen Dank, dass Sie an unserer Untersuchung teilnehmen! Bitte geben Sie zunächst das Datum der Veranstaltung und anschließend ihre Matrikelnummer an. Kreuzen Sie dann bitte bei den Fragen diejenige Aussage an, der Sie am meisten zustimmen können.

Datum der Veranstaltung: _____

Matrikelnummer: _____

Im Moment fühle ich mich...

	überhaupt nicht				sehr
glücklich	○	○	○	○	○
wohl	○	○	○	○	○
unglücklich	○	○	○	○	○
unzufrieden	○	○	○	○	○

Im Moment fühle ich mich...

	überhaupt nicht				sehr
zufrieden	○	○	○	○	○
schlecht	○	○	○	○	○
gut	○	○	○	○	○
unwohl	○	○	○	○	○

	überhaupt nicht verständlich	eher nicht verständlich	etwas verständlich	ziemlich gut verständlich	sehr gut verständlich
Wie verständlich waren für Sie die vorgetragenen Inhalte insgesamt?	○	○	○	○	○

	überhaupt nicht verständlich	eher nicht verständlich	etwas verständlich	ziemlich gut verständlich	sehr gut verständlich
Wie verständlich waren die verwendeten Lehrmittel (z.B. Folien, Präsentationen, Beispiele)?	○	○	○	○	○

	zu langsam	eher langsam	genau richtig	ziemlich schnell	zu schnell
Wie schätzen Sie das Tempo ein, das der Vermittlung zugrunde lag?	○	○	○	○	○

	überhaupt nicht	eher nicht	etwas	ziemlich	sehr
Wie sehr wünschen Sie, dass die Lehrveranstaltung anders gestaltet werden sollte?	○	○	○	○	○

	trifft überhaupt nicht zu	trifft eher nicht zu	trifft etwas zu	trifft eher zu	trifft voll und ganz zu
Alles in allem hat sich der Besuch der Lehrveranstaltung für mich gelohnt.	○	○	○	○	○

	trifft überhaupt nicht zu	trifft eher nicht zu	trifft etwas zu	trifft eher zu	trifft voll und ganz zu
Meiner Meinung nach war die Lehrveranstaltung inhaltlich gut strukturiert.	○	○	○	○	○

	trifft überhaupt nicht zu	trifft eher nicht zu	trifft etwas zu	trifft eher zu	trifft voll und ganz zu
Ich habe mich auf die Stunde vorbereitet.	○	○	○	○	○

Vielen Dank!

Anhang F: Bildschirmfotos des Fragebogens

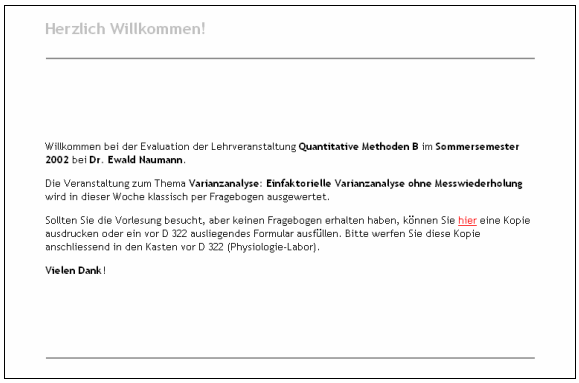


Abbildung F.1: Der Begrüßungsbildschirm.

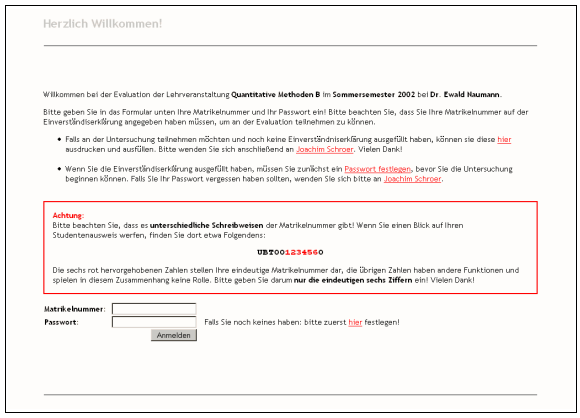


Abbildung F.2: Anmelden für die Evaluation.

Abbildung F.3: Fehlermeldung bei falschem Passwort oder Benutzernamen. Eine ähnliche Meldung erscheint, wenn ein Teilnehmer in der aktuellen Woche schon an der Befragung teilgenommen hat.

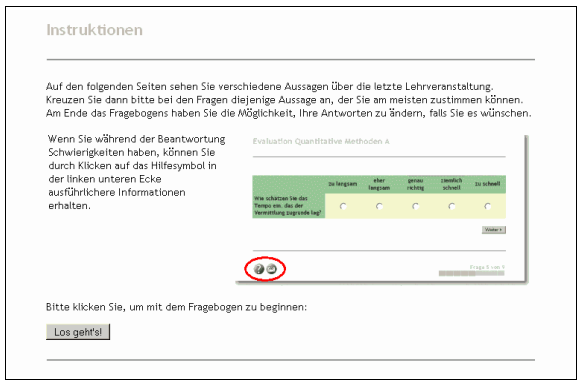


Abbildung F.4: Instruktionen und Hilfe vor der Befragung.

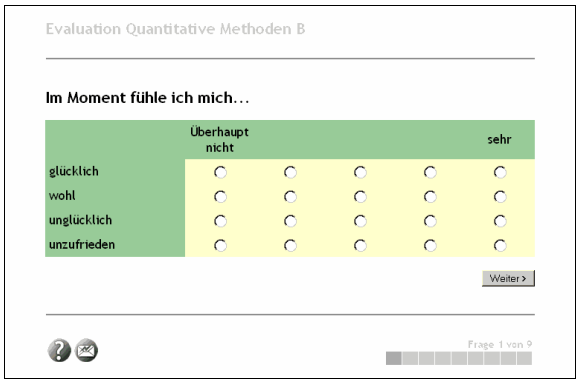


Abbildung F.5: Befindlichkeitsfragen (1).

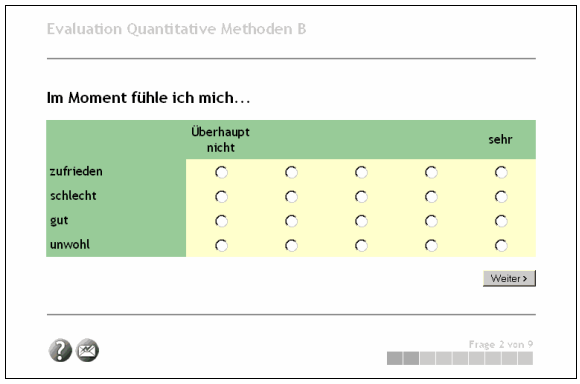


Abbildung F.6: Befindlichkeitsfragen (2).

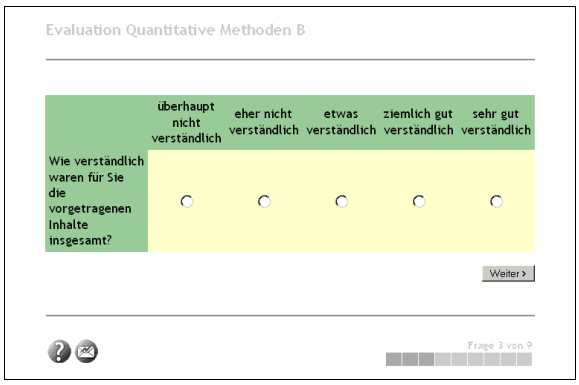


Abbildung F.7: Verständlichkeit der Inhalte.

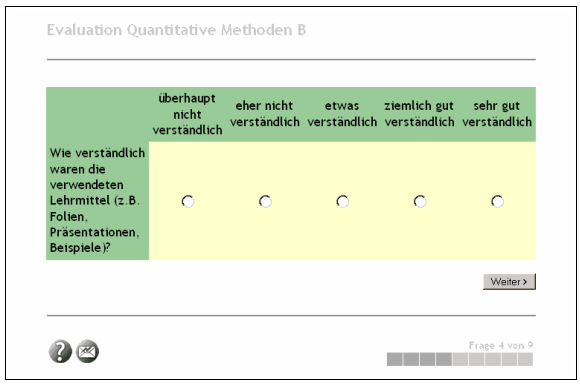


Abbildung F.8: Verständlichkeit der Lehrmittel.

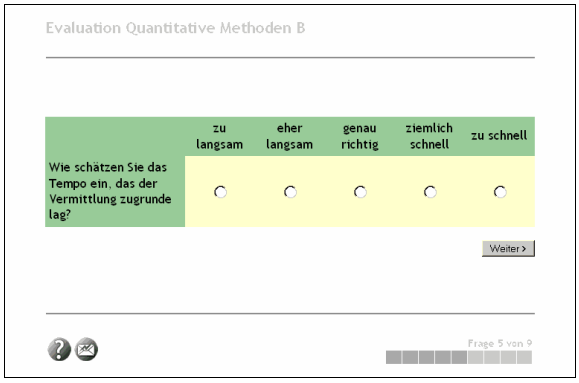


Abbildung F.9: Tempo.

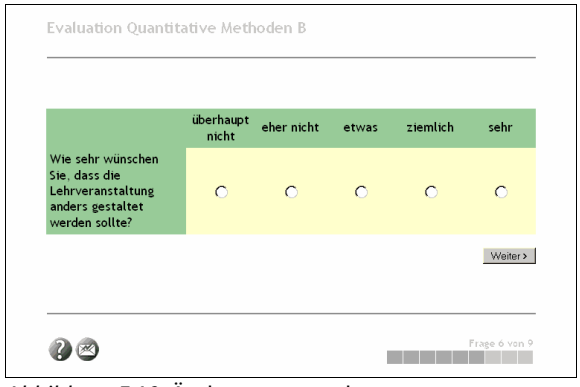


Abbildung F.10: Änderungswunsch.

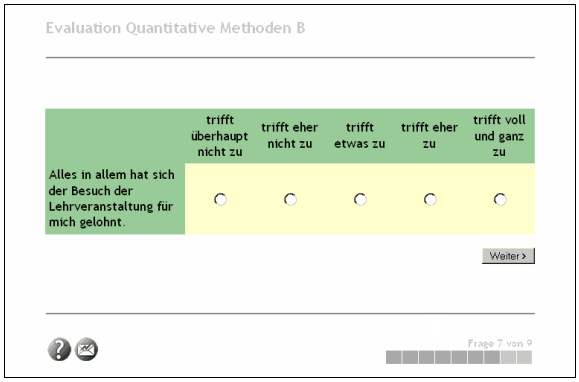


Abbildung F.11: Insgesamt lohnend.

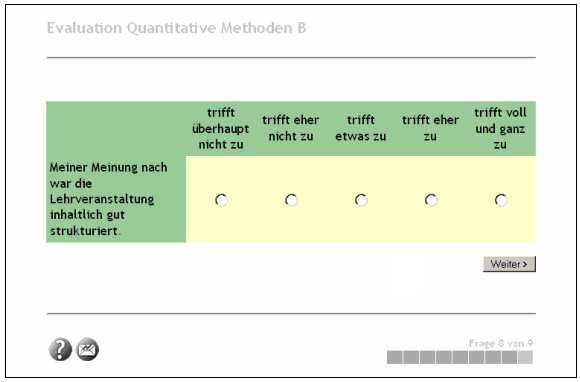


Abbildung F.12: Strukturierung.

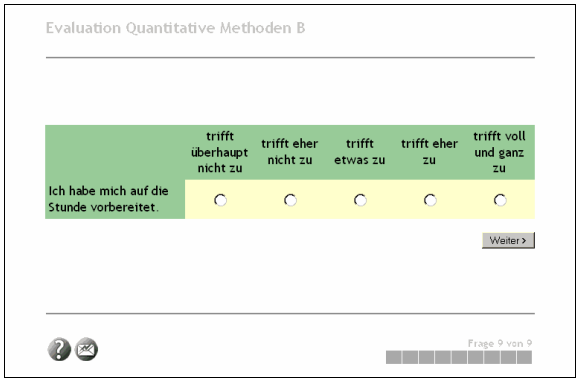


Abbildung F.13: Eigene Vorbereitung auf die Stunde.

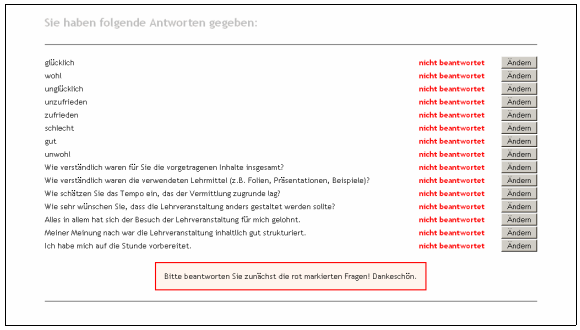
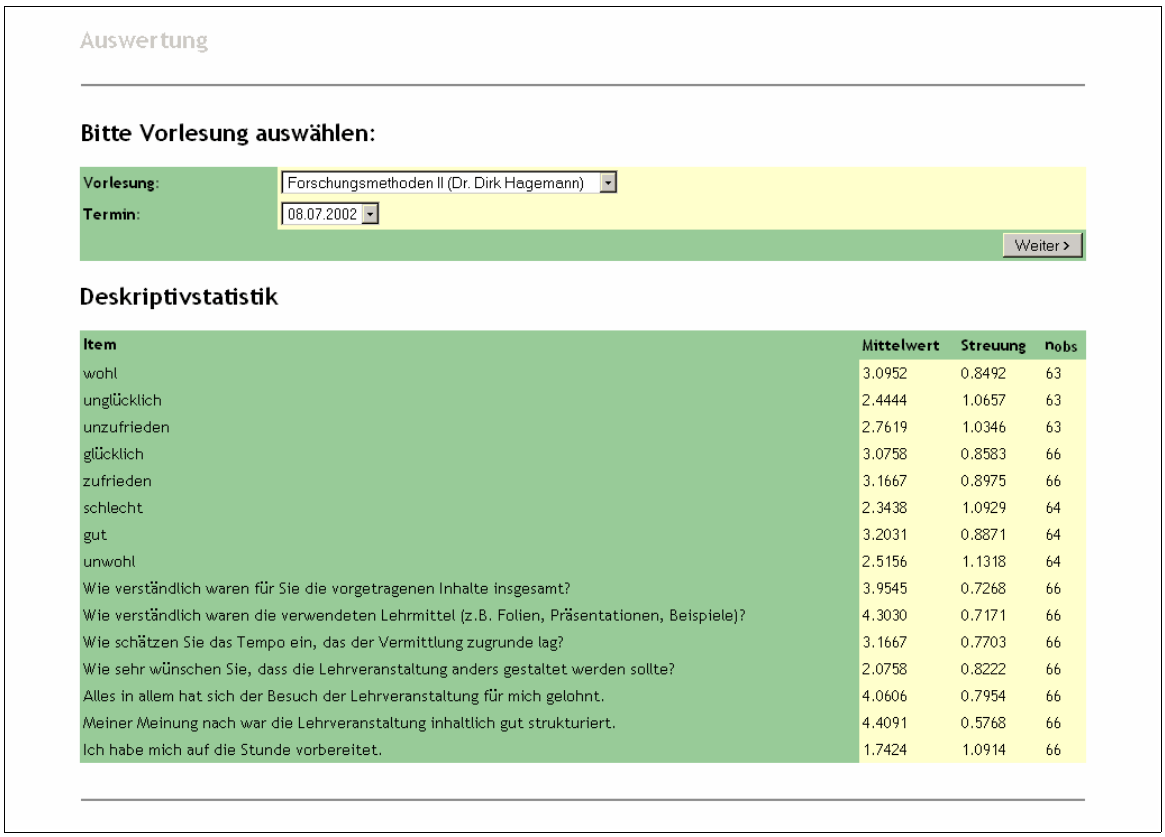


Abbildung F.14: Abschließende Antwortkontrolle mit Fehlermeldungen.

Abbildung F.15: Abschließende Antwortkontrolle ohne Fehlermeldungen.

Abbildung F.16: Abschlussbildschirm. Gleichzeitig werden die Antworten in der Datenbank gespeichert; eine nochmalige Teilnahme ist für diese Woche nicht mehr möglich.



Deskriptivstatistik

Item	Mittelwert	Streuung	nobs
wohl	3.0952	0.8492	63
unglücklich	2.4444	1.0657	63
unzufrieden	2.7619	1.0346	63
glücklich	3.0758	0.8583	66
zufrieden	3.1667	0.8975	66
schlecht	2.3438	1.0929	64
gut	3.2031	0.8871	64
unwohl	2.5156	1.1318	64
Wie verständlich waren für Sie die vorgetragenen Inhalte insgesamt?	3.9545	0.7268	66
Wie verständlich waren die verwendeten Lehrmittel (z.B. Folien, Präsentationen, Beispiele)?	4.3030	0.7171	66
Wie schätzen Sie das Tempo ein, das der Vermittlung zugrunde lag?	3.1667	0.7703	66
Wie sehr wünschen Sie, dass die Lehrveranstaltung anders gestaltet werden sollte?	2.0758	0.8222	66
Alles in allem hat sich der Besuch der Lehrveranstaltung für mich gelohnt.	4.0606	0.7954	66
Meiner Meinung nach war die Lehrveranstaltung inhaltlich gut strukturiert.	4.4091	0.5768	66
Ich habe mich auf die Stunde vorbereitet.	1.7424	1.0914	66

Abbildung F.17: Beispiel einer Auswertung mit Hilfe des Auswertungsskriptes. Für jede Vorlesung wurden (nach Item sortiert) der Mittelwert, die Streuung und die Anzahl der Beobachtungen angegeben.

Anhang G: Ergänzungsfragebogen

Ergänzungsfragebogen zur Veranstaltung Quantitative Methoden B

Matrikelnummer: _____

Denken Sie bitte an eine typische Vorlesungswoche im vergangenen Semester:

- An wievielen Tagen der Woche haben Sie am Computer gearbeitet (außer Internetnutzung)? _____ Tage
- Wieviel Zeit haben Sie **insgesamt** am Computer gearbeitet (außer Internetnutzung)? _____ Stunden
- An wievielen Tagen der Woche haben Sie das Internet genutzt? _____ Tage
- Wieviel Zeit haben Sie **insgesamt** im Internet verbracht? _____ Stunden

Spontan finde ich den Dozenten eher...

	1	2	3	4	5	
sympathisch	○	○	○	○	○	unsympathisch

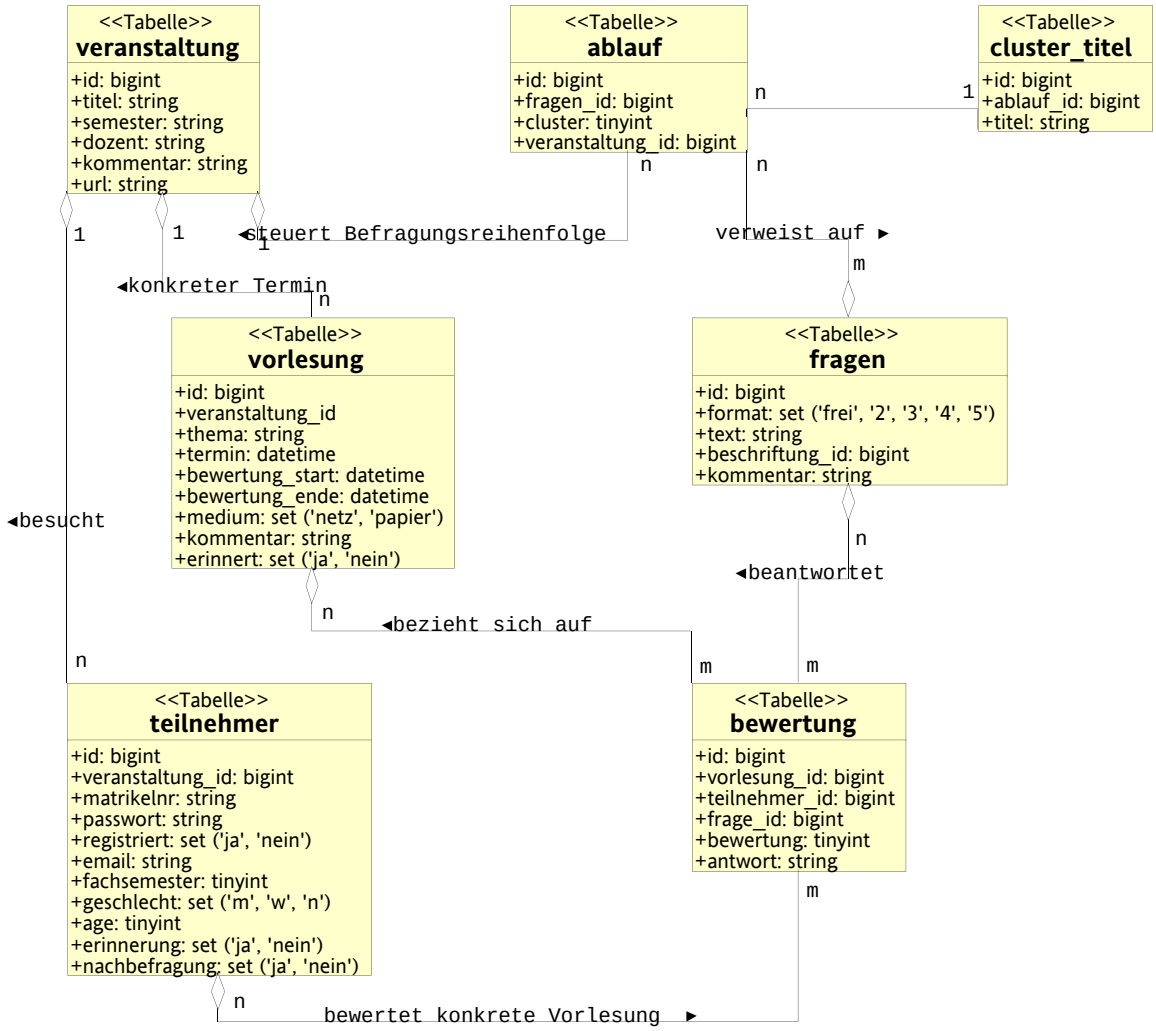
Denken Sie nun bitte an die Vorlesung über das gesamte Semester:

- **Was hat Ihnen konkret an der Vorlesung gefallen?**
  (etwa hinsichtlich der inhaltlichen Struktur, Geschwindigkeit, eingesetzten Medien, etc. ...)

- **Was hat Ihnen konkret an der Vorlesung nicht gefallen oder Sie gestört?**
  (etwa hinsichtlich der inhaltlichen Struktur, Geschwindigkeit, eingesetzten Medien, etc. ...)

Vielen Dank!

Anhang H: Datenbankstruktur



Die Darstellung verwendet die *Unified Modeling Language* (UML, Object Management Group, 2001). Die Elemente der Tabelle `Teilnehmer` sind direkt an ein Element der Tabelle `Vorlesung` gebunden. Für eine allgemeinere Form der Datenbank sollte man an dieser Stelle keinen direkten Schlüssel, sondern eine Relation verwenden.

Anhang I: E-Mail-Vorlagen

1 Erinnerungsmails

```
Von:      Joachim Schroer <schr1305@uni-trier.de>
Betreff: [Evaluation] Erinnerung an die Evaluation der Vorlesung
         "<Vorlesung>" vom <Datum>.

Hallo,

vielen Dank, dass Du an unserer Untersuchung zur Evaluation von
Lehrveranstaltungen teilnimmst! Die Vorlesung vom <Datum> kann noch bis
<Wochentag> unter folgender Adresse bewertet werden:

http://psychologie.uni-trier.de/evaluation/<Nachname des Dozenten>/

Schöne Grüße,

         Joachim Schroer


PS: Falls Du von diesem Projekt nichts gehört haben solltest, gab es
wahrscheinlich einen Fehler beim Abtippen des Fragebogens. In diesem Fall
würde ich mich über eine kurze Mail freuen.
```

2 Vergabe der Versuchspersonenstunden

```
Von:      Joachim Schroer <schr1305@uni-trier.de>
Betreff: Teilnahmebescheinigung

                        Teilnahmebescheinigung

Hiermit bestätige ich, dass <der Student/die Studentin> mit der Ma-
trikelnummer <Matrikelnummer> im Sommersemester 2002 an insgesamt
<Anzahl> Terminen an unserer Studie zur Lehrevaluation teilgenommen
hat.

<Er/Sie> erhält dafür <eine volle/eine halbe> Versuchspersonenstunde.

Bei Fragen wenden Sie sich bitte an mich oder Joachim Schroer
(schr1305@uni-trier.de).

Gez.
     Dr. Ewald Naumann
```

Anhang J: Korrelationsmatrizen

Tabelle J.1: Matrix der bivariaten Korrelationen in der Kalenderwoche 28 (Vorlesung *Quantitative Methoden B* insgesamt).

Quantitative Methoden B (insgesamt)

		allgemeine Computer-nutzung	allgemeine Internet-nutzung	Dozent sympathisch	Eigene Vorbereitung (KW 28)	allgemeine Befindlichkeit (KW 28)	Inhalt (KW 28)	Lehrmittel (KW 28)	Tempo (KW 28)	Änderungs-wunsch (KW 28)	insgesamt lohnend (KW 28)	Strukturiert-heit (KW 28)
allgemeine Computernutzung	Korrelation nach Pearson	1	,401**	-,104	,115	,016	-,028	,044	-,094	,013	,104	,066
	Signifikanz (2-seitig)	,	,000	,273	,281	,883	,792	,685	,379	,905	,330	,540
	N	113	113	113	89	89	89	89	89	89	89	89
allgemeine Internetnutzung	Korrelation nach Pearson	,401**	1	-,047	-,046	-,026	,098	,045	-,057	,041	-,129	-,016
	Signifikanz (2-seitig)	,000	,	,618	,665	,809	,360	,676	,596	,705	,229	,879
	N	113	113	113	89	89	89	89	89	89	89	89
Dozent sympathisch	Korrelation nach Pearson	-,104	-,047	1	,240*	,123	,268*	,240*	-,131	-,300**	,258*	,300**
	Signifikanz (2-seitig)	,273	,618	,	,023	,251	,011	,024	,223	,004	,015	,004
	N	113	113	113	89	89	89	89	89	89	89	89
Eigene Vorbereitung (KW 28)	Korrelation nach Pearson	,115	-,046	,240*	1	,148	,402**	,156	-,211*	-,198	,406**	,189
	Signifikanz (2-seitig)	,281	,665	,023	,	,151	,000	,129	,039	,053	,000	,065
	N	89	89	89	96	96	96	96	96	96	96	96
allgemeine Befindlichkeit (KW 28)	Korrelation nach Pearson	,016	-,026	,123	,148	1	,117	-,028	-,221*	-,130	,087	,103
	Signifikanz (2-seitig)	,883	,809	,251	,151	,	,257	,787	,031	,208	,397	,319
	N	89	89	89	96	96	96	96	96	96	96	96
Inhalt (KW 28)	Korrelation nach Pearson	-,028	,098	,268*	,402**	,117	1	,516**	-,294**	-,341**	,333**	,364**
	Signifikanz (2-seitig)	,792	,360	,011	,000	,257	,	,000	,004	,001	,001	,000
	N	89	89	89	96	96	96	96	96	96	96	96
Lehrmittel (KW 28)	Korrelation nach Pearson	,044	,045	,240*	,156	-,028	,516**	1	-,128	-,481**	,305**	,515**
	Signifikanz (2-seitig)	,685	,676	,024	,129	,787	,000	,	,212	,000	,003	,000
	N	89	89	89	96	96	96	96	96	96	96	96
Tempo (KW 28)	Korrelation nach Pearson	-,094	-,057	-,131	-,211*	-,221*	-,294**	-,128	1	,172	-,101	-,053
	Signifikanz (2-seitig)	,379	,596	,223	,039	,031	,004	,212	,	,095	,329	,610
	N	89	89	89	96	96	96	96	96	96	96	96
Änderungswunsch (KW 28)	Korrelation nach Pearson	,013	,041	-,300**	-,198	-,130	-,341**	-,481**	,172	1	-,500**	-,502**
	Signifikanz (2-seitig)	,905	,705	,004	,053	,208	,001	,000	,095	,	,000	,000
	N	89	89	89	96	96	96	96	96	96	96	96
insgesamt lohnend (KW 28)	Korrelation nach Pearson	,104	-,129	,258*	,406**	,087	,333**	,305**	-,101	-,500**	1	,450**
	Signifikanz (2-seitig)	,330	,229	,015	,000	,397	,001	,003	,329	,000	,	,000
	N	89	89	89	96	96	96	96	96	96	96	96
Strukturiertheit (KW 28)	Korrelation nach Pearson	,066	-,016	,300**	,189	,103	,364**	,515**	-,053	-,502**	,450**	1
	Signifikanz (2-seitig)	,540	,879	,004	,065	,319	,000	,000	,610	,000	,000	,
	N	89	89	89	96	96	96	96	96	96	96	96

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Tabelle J.2: Matrix der bivariaten Korrelationen in der Kalenderwoche 28 (Vorlesung *Forschungsmethoden II* insgesamt). Die Korrelationen der Kontrollvariablen mit den abhängigen Variablen sind hervorgehoben.

Forschungsmethoden II (insgesamt)

		allgemeine Computer-nutzung	allgemeine Internet-nutzung	Dozent sympathisch	Eigene Vorbereitung (KW 28)	allgemeine Befindlichkeit (KW 28)	Inhalt (KW 28)	Lehrmittel (KW 28)	Tempo (KW 28)	Änderungs-wunsch (KW 28)	insgesamt lohnend (KW 28)	Strukturiert-heit (KW 28)
allgemeine Computernutzung	Korrelation nach Pearson	1	,248*	,041	,282	,044	,193	,051	,131	,161	,109	-,128
	Signifikanz (2-seitig)	,	,038	,733	,054	,767	,193	,735	,380	,279	,464	,391
	N	70	70	70	47	47	47	47	47	47	47	47
allgemeine Internetnutzung	Korrelation nach Pearson	,248*	1	-,180	,135	-,134	-,227	-,165	,110	,246	-,110	-,266
	Signifikanz (2-seitig)	,038	,	,136	,364	,370	,125	,269	,463	,096	,463	,071
	N	70	70	70	47	47	47	47	47	47	47	47
Dozent sympathisch	Korrelation nach Pearson	,041	-,180	1	,050	,050	,062	,142	,150	-,306*	,121	,269
	Signifikanz (2-seitig)	,733	,136	,	,737	,737	,680	,342	,315	,037	,419	,068
	N	70	70	70	47	47	47	47	47	47	47	47
Eigene Vorbereitung (KW 28)	Korrelation nach Pearson	,282	,135	,050	1	,027	,010	-,118	-,010	-,004	-,328*	-,014
	Signifikanz (2-seitig)	,054	,364	,737	,	,847	,941	,403	,946	,976	,018	,922
	N	47	47	47	52	52	52	52	52	52	52	52
allgemeine Befindlichkeit (KW 28)	Korrelation nach Pearson	,044	-,134	,050	,027	1	,365*[1]	,200	-,126	-,058	-,014	,154
	Signifikanz (2-seitig)	,767	,370	,737	,847	,	,008	,155	,375	,683	,920	,275
	N	47	47	47	52	52	52	52	52	52	52	52
Inhalt (KW 28)	Korrelation nach Pearson	,193	-,227	,062	,010	,365*[1]	1	,617*[1]	-,498*[1]	-,473*[1]	,531*[1]	,241
	Signifikanz (2-seitig)	,193	,125	,680	,941	,008	,	,000	,000	,000	,000	,085
	N	47	47	47	52	52	52	52	52	52	52	52
Lehrmittel (KW 28)	Korrelation nach Pearson	,051	-,165	,142	-,118	,200	,617*[1]	1	-,205	-,551*[1]	,396*[1]	,437**[1]
	Signifikanz (2-seitig)	,735	,269	,342	,403	,155	,000	,	,144	,000	,004	,001
	N	47	47	47	52	52	52	52	52	52	52	52
Tempo (KW 28)	Korrelation nach Pearson	,131	,110	,150	-,010	-,126	-,498*[1]	-,205	1	,372*[1]	-,151	-,126
	Signifikanz (2-seitig)	,380	,463	,315	,946	,375	,000	,144	,	,007	,285	,374
	N	47	47	47	52	52	52	52	52	52	52	52
Änderungswunsch (KW 28)	Korrelation nach Pearson	,161	,246	-,306*	-,004	-,058	-,473*[1]	-,551*[1]	,372*[1]	1	-,414*[1]	-,395**[1]
	Signifikanz (2-seitig)	,279	,096	,037	,976	,683	,000	,000	,007	,	,002	,004
	N	47	47	47	52	52	52	52	52	52	52	52
insgesamt lohnend (KW 28)	Korrelation nach Pearson	,109	-,110	,121	-,328*	-,014	,531*[1]	,396*[1]	-,151	-,414*[1]	1	,197
	Signifikanz (2-seitig)	,464	,463	,419	,018	,920	,000	,004	,285	,002	,	,161
	N	47	47	47	52	52	52	52	52	52	52	52
Strukturiertheit (KW 28)	Korrelation nach Pearson	-,128	-,266	,269	-,014	,154	,241	,437**[1]	-,126	-,395*[1]	,197	1
	Signifikanz (2-seitig)	,391	,071	,068	,922	,275	,085	,001	,374	,004	,161	,
	N	47	47	47	52	52	52	52	52	52	52	52

*[1]: Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

**[1]: Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Tabelle J.3: Matrix der bivariaten Korrelationen in der Kalenderwoche 28 (Vorlesung *Quantitative Methoden B*, nur Papierfragebogen).

Quantitative Methoden B (Papierfragebogen)

		allgemeine Computer-nutzung	allgemeine Internet-nutzung	Dozent sympathisch	Eigene Vorbereitung (KW 28)	allgemeine Befindlichkeit (KW 28)	Inhalt (KW 28)	Lehrmittel (KW 28)	Tempo (KW 28)	Änderungs-wunsch (KW 28)	insgesamt lohnend (KW 28)	Strukturiert-heit (KW 28)
allgemeine Computernutzung	Korrelation nach Pearson	1	,452**[1]	,019	,287	,058	,032	,087	-,184	,124	,076	,096
	Signifikanz (2-seitig)	,	,001	,892	,050	,698	,829	,563	,217	,408	,612	,520
	N	51	51	51	47	47	47	47	47	47	47	47
allgemeine Internetnutzung	Korrelation nach Pearson	,452**[1]	1	-,239	-,044	-,153	-,065	-,027	,040	,052	-,061	-,099
	Signifikanz (2-seitig)	,001	,	,091	,768	,305	,666	,855	,788	,727	,685	,509
	N	51	51	51	47	47	47	47	47	47	47	47
Dozent sympathisch	Korrelation nach Pearson	,019	-,239	1	,175	,098	,374**[1]	,365*	-,261	-,362*	,301*	,275
	Signifikanz (2-seitig)	,892	,091	,	,239	,511	,010	,012	,077	,012	,040	,061
	N	51	51	51	47	47	47	47	47	47	47	47
Eigene Vorbereitung (KW 28)	Korrelation nach Pearson	,287	-,044	,175	1	,274	,321*	,173	-,241	-,249	,412**[1]	,290*
	Signifikanz (2-seitig)	,050	,768	,239	,	,060	,026	,240	,098	,088	,004	,046
	N	47	47	47	48	48	48	48	48	48	48	48
allgemeine Befindlichkeit (KW 28)	Korrelation nach Pearson	,058	-,153	,098	,274	1	,027	-,067	-,063	-,131	,195	,131
	Signifikanz (2-seitig)	,698	,305	,511	,060	,	,854	,653	,671	,377	,185	,376
	N	47	47	47	48	48	48	48	48	48	48	48
Inhalt (KW 28)	Korrelation nach Pearson	,032	-,065	,374**[1]	,321*	,027	1	,531**[1]	-,275	-,459**[1]	,468**[1]	,422**[1]
	Signifikanz (2-seitig)	,829	,666	,010	,026	,854	,	,000	,059	,001	,001	,003
	N	47	47	47	48	48	48	48	48	48	48	48
Lehrmittel (KW 28)	Korrelation nach Pearson	,087	-,027	,365*	,173	-,067	,531**[1]	1	-,318*	-,515**[1]	,336*	,501**[1]
	Signifikanz (2-seitig)	,563	,855	,012	,240	,653	,000	,	,028	,000	,020	,000
	N	47	47	47	48	48	48	48	48	48	48	48
Tempo (KW 28)	Korrelation nach Pearson	-,184	,040	-,261	-,241	-,063	-,275	-,318*	1	,034	-,088	-,080
	Signifikanz (2-seitig)	,217	,788	,077	,098	,671	,059	,028	,	,819	,553	,589
	N	47	47	47	48	48	48	48	48	48	48	48
Änderungswunsch (KW 28)	Korrelation nach Pearson	,124	,052	-,362*	-,249	-,131	-,459**[1]	-,515**[1]	,034	1	-,487**[1]	-,624**[1]
	Signifikanz (2-seitig)	,408	,727	,012	,088	,377	,001	,000	,819	,	,000	,000
	N	47	47	47	48	48	48	48	48	48	48	48
insgesamt lohnend (KW 28)	Korrelation nach Pearson	,076	-,061	,301*	,412**[1]	,195	,468**[1]	,336*	-,088	-,487**[1]	1	,581**[1]
	Signifikanz (2-seitig)	,612	,685	,040	,004	,185	,001	,020	,553	,000	,	,000
	N	47	47	47	48	48	48	48	48	48	48	48
Strukturiertheit (KW 28)	Korrelation nach Pearson	,096	-,099	,275	,290*	,131	,422**[1]	,501**[1]	-,080	-,624**[1]	,581**[1]	1
	Signifikanz (2-seitig)	,520	,509	,061	,046	,376	,003	,000	,589	,000	,000	,
	N	47	47	47	48	48	48	48	48	48	48	48

**[1]. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

*[1]. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Tabelle J.4: Matrix der bivariaten Korrelationen in der Kalenderwoche 28 (Vorlesung *Quantitative Methoden B*, nur Internetfragebogen).

Quantitative Methoden B (Internetfragebogen)

		allgemeine Computer-nutzung	allgemeine Internet-nutzung	Dozent sympathisch	Eigene Vorbereitung (KW 28)	allgemeine Befindlichkeit (KW 28)	Inhalt (KW 28)	Lehrmittel (KW 28)	Tempo (KW 28)	Änderungs-wunsch (KW 28)	insgesamt lohnend (KW 28)	Strukturiert-heit (KW 28)
allgemeine Computernutzung	Korrelation nach Pearson	1	,359**[1]	-,286*	-,058	,002	-,086	,018	-,039	-,091	,112	-,016
	Signifikanz (2-seitig)	,	,009	,040	,718	,992	,589	,910	,807	,568	,482	,918
	N	52	52	52	42	42	42	42	42	42	42	42
allgemeine Internetnutzung	Korrelation nach Pearson	,359**[1]	1	,102	-,029	,097	,297	,128	-,161	,002	-,214	,164
	Signifikanz (2-seitig)	,009	,	,474	,856	,542	,056	,420	,307	,992	,174	,300
	N	52	52	52	42	42	42	42	42	42	42	42
Dozent sympathisch	Korrelation nach Pearson	-,286*	,102	1	,295	,193	,171	,133	-,028	-,189	,157	,264
	Signifikanz (2-seitig)	,040	,474	,	,058	,220	,280	,400	,859	,231	,320	,091
	N	52	52	52	42	42	42	42	42	42	42	42
Eigene Vorbereitung (KW 28)	Korrelation nach Pearson	-,058	-,029	,295	1	,076	,453**[1]	,145	-,192	-,154	,401**[1]	,100
	Signifikanz (2-seitig)	,718	,856	,058	,	,610	,001	,325	,192	,296	,005	,500
	N	42	42	42	48	48	48	48	48	48	48	48
allgemeine Befindlichkeit (KW 28)	Korrelation nach Pearson	,002	,097	,193	,076	1	,173	-,008	-,316*	-,138	,020	,125
	Signifikanz (2-seitig)	,992	,542	,220	,610	,	,238	,957	,028	,350	,892	,395
	N	42	42	42	48	48	48	48	48	48	48	48
Inhalt (KW 28)	Korrelation nach Pearson	-,086	,297	,171	,453**[1]	,173	1	,512**[1]	-,312*	-,254	,233	,324*
	Signifikanz (2-seitig)	,589	,056	,280	,001	,238	,	,000	,031	,082	,111	,025
	N	42	42	42	48	48	48	48	48	48	48	48
Lehrmittel (KW 28)	Korrelation nach Pearson	,018	,128	,133	,145	-,008	,512**[1]	1	,019	-,455**[1]	,282	,576**[1]
	Signifikanz (2-seitig)	,910	,420	,400	,325	,957	,000	,	,896	,001	,052	,000
	N	42	42	42	48	48	48	48	48	48	48	48
Tempo (KW 28)	Korrelation nach Pearson	-,039	-,161	-,028	-,192	-,316*	-,312*	,019	1	,301*	-,117	-,052
	Signifikanz (2-seitig)	,807	,307	,859	,192	,028	,031	,896	,	,037	,428	,727
	N	42	42	42	48	48	48	48	48	48	48	48
Änderungswunsch (KW 28)	Korrelation nach Pearson	-,091	,002	-,189	-,154	-,138	-,254	-,455**[1]	,301*	1	-,514**[1]	-,385**[1]
	Signifikanz (2-seitig)	,568	,992	,231	,296	,350	,082	,001	,037	,	,000	,007
	N	42	42	42	48	48	48	48	48	48	48	48
insgesamt lohnend (KW 28)	Korrelation nach Pearson	,112	-,214	,157	,401**[1]	,020	,233	,282	-,117	-,514**[1]	1	,316*
	Signifikanz (2-seitig)	,482	,174	,320	,005	,892	,111	,052	,428	,000	,	,029
	N	42	42	42	48	48	48	48	48	48	48	48
Strukturiertheit (KW 28)	Korrelation nach Pearson	-,016	,164	,264	,100	,125	,324*	,576**[1]	-,052	-,385**[1]	,316*	1
	Signifikanz (2-seitig)	,918	,300	,091	,500	,395	,025	,000	,727	,007	,029	,
	N	42	42	42	48	48	48	48	48	48	48	48

**[1]. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

*[1]. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Tabelle J.5: Matrix der bivariaten Korrelationen in der Kalenderwoche 28 (Vorlesung *Forschungsmethoden II*, nur Papierfragebogen).

Forschungsmethoden II (Papierfragebogen)

		allgemeine Computer-nutzung	allgemeine Internet-nutzung	Dozent sympathisch	Eigene Vorbereitung (KW 28)	allgemeine Befindlichkeit (KW 28)	Inhalt (KW 28)	Lehrmittel (KW 28)	Tempo (KW 28)	Änderungs-wunsch (KW 28)	insgesamt lohnend (KW 28)	Strukturiert-heit (KW 28)
allgemeine Computernutzung	Korrelation nach Pearson	1	,461*	-,359	,678**[1]	-,258	-,030	-,244	-,194	,427*	-,019	-,333
	Signifikanz (2-seitig)	,	,020	,078	,001	,247	,896	,275	,387	,048	,934	,130
	N	25	25	25	22	22	22	22	22	22	22	22
allgemeine Internetnutzung	Korrelation nach Pearson	,461*	1	-,165	,461*	-,063	-,380	-,503*	,238	,303	,010	-,216
	Signifikanz (2-seitig)	,020	,	,431	,031	,781	,081	,017	,286	,170	,966	,334
	N	25	25	25	22	22	22	22	22	22	22	22
Dozent sympathisch	Korrelation nach Pearson	-,359	-,165	1	-,118	,140	-,071	,206	,152	-,351	,048	,237
	Signifikanz (2-seitig)	,078	,431	,	,602	,535	,754	,357	,499	,109	,834	,288
	N	25	25	25	22	22	22	22	22	22	22	22
Eigene Vorbereitung (KW 28)	Korrelation nach Pearson	,678**[1]	,461*	-,118	1	,214	,157	-,288	-,505**[1]	-,020	-,396*	,011
	Signifikanz (2-seitig)	,001	,031	,602	,	,305	,453	,163	,010	,923	,050	,958
	N	22	22	22	25	25	25	25	25	25	25	25
allgemeine Befindlichkeit (KW 28)	Korrelation nach Pearson	-,258	-,063	,140	,214	1	,303	,226	-,310	-,250	-,204	,182
	Signifikanz (2-seitig)	,247	,781	,535	,305	,	,141	,278	,131	,229	,327	,384
	N	22	22	22	25	25	25	25	25	25	25	25
Inhalt (KW 28)	Korrelation nach Pearson	-,030	-,380	-,071	,157	,303	1	,596**[1]	-,624**[1]	-,561**[1]	,465*	,272
	Signifikanz (2-seitig)	,896	,081	,754	,453	,141	,	,002	,001	,004	,019	,188
	N	22	22	22	25	25	25	25	25	25	25	25
Lehrmittel (KW 28)	Korrelation nach Pearson	-,244	-,503*	,206	-,288	,226	,596**[1]	1	-,216	-,550**[1]	,424*	,517**[1]
	Signifikanz (2-seitig)	,275	,017	,357	,163	,278	,002	,	,300	,004	,035	,008
	N	22	22	22	25	25	25	25	25	25	25	25
Tempo (KW 28)	Korrelation nach Pearson	-,194	,238	,152	-,505**[1]	-,310	-,624**[1]	-,216	1	,519**[1]	-,005	-,243
	Signifikanz (2-seitig)	,387	,286	,499	,010	,131	,001	,300	,	,008	,981	,243
	N	22	22	22	25	25	25	25	25	25	25	25
Änderungswunsch (KW 28)	Korrelation nach Pearson	,427*	,303	-,351	-,020	-,250	-,561**[1]	-,550**[1]	,519**[1]	1	-,413*	-,487*
	Signifikanz (2-seitig)	,048	,170	,109	,923	,229	,004	,004	,008	,	,040	,013
	N	22	22	22	25	25	25	25	25	25	25	25
insgesamt lohnend (KW 28)	Korrelation nach Pearson	-,019	,010	,048	-,396*	-,204	,465*	,424*	-,005	-,413*	1	,090
	Signifikanz (2-seitig)	,934	,966	,834	,050	,327	,019	,035	,981	,040	,	,667
	N	22	22	22	25	25	25	25	25	25	25	25
Strukturiertheit (KW 28)	Korrelation nach Pearson	-,333	-,216	,237	,011	,182	,272	,517**[1]	-,243	-,487*	,090	1
	Signifikanz (2-seitig)	,130	,334	,288	,958	,384	,188	,008	,243	,013	,667	,
	N	22	22	22	25	25	25	25	25	25	25	25

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Tabelle J.6: Matrix der bivariaten Korrelationen in der Kalenderwoche 28 (Vorlesung *Forschungsmethoden II*, nur Internetfragebogen).

Forschungsmethoden II (Internetfragebogen)

		allgemeine Computer-nutzung	allgemeine Internet-nutzung	Dozent sympathisch	Eigene Vorbereitung (KW 28)	allgemeine Befindlichkeit (KW 28)	Inhalt (KW 28)	Lehrmittel (KW 28)	Tempo (KW 28)	Änderungs-wunsch (KW 28)	insgesamt lohnend (KW 28)	Strukturiert-heit (KW 28)
allgemeine Computernutzung	Korrelation nach Pearson	1	,212	,113	,101	,258	,394	,358	,367	-,147	,251	,108
	Signifikanz (2-seitig)	,	,271	,560	,632	,213	,051	,079	,071	,485	,227	,606
	N	29	29	29	25	25	25	25	25	25	25	25
allgemeine Internetnutzung	Korrelation nach Pearson	,212	1	-,209	-,022	-,164	-,099	,165	,041	,190	-,235	-,328
	Signifikanz (2-seitig)	,271	,	,276	,915	,433	,639	,431	,846	,364	,258	,109
	N	29	29	29	25	25	25	25	25	25	25	25
Dozent sympathisch	Korrelation nach Pearson	,113	-,209	1	,157	,013	,257	,092	,198	-,244	,241	,333
	Signifikanz (2-seitig)	,560	,276	,	,453	,950	,215	,662	,343	,240	,245	,103
	N	29	29	29	25	25	25	25	25	25	25	25
Eigene Vorbereitung (KW 28)	Korrelation nach Pearson	,101	-,022	,157	1	-,096	-,125	,069	,461*	,009	-,259	-,045
	Signifikanz (2-seitig)	,632	,915	,453	,	,633	,535	,732	,016	,964	,192	,823
	N	25	25	25	27	27	27	27	27	27	27	27
allgemeine Befindlichkeit (KW 28)	Korrelation nach Pearson	,258	-,164	,013	-,096	1	,399*	,139	-,033	,157	,232	,149
	Signifikanz (2-seitig)	,213	,433	,950	,633	,	,039	,490	,870	,435	,244	,457
	N	25	25	25	27	27	27	27	27	27	27	27
Inhalt (KW 28)	Korrelation nach Pearson	,394	-,099	,257	-,125	,399*	1	,633**[1]	-,400*	-,367	,655**[1]	,210
	Signifikanz (2-seitig)	,051	,639	,215	,535	,039	,	,000	,039	,060	,000	,294
	N	25	25	25	27	27	27	27	27	27	27	27
Lehrmittel (KW 28)	Korrelation nach Pearson	,358	,165	,092	,069	,139	,633**[1]	1	-,216	-,548**[1]	,396*	,341
	Signifikanz (2-seitig)	,079	,431	,662	,732	,490	,000	,	,280	,003	,041	,082
	N	25	25	25	27	27	27	27	27	27	27	27
Tempo (KW 28)	Korrelation nach Pearson	,367	,041	,198	,461*	-,033	-,400*	-,216	1	,225	-,324	,009
	Signifikanz (2-seitig)	,071	,846	,343	,016	,870	,039	,280	,	,260	,099	,966
	N	25	25	25	27	27	27	27	27	27	27	27
Änderungswunsch (KW 28)	Korrelation nach Pearson	-,147	,190	-,244	,009	,157	-,367	-,548**[1]	,225	1	-,438*	-,260
	Signifikanz (2-seitig)	,485	,364	,240	,964	,435	,060	,003	,260	,	,022	,190
	N	25	25	25	27	27	27	27	27	27	27	27
insgesamt lohnend (KW 28)	Korrelation nach Pearson	,251	-,235	,241	-,259	,232	,655**[1]	,396*	-,324	-,438*	1	,371
	Signifikanz (2-seitig)	,227	,258	,245	,192	,244	,000	,041	,099	,022	,	,057
	N	25	25	25	27	27	27	27	27	27	27	27
Strukturiertheit (KW 28)	Korrelation nach Pearson	,108	-,328	,333	-,045	,149	,210	,341	,009	-,260	,371	1
	Signifikanz (2-seitig)	,606	,109	,103	,823	,457	,294	,082	,966	,190	,057	,
	N	25	25	25	27	27	27	27	27	27	27	27

*[1]. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

**[1]. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Kalender-woche	Quantitative Methoden B						Forschungsmethoden II					
	Inhalt	Lehrmittel	Tempo	Änderungs-wunsch	insgesamt lohnend	Struktur	Inhalt	Lehrmittel	Tempo	Änderungs-wunsch	insgesamt lohnend	Struktur
18	0,22	0,17	0,04	-0,01	0,12	0,05	-0,29	-0,24	0,25	0,25	-0,01	0,08
19	0,17	0,14	-0,09	0,14	0,07	0,09	0,02	-0,02	0,14	0,11	0,14	-0,11
20	0,46	0,23	0,04	-0,13	0,28	0,28	0,16	0,26	0,08	-0,05	0,17	0,18
22	0,28	0,12	-0,11	-0,02	0,21	0,03	0,22	0,19	0	-0,04	0,1	0,13
24	0,19	0,22	-0,1	-0,17	0,23	0,1	0,37	0,29	-0,21	0,19	0,05	0,23
25	0,28	0,05	0	0,1	0,26	0,18	0,08	0,2	0,12	0,09	0,06	0,2
26	0,05	0,08	0,07	0	0,12	-0,01	0,31	0,16	0,03	-0,05	0,2	0,29
27	0,2	0,02	-0,05	0,02	0,31	0,13	0,19	0,27	-0,23	-0,13	0,09	0,13
28	0,4	0,16	-0,21	-0,2	0,41	0,19	0,01	-0,12	-0,01	0	-0,33	-0,01
Mittel 18-27	0,23	0,13	-0,03	-0,01	0,2	0,11	0,13	0,14	0,02	0,05	0,1	0,14

Tabelle J.7: Matrix der bivariaten Korrelationen zwischen der Kontrollvariable *eigene Vorbereitung* und den studentischen Einschätzungen über den Semesterverlauf, aufgeteilt nach der besuchten Veranstaltung. Zur Berechnung der Mittelwerte wurden die einzelnen Korrelationskoeffizienten Fisher-Z-transformiert, gemittelt und anschließend wieder zurücktransformiert.

Kalender-woche	Quantitative Methoden B						Forschungsmethoden II					
	Inhalt	Lehrmittel	Tempo	Änderungs-wunsch	insgesamt lohnend	Struktur	Inhalt	Lehrmittel	Tempo	Änderungs-wunsch	insgesamt lohnend	Struktur
18	0,21	0,04	-0,07	0,17	-0,13	-0,08	0,41	0,21	-0,09	-0,38	0,4	0,22
19	0,34	0,38	-0,28	0	0,29	0,15	0,22	0,27	-0,02	-0,24	0,25	0,36
20	0,32	0,26	-0,12	-0,18	0,13	0,06	0,4	0,36	-0,25	-0,32	0,41	0,31
22	0,38	0,27	-0,21	-0,16	0,31	0,12	0,26	0,3	0,07	-0,09	0,41	0,18
24	0,4	0,29	-0,27	-0,27	0,41	0,2	0,03	-0,15	0	0,1	0,13	0,01
25	0,24	0,36	0,02	-0,24	0,44	0,27	0,29	0,32	-0,28	-0,28	0,23	0,3
26	0,32	0,46	-0,23	-0,28	0,44	0,28	0,16	0,13	0,28	-0,18	0,38	0,28
27	0,1	0,22	-0,14	-0,13	0,17	0,09	0,32	0,23	-0,16	-0,29	0,27	0,14
28	0,12	-0,03	-0,22	-0,13	0,09	0,1	0,37	0,2	-0,13	-0,06	-0,01	0,15
Mittel 18-27	0,29	0,29	-0,16	-0,14	0,26	0,14	0,26	0,21	-0,06	-0,21	0,31	0,23

Tabelle J.8: Bivariate Korrelationen zwischen der Kontrollvariable *Befindlichkeit* und den studentischen Einschätzungen über den Semesterverlauf, aufgeteilt nach der besuchten Veranstaltung. Auch hier wurden die einzelnen Korrelationskoeffizienten aus den verschiedenen Veranstaltungen Fisher-Z-transformiert, gemittelt und anschließend wieder zurücktransformiert.

Anhang K: Prüfung der statistischen Annahmen

1 These 1: Äquivalenz der Internet- und Papierversion

Tabelle K.1.1: Deskriptive Statistik der mittleren studentischen Bewertungen im Auswertungsdesign 1.

Gemittelte Variablen, Quantitative Methoden B

	N	Mittelwert	Std.-Abw.	Schiefe		Kurtosis	
	Statistik	Statistik	Statistik	Statistik	Standardfehler	Statistik	Standardfehler
Mittelwert Inhalt | Netz	120	3,3549	,65621	-,200	,221	-,347	,438
Mittelwert Inhalt | Papier	121	3,4635	,61988	,019	,220	,145	,437
Mittelwert Lehrmittel | Netz	120	3,7160	,67072	-,353	,221	,490	,438
Mittelwert Lehrmittel | Papier	121	3,7376	,65508	-,243	,220	,118	,437
Mittelwert Tempo | Netz	120	3,3722	,61265	-,064	,221	1,756	,438
Mittelwert Tempo | Papier	121	3,3636	,49849	-,290	,220	,604	,437
Mittelwert Änderungswunsch | Netz	120	2,5910	,77288	,507	,221	,244	,438
Mittelwert Änderungswunsch | Papier	121	2,5055	,68505	,645	,220	1,526	,437
insgesamt lohnend | Netz	120	3,5722	,75623	-,566	,221	-,024	,438
insgesamt lohnend | Papier	121	3,6412	,67098	-,304	,220	-,172	,437
Mittelwert Strukturiertheit | Netz	120	3,9958	,63931	-,353	,221	-,211	,438
Mittelwert Strukturiertheit | Papier	121	3,9752	,57912	-,713	,220	,767	,437
Gültige Werte (Listenweise)	117						

Tabelle K.1.2: Deskriptive Statistik der mittleren studentischen Bewertungen im Auswertungsdesign 3.

Gemittelte Variablen, Forschungsmethoden II

	N	Mittelwert	Std.-Abw.	Schiefe		Kurtosis	
	Statistik	Statistik	Statistik	Statistik	Standardfehler	Statistik	Standardfehler
Mittelwert Inhalt | Netz	87	3,7289	,71032	-,641	,258	,535	,511
Mittelwert Inhalt | Papier	88	3,8873	,64994	-,580	,257	,516	,508
Mittelwert Lehrmittel | Netz	87	4,0450	,67924	-,600	,258	,524	,511
Mittelwert Lehrmittel | Papier	88	4,2036	,50395	-,509	,257	,444	,508
Mittelwert Tempo | Netz	87	3,6312	,72265	,199	,258	-,304	,511
Mittelwert Tempo | Papier	88	3,5284	,59316	,211	,257	-,065	,508
Mittelwert Änderungswunsch | Netz	87	2,2203	,79036	,203	,258	-,328	,511
Mittelwert Änderungswunsch | Papier	88	2,1913	,67773	,231	,257	-,393	,508
insgesamt lohnend | Netz	87	4,0814	,72637	-,585	,258	,200	,511
insgesamt lohnend | Papier	88	4,1013	,66708	-,732	,257	,948	,508
Mittelwert Strukturiertheit | Netz	87	4,4310	,63072	-1,542	,258	3,810	,511
Mittelwert Strukturiertheit | Papier	88	4,4858	,46736	-,615	,257	-,102	,508
Gültige Werte (Listenweise)	85						

Tabelle K.1.3: Deskriptive Statistik der mittleren studentischen Bewertungen im Auswertungsdesign 2.

Bewertungen in Kalenderwoche 28, Quantitative Methoden B

	N	Mittelwert	Std.-Abw.	Schiefe		Kurtosis	
	Statistik	Statistik	Statistik	Statistik	Standardfehler	Statistik	Standardfehler
Inhalt (KW 28)	96	3,82	,632	-,097	,246	-,002	,488
Lehrmittel (KW 28)	96	4,08	,627	-,584	,246	1,615	,488
Tempo (KW 28)	96	3,21	,479	,535	,246	,206	,488
Änderungswunsch (KW 28)	96	2,25	,768	,107	,246	-,383	,488
insgesamt lohnend (KW 28)	96	3,98	,781	-,640	,246	,382	,488
Strukturiertheit (KW 28)	96	4,28	,676	-,618	,246	,221	,488
Eigene Vorbereitung (KW 28)	96	3,41	1,157	-,517	,246	-,447	,488
Gültige Werte (Listenweise)	96						

Tabelle K.1.4: Deskriptive Statistik der mittleren studentischen Bewertungen im Auswertungsdesign 4.

Bewertungen in Kalenderwoche 28, Forschungsmethoden II

	N	Mittelwert	Std.-Abw.	Schiefe		Kurtosis	
	Statistik	Statistik	Statistik	Statistik	Standardfehler	Statistik	Standardfehler
Inhalt (KW 28)	52	3,81	,768	,078	,330	-,738	,650
Lehrmittel (KW 28)	52	4,23	,757	-,699	,330	,054	,650
Tempo (KW 28)	52	3,81	,971	-,532	,330	,014	,650
Änderungswunsch (KW 28)	52	2,35	,947	,391	,330	-,015	,650
insgesamt lohnend (KW 28)	52	4,04	,885	-,783	,330	,117	,650
Strukturiertheit (KW 28)	52	4,44	,574	-,412	,330	-,747	,650
Eigene Vorbereitung (KW 28)	52	1,85	1,127	1,341	,330	1,009	,650
Gültige Werte (Listenweise)	52						

Die MANOVA macht drei Annahmen (Stevens, 1996, S. 238):

1. *Unabhängige Beobachtungen*. Da die Bewertung der Vorlesung keine Interaktion unter den Studierenden erfordert, kann diese Voraussetzung als erfüllt gelten.

2. *Multivariate Normalverteilung* der abhängigen Variablen in allen Gruppen. Eine notwendige, aber nicht hinreichende Bedingung für die multivariate Normalverteilung ist zunächst die univariate Normalverteilung. Darüber hinaus müssen alle Linearkombinationen der Variablen und alle Untermengen aus der Menge der Variablen multivariat normalverteilt sein (Stevens, 1996, S. 243). Da kein statistischer Test der multivariaten Normalverteilung existiert, empfiehlt Stevens (ebd., S. 244), zunächst die univariate Normalverteilung der Variablen zu prüfen und dann grafisch die bivariate Normalverteilung abzuschätzen.

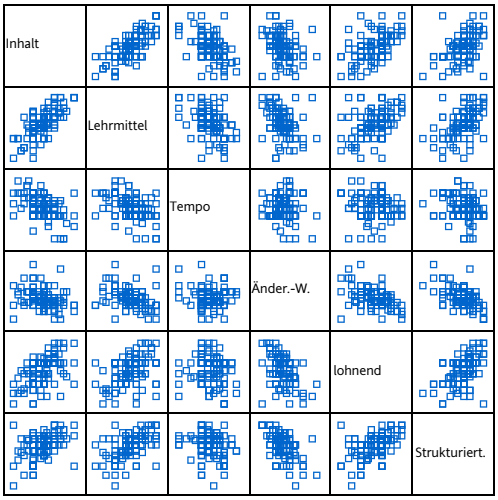
   Zur Überprüfung der *univariaten Normalverteilung* bei kleinen Stichproben ($n \leq 20$) raten Shapiro, Wilk und Chen (1968) sowohl Schiefe und Kurtosis als auch die Ergebnisse des Shapiro-Wilk-Tests zu berücksichtigen. West, Finch und Curran (1995) empfehlen, die Annahme der univariaten Normalverteilung bei Stichproben unter 200 Versuchspersonen nur durch den

Vergleich mit Grenzwerten von Schiefe und Kurtosis zu überprüfen. Die Schiefe sollte nicht größer als 2 sein, die Kurtosis den Wert 7 nicht überschreiten. Diese kritischen Werte wurden von den abhängigen Variablen in den Auswertungsdesigns 1 bis 4 nicht überschritten (Tabelle K.1.1 bis Tabelle K.1.4).

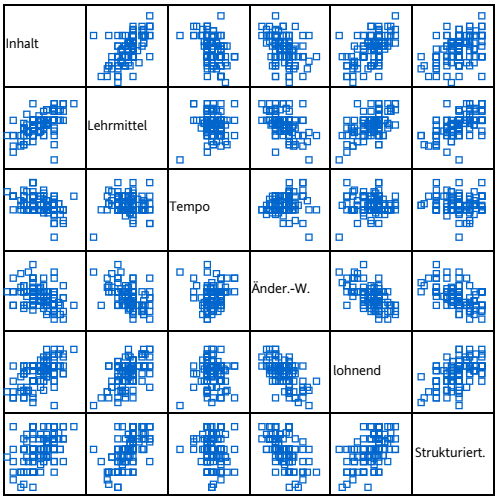
Zum Test der *bivariaten Normalverteilungen* schlägt Stevens (1996, S. 243) einen grafischen Test vor: Die Streudiagramme der Variablen sollten ungefähr Ellipsen bilden. Das trifft für die Mehrzahl der Streudiagramme in Abbildung K.1 und Abbildung K.2 zu. Damit kann die Annahme der multivariaten Normalverteilung als ausreichend belegt angesehen werden.

3. Die *Homogenität der Varianz-Kovarianzmatrizen* der abhängigen Variablen wurde in den Auswertungsdesigns 2 und 4 jeweils durch den Box-M-Test geprüft. Wenn dieser Test signifikant wird, muss die Annahme der homogenen Varianz-Kovarianzmatrizen verworfen werden. Das Ergebnis war jedoch in beiden Fällen nicht signifikant (Auswertungsdesign 2: $F(28, 30789.69) = 1.03$, $p = .42 > .05$; Auswertungsdesign 4: $F(28, 8597.75) = 1.14$, $p = .29 > .05$). Damit kann man die Annahme der Homogenität der Varianz-Kovarianzmatrizen beibehalten. Für die messwiederholten Hotellings T^2-Tests in den Auswertungsdesigns 1 und 3 ist diese Annahme ohnehin nicht erforderlich (Stevens, 1996, S. 459).

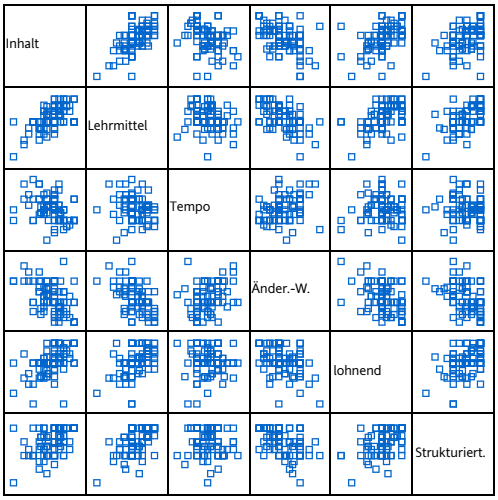
Papierfragebogen, Quantitative Methoden B



Internetfragebogen, Quantitative Methoden B



Papierfragebogen, Forschungsmethoden II



Internetfragebogen, Forschungsmethoden II

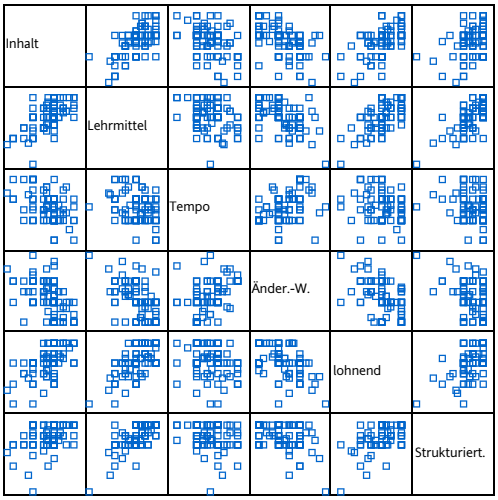
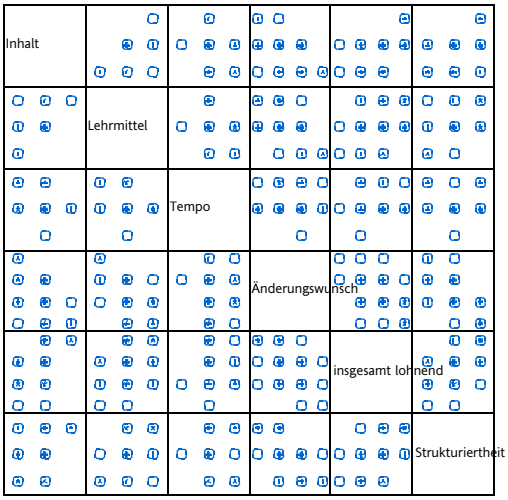
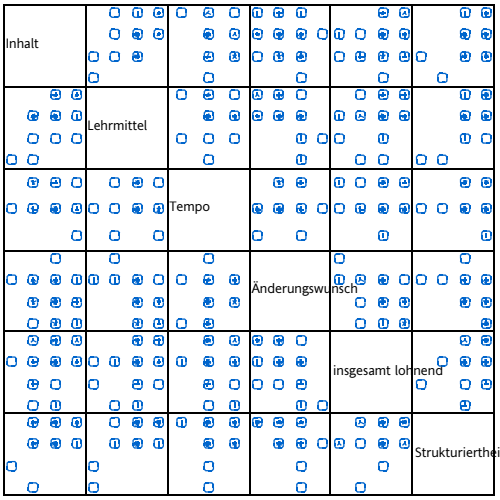


Abbildung K.1: Streudiagramme der abhängigen Variablen in den Auswertungsdesigns 1 und 3, aufgeteilt nach Papier- und Internetbefragungen.

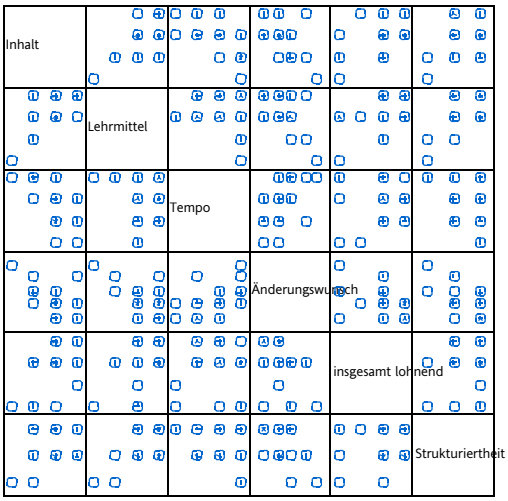
Papierfragebogen, Quantitative Methoden B



Internetfragebogen, Quantitative Methoden B



Papierfragebogen, Forschungsmethoden II



Internetfragebogen, Forschungsmethoden II

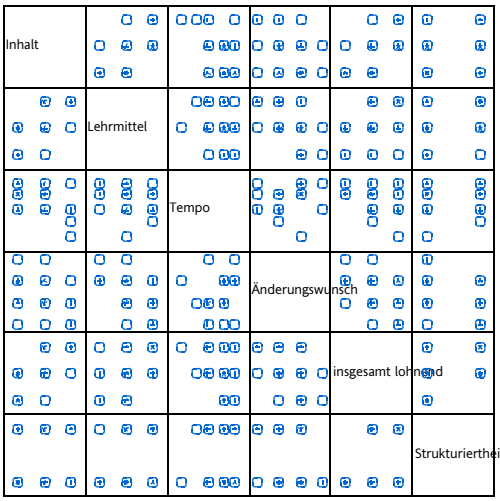


Abbildung K.2: Streudiagramme der abhängigen Variablen in den Auswertungsdesigns 2 und 4 (d.h. Kalenderwoche 28), aufgeteilt nach Papier- und Internetbefragungen. Da es sich bei den Werten in diesem Diagramm nicht um Mittelwerte handelt, liegen viele Antworten übereinander, so dass die Bewertung erschwert wird. Um die Verteilung trotzdem zu beurteilen, sollte man beachten, wie stark die Punkte ausgefüllt sind.

2 These 3: Beteiligungsquote

Vor der statistischen Auswertung wurden zunächst die Voraussetzungen zur Berechnung eines *t*-Test unabhängige Stichproben geprüft (Bortz, 1993, S. 132):

- Die beiden Stichproben sind voneinander *unabhängig*.
- Die *Normalverteilung der Variablen in beiden Gruppen* wurde per Kolmogorov-Smirnov- und Shapiro-Wilk-Test geprüft. Zusätzlich wurden deskriptive Maße herangezogen, da nach West, Finch und Curran (1995) die Normalverteilungsannahme verworfen werden sollte, sobald die Schiefe einer Verteilung > 2 und die Kurtosis > 7 ist. Zudem ist bei kleinen Stichproben (z. B. $n = 20$) die Kombination aus der Betrachtung von Schiefe und Kurtosis und der Berechnung des Shapiro-Wilk-Testes nach Shapiro, Wilk und Chen (1968) am sensibelsten für Abweichungen von der Normalverteilung. Die beiden statistischen Tests wurden nicht signifikant, und auch Schiefe und Kurtosis überschreiten nicht die kritischen Werte (Tabelle K.2.1). Aufgrund dieser Ergebnisse kann man die Variablen als normalverteilt ansehen.
- Die *Homogenität der Varianzen* der beiden Stichproben wurde mit dem Levene-Test der Varianzgleichheit geprüft. Dieser Test war nicht signifikant ($F(1,14) = 0.70$; $p = .70 > .05$). Die Annahme der Varianzhomogenität kann deshalb nicht verworfen werden.

Tabelle K.2.1: Prüfung der Normalverteilung der Beteiligungsquote. Keine der Prüfgrößen spricht gegen die Annahme der Normalverteilung.

	Kolmogorov-Smirnov			Shapiro-Wilk			Schiefe		Kurtosis	
Veranstaltung	**Statistik**	***df***	***p***	**Statistik**	***df***	***p***	**Statistik**	***SE***	**Statistik**	***SE***
Quantitative Methoden B	0.216	8	.20	0.892	8	.25	-0.66	0.75	-1.02	1.48
Forschungsmethoden II	0.242	8	.19	0.915	8	.39	0.02	0.75	-1.02	1.48

3 These 4: Variabilität der Befindlichkeit

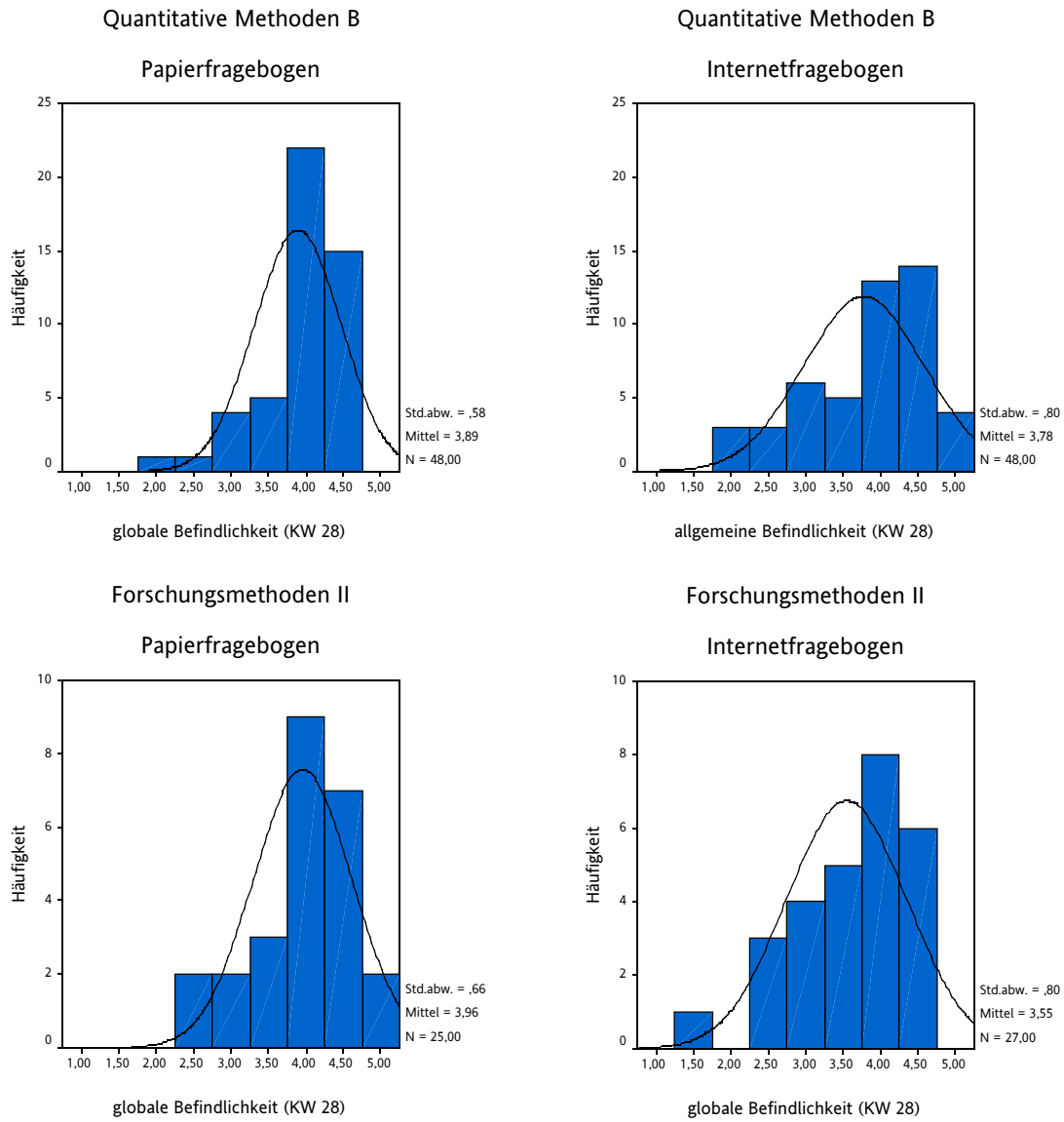


Abbildung K.3: Histogramme der allgemeinen Befindlichkeit im Kalenderwoche 28 nach besuchter Vorlesung und Medium des Fragebogens.

Tabelle K.3.1: Tests auf Normalverteilung in den beiden Vorlesungen

Quantitative Methoden B

		Kolmogorov-Smirnov			Shapiro-Wilk		
	Medium	Statistik	df	Signifikanz	Statistik	df	Signifikanz
allgemeine Befindlichkeit (KW 28)	gerade	,179	48	,001	,865	48	,000
	ungerade	,153	48	,007	,936	48	,012

Forschungsmethoden II

	Matrikelnummer gerade / ungerade	Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistik	df	Signifikanz	Statistik	df	Signifikanz
allgemeine Befindlichkeit (KW 28)	gerade	,207	25	,007	,936	25	,118
	ungerade	,141	27	,177	,947	27	,185

4 These 5: Einfluss von Kovariaten

Tabelle K.4.1: Deskriptive Statistiken der Rohwerte der Computer- und Internetnutzung nach besuchter Vorlesung und eingesetztem Medium. Kritische Werte nach West, Finch und Curran (1995) sind hervorgehoben.

			Computernutzung ohne Internet		**Internetnutzung**	
Vorlesung	**Medium**	**Kennwert**	**Tage/Woche**	**Stunden/Woche**	**Tage/Woche**	**Stunden/Woche**
Quantitative Methoden B	Papier	*n*	51	51	51	51
		M	2..0	7.00	4.62	5.82
		Minimum	0.0	0.00	0.0	0.00
		Maximum	7.0	150.00	7.0	30.00
		s	1.71	21.21	1.94	6.80
		Schiefe	1.17	6.45	-0.49	**2.40**
		Kurtosis	1.16	**43.56**	-0.755	5.77
	Internet	*n*	52	52	52	52
		M	2.08	5.71	4.21	5.19
		Minimum	0.0	0.00	1.0	0.50
		Maximum	7.0	30.00	7.0	50.00
		s	1.77	5.95	1.78	7.40
		Schiefe	1.45	**2.16**	-0.22	**4.74**
		Kurtosis	1.39	5.22	-0.86	**26.97**
Forschungs-methoden II	Papier	*n*	25	25	25	25
		M	3.16	8.14	4.74	3.66
		Minimum	0.5	2.00	1.5	.50
		Maximum	7.0	25.00	7.0	12.00
		s	2.02	7.13	1.89	2.98
		Schiefe	0.67	1.32	-0.34	1.72
		Kurtosis	-0.81	0.77	-1.22	2.62
	Internet	*n*	29	29	29	29
		M	2.93	8.17	4.85	4.13
		Minimum	1.0	1.00	0.5	0.20
		Maximum	6.0	30.00	7.0	12.00
		s	1.58	8.19	2.10	3.16
		Schiefe	0.59	1.77	-0.86	0.80
		Kurtosis	-0.87	2.25	-0.47	-0.16

5 Frage 6: Prüfungsnoten

Tabelle K.5.1: Deskriptive Statistiken der drei Gruppen, in die die Studierenden im Grundstudium je nach ihrer Teilnahmehäufigkeit eingeordnet wurden. Kritische Werte für Schiefe und Kurtosis nach West, Finch und Curran (1995) sind hervorgehoben.

Deskriptive Statistiken der einzelnen Gruppen

Erreichte Punktzahl in der Klausur

Gruppe nach Teilnahme	N	Mittelwert	Median	Std.-Abw.	Schiefe	Kurtosis
nicht teilgenommen	47	43,63	47,50	13,646	-1,09	,60
bis sechsmal	50	51,44	53,25	8,926	-,43	-,40
mehr als sechsmal	74	53,07	56,00	9,212	**-2,49**	**9,16**
Insgesamt	171	50,00	53,00	11,214	-1,58	3,09

Die Voraussetzungen zur Berechnung einer ANOVA zum Vergleich der Prüfungsnoten zwischen den Gruppen waren nicht erfüllt. Dagegen sprachen folgende Ergebnisse:

1. Die Verteilungen verletzen die von West, Finch und Curran (1995) empfohlenen Kennwerte zur Bestimmung der Normalverteilung in der Gruppe *mehr als sechsmal teilgenommen* (Tabelle K.5.1).
2. Die Fehlervarianzen der verschiedenen Gruppen können nicht als homogen betrachtet werden, da der Levene Test ein signifikantes Ergebnis ergab ($F(2,168) = 6{,}67$; $p < 0.01$).

Da zudem die Stichprobenumfänge in den einzelnen Gruppen unterschiedlich waren, konnte die Robustheit der ANOVA gegen Verletzung ihrer Voraussetzungen nicht angenommen werden. Als Ersatz empfiehlt Bortz (1993, S. 263) die verteilungsfreie Kruskal-Wallis-Rangvarianzanalyse.

Anhang L: Regressionsmodelle

Tabelle L.1: Regressionsmodell der Teilnahmehäufigkeit bei Internetbefragungen auf Computer- und Internetnutzung im Grundstudium.

Teilnaheme Internet, Quantitative Methoden B[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	2,561	,110		23,365	,000
	allgemeine Computernutzung	-4,537E-02	,146	-,033	-,311	,756
	allgemeine Internetnutzung	,262	,149	,189	1,762	,081

a. Abhängige Variable: Teilnahme Internet (Semesterverlauf, ohne KW 23)

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,178[a]	,032	,012	1,110

a. Einflußvariablen : (Konstante), allgemeine Internetnutzung, allgemeine Computernutzung

Tabelle L.2: Regressionsmodell der Teilnahmehäufigkeit bei Internetbefragungen auf Computer- und Internetnutzung im Hauptstudium.

Teilnahme Internet, Forschungsmethoden II[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	2,508	,157		15,928	,000
	allgemeine Computernutzung	1,877E-02	,211	,013	,089	,929
	allgemeine Internetnutzung	,119	,188	,093	,631	,531

a. Abhängige Variable: Teilnahme Internet (Semesterverlauf, ohne KW 23)

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,098[a]	,010	-,029	1,145

a. Einflußvariablen : (Konstante), allgemeine Internetnutzung, allgemeine Computernutzung

Tabelle L.3: Regressionsmodell der Teilnahmehäufigkeit bei papierbasierten Befragungen auf Computer- und Internetnutzung im Grundstudium.

Teilnahme Papier, Quantitative Methoden B[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	3,033	,100		30,334	,000
	allgemeine Computernutzung	-3,210E-02	,133	-,026	-,241	,810
	allgemeine Internetnutzung	-5,788E-02	,136	-,046	-,426	,671

a. Abhängige Variable: Teilnahme Papier (Semesterverlauf, ohne KW 23)

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,061[a]	,004	-,016	1,013

a. Einflußvariablen : (Konstante), allgemeine Internetnutzung, allgemeine Computernutzung

Tabelle L.4: Regressionsmodell der Teilnahmehäufigkeit bei papierbasierten Befragungen auf Computer- und Internetnutzung im Hauptstudium.

Teilnahme Papier, Forschungsmethoden IF[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	3,580	,098		36,661	,000
	allgemeine Computernutzung	-,131	,131	-,144	-1,001	,322
	allgemeine Internetnutzung	-,117	,117	-,144	-1,000	,322

a. Abhängige Variable: Teilnahme Papier (Semesterverlauf, ohne KW 23)

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,234[a]	,055	,017	,710

a. Einflußvariablen : (Konstante), allgemeine Internetnutzung, allgemeine Computernutzung

Anhang M: Mittelwertsverläufe über das Semester

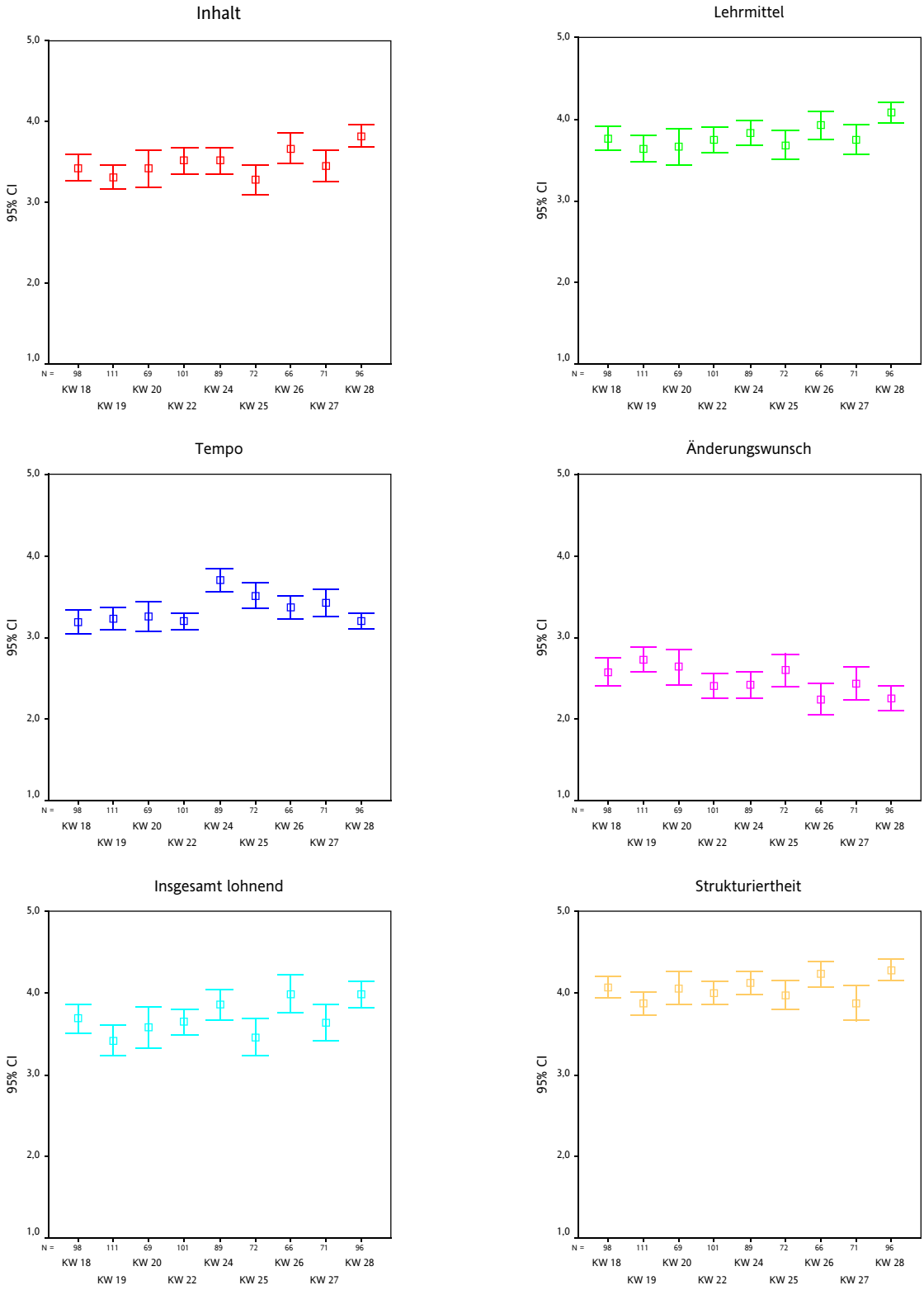


Abbildung M.1: Mittelwertsverläufe der abhängigen Variablen im Grundstudium (Auswertungsdesign 1 und 2).

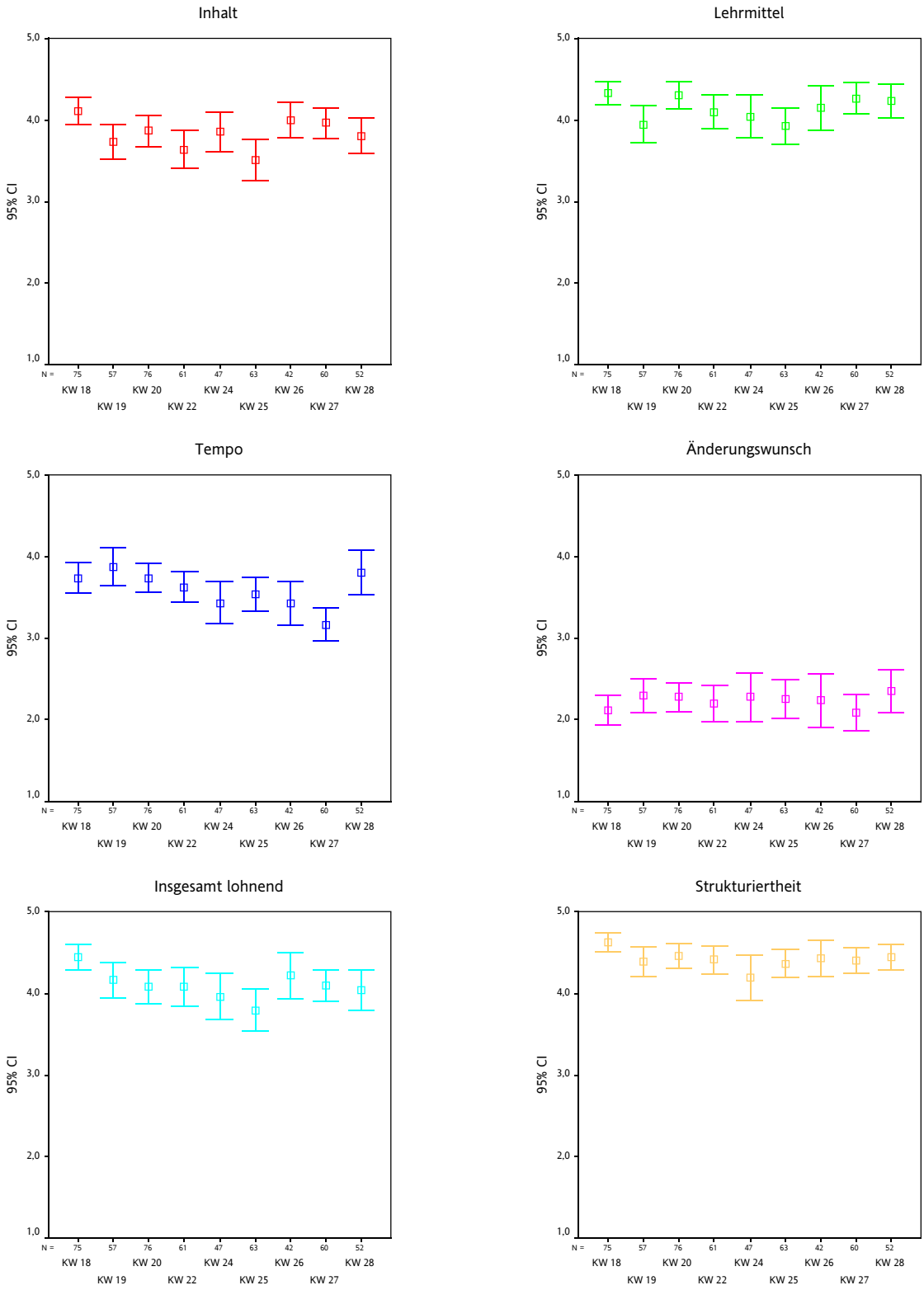


Abbildung M.2: Mittelwertsverläufe der abhängigen Variablen im Hauptstudium (Auswertungsdesign 3 und 4).

Versicherung nach § 19 Abs. 7 der Prüfungsordnung

Hiermit versichere ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Trier, am 31.08.2003