TU

TECHISCHE UNIVERSITÄT WIEN

DIPLOMARBEIT

Qualität von Studium und Lehre an der Universität

Ein Handbuch und eine Anleitung zur möglichen Berechnung

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Diplom-Ingenieurs unter der Leitung von

Professor H. P. Osanna
Institutsnummer: 311
FERTIGUNGSTECHNIK

eingereicht an der Technischen Universität Wien

Fakultät für Maschinenwesen und Betriebswissenschaften

von

Stefan Keckeis
Matrikelnummer: 9126553
Pontenstr. 20, 6890 Lustenau

Wien, am _____

_____
eigenhändige Unterschrift

TU

TECHISCHE UNIVERSITÄT WIEN

Ich habe zur Kenntnis genommen, dass ich zur Drucklegung meiner Arbeit unter der Bezeichnung

DIPLOMARBEIT

nur mit Bewilligung der Prüfungskommission berechtigt bin.

Ich erkläre weiters an Eides statt, dass ich meine Diplomarbeit nach den erkannten Grundsätzen für wissenschaftliche Abhandlung selbständig ausgeführt habe und alle verwendeten Hilfsmittel, insbesondere die zugrunde gelegte Literatur genannt habe.

Weiters erkläre ich, dass ich dieses Diplomarbeitsthema bisher weder im In- noch im Ausland (einer Beurteilerin / einem Beurteiler zur Begutachtung) in irgendeiner Form als Prüfungsarbeit vorgelegt habe und dass diese Arbeit mit der vom Begutachter beurteilten Arbeit übereinstimmt.

Wien, am _____ _____

eigenhändige Unterschrift

Qualität von Studium und Lehre an der Universität

1. Einleitung:

Grundsätzlich stellt sich die Frage, warum gerade in den letzten Jahren die Diskussion und die Aufmerksamkeit gegenüber den Universitäten, genauer gesagt gegenüber ihrem Aufgabengebiet, der Lehre, in verstärktem Maße zugenommen haben. Um diese Frage ausreichend zu beantworten, muss auf die geschichtliche Entwicklung der Universität und der Rahmenbedingungen eingegangen werden. War die Universität früher gekennzeichnet durch die Vermittlung relativ „elitärer Bildungsgüter“ an einen relativ kleinen Adressatenkreis, so ist ihr Bild heute doch im wesentlichen geprägt durch eine berufsausbildende Funktion für eine wachsende Studentenschaft. Genau diese Tatsache aber findet ihren Niederschlag in den unterschiedlichsten Bereichen universitärer Aufgabenerfüllung.
Es darf nicht wundern, wenn Universitäten unter der immer größer werdenden Last von Studierenden, nebst der Verpflichtung den immer schneller werdenden wissenschaftlichen Fortschritt und eine qualitativ hochwertige Ausbildung sicherzustellen, an kapazitive Schranken, sowohl personeller, finanzieller als auch verwaltungstechnischer Art stoßen. Diese prekäre Situation wird auch nicht durch den Umstand gebessert, dass sich das Bildungssystem mit in den letzten Jahren immer geringeren Zuwendungen seitens der öffentlichen Hand gegenüber sieht. Um dieser sorgenvollen Entwicklung entgegenzuwirken, bedarf es des umsichtigen und richtigen Einsatzes beschränkter Ressourcen, dass heißt einer Verwendung derselben, die für den Entscheidungsträger deutlich macht, warum, wo und wie viel investiert wird und mit welchem Erfolg. Die Wissenschaftsbereiche sollen also einer Art Prüfung bezüglich ihres Bedarfes und Leistungsfähigkeit unterzogen werden und dies soll als Grundlage zukünftiger Mittelzuwendungen dienen. Diese nicht neue Idee der Evaluation der Leistungsfähigkeit in Forschung und Lehre, eines formellen Bewertungsverfahren, das genau definierten Prinzipien gehorcht, birgt den Vorteil in sich, nicht nur gegenüber sich selbst und dem Steuerzahler Rechenschaft über den Einsatz der verwendeten Mittel abzulegen, sondern bietet auch die Möglichkeit den tatsächlichen Stand heutiger Wissenschaft zu verifizieren und erlaubt strategische Planung und Schwerpunktsentscheidungen basierend auf den Daten der Erhebung. Dieses Controlling der Mittelzuweisung kann und wird aber nur funktionieren, wenn Evaluation als kritische Untersuchung und Bewertung von Leistungen, Personen, Personengruppen oder Institutionen verstanden wird deren praktische Relevanz und Durchführung sich nach den Zielen, Funktionen, Strukturen und Aufgaben der zu untersuchenden Einheiten richtet.
Evaluation bedeutet also: Erstens Leistung nach bestimmten vorher festgesetzten Kriterien zu messen und zweitens ein Bewerten, ob einerseits die Kennzahlen den (Qualitäts-)Erwartungen und Anforderungen genügen, und andererseits ob sich die Messungen innerhalb der festgesetzten (Qualitäts-)Toleranzen bewegen. Dabei ist zu beachten, dass jede Evaluation ad Absurdum geführt und in keinem Verhältnis zu ihrem Aufwand stehen würde, wenn ihre Aussagekraft auf der bloßen Feststellung des Ist-Zustandes beschränkt wäre, um in weiterer Folge lapidare Empfehlungen etwaiger Korrekturmaßnahmen einzuleiten. Vielmehr ist ihr Aufgabengebiet so zu sehen, dass sie gleichermaßen Kontroll- als auch Planungsinstrument darstellt, mit dessen Hilfe gezielt Entwicklungen gefördert und überwacht werden können. Dafür müssen auch die nötigen Rahmenbedingungen, Strukturen und Mechanismen

geschaffen werden, die eine Entwicklung initiieren, aufrechterhalten und unterstützen können.

2. Ziele der Universität und ihre Interessensgruppen:

Solange nicht eine einigermaßen tragfähige Übereinstimmung der relevanten Gruppen über die Ziele der jeweiligen Institution oder ihrer Untereinheiten hergestellt ist, sind Evaluationen sinnlos, weil sie nicht in Entwicklung übergeführt werden können. Das Problem hierbei ist nicht das Fehlen von Zielen, sondern das Existieren von Zielen in einer lediglich sehr globalen, allgemein gehaltenen und teilweise unscharf formulierten Form, dass daraus in der Regel keine oder nur schwierig aussagekräftige Qualitätskriterien ableitbar sein werden. Um nun zu besagten Zielvorgaben und in weiterer Folge zu Kriterien zu gelangen, sind zwei Schritte von Nöten:

- Feststellen der relevanten Gruppen für den Zielvorgabeprozess
- Einladung der relevanten Parteien und Schaffung eines Plenums, in dem Zielvorstellungen transparent und als Basis zur Diskussion zugänglich gemacht werden.

2.1. Feststellen der relevanten Gruppen für den Zielvorgabeprozess:

Dies scheint noch die einigermaßen einfachere Aufgabe von beiden Problemlösungsschritten zu sein. Zunächst aber sollen für die nachstehende Ausführung die Begriffe Grundsatz, Ziel und „relevante“ Interessensgruppe hinsichtlich einer Zielvorgabenfestsetzung der unmissverständlichen Klarheit wegen „neu“ definiert und voneinander abgegrenzt werden:
In weiterer Folge soll unter **Grundsatz** die von der Gesellschaft oder Öffentlichkeit gemeinhin als richtig empfundene, nicht prüfbare Richtlinie verstanden werden, nach der man sein Handeln auslegt. Er gilt als quasi zeitlich stationär. Diese Überlegung scheint legitim, da es zumindest zu seiner Änderung einer oft langwierigen Neuordnung der Wertigkeiten in der Gesellschaft bedarf.
Dagegen sind **Ziele** die prüfbaren Vorstellungen eines zukünftigen Ergebnisses von bestimmten Interessensgruppen, deren Festlegungen sich mit der Zeit ändern können und daher wandelbar sind.
Als **relevant** werden jene Interessensgruppen angesehen, bei denen eine Änderung der Festsetzung (Ziele) zu einer direkten Betroffenheit führen würde.
Prüfbar sind nur jene Festsetzungen (Ziele), die über relevante Interessensgruppen verfügen und über definierte (Prüf-) Kriterien verfügen.
Aus diesen Definitionen ist folgendes ersichtlich:
Grundsätze sind im Gegensatz zu Zielen nicht prüfbar. Sie verfügen aufgrund der oben festgesetzten Definitionen nicht über eine für die Prüfung notwendige, relevante Interessensgruppe, da sie als über die Zeit beständig, quasi unveränderbar angesehen werden, oder besitzen keine erforderlichen Prüfkriterien. Auch scheint diese Art von Festlegung sinnvoll, da die Interessensgruppe von Grundsätzen die Öffentlichkeit, die Gesellschaft darstellt, deren Einladung zur Zielvorgabenbestimmung nicht praktikabel anmutet.
Ziele freilich sind aufgrund meist spezifischerer und exakterer Vorgaben bedingt durch kleinerer Interessensgruppen prinzipiell einer Prüfung zugänglich und bedürfen nicht zuletzt wegen ihrer Wandelbarkeit der Anforderungen einer ständigen Überprüfung erstens der Zielvorstellungen und zweitens der Erfüllung dieser, welches die Prüfung im eigentlichen Sinne nach übereingekommenen Qualitätskriterien ausmacht.

Es können fünf Interessensgruppen, die mit dem Aufgabengebiet universitäre Lehre und Ausbildung mehr oder weniger in Kontakt kommen, unterschieden werden. Bezeichnender Weise verknüpfen diese Gruppen oft ungleiche, teilweise konträre Zielvorstellungen mit dieser Institution –Universität- oder aber ihre verschiedenen Prioritätensetzungen und Gewichtungen bei doch einigermaßen homogenen Zielen verhindern oder stören ein gemeinsames Vorgehen und Durchsetzen derselben. Auch ihre Möglichkeiten der Berücksichtigung und Einflussnahme sowohl auf universitäre Zielformulierung, wie auch bei der Überprüfung und Steuerung sind im höchsten Maße durch eine gewisse Unverhältnismäßigkeit und Ungleichgewicht gekennzeichnet.

Der **Lehrkörper** als erstgenannte Interessensgruppe bildet zusammen mit der Studentenschaft als weiterer Partei den Kernbereich universitärer Lehre. Auf ihm lastet der Großteil der Verantwortung bezüglich der qualitativ wie auch quantitativ gestiegenen Herausforderung im Bereich der Lehre, dessen Hauptaugenmerk vor allem auf der Aufgabenerfüllung, nämlich den Anforderungen entsprechend ausgebildete Akademiker für die Wirtschaft bereit zu stellen, liegen sollte. Das heißt, der Lehrkörper sollte dafür Sorge tragen, universitäre Lehre aktuell bezüglich ihres Inhaltes am Stand des wissenschaftlichen Fortschrittes orientiert zu halten und Studieninhalte an neue Anforderungen der Wirtschaft und Technik anzupassen oder gar neue Studiengänge einzuführen. Verbunden mit der Aufgabenerfüllung sind also der Lehre gewissermaßen schon Zielvorstellungen inhärent, die als Randbedingungen für die Lösung des Problems richtige Lehre angesehen werden können. Neben dieser Funktion verfolgt der Lehrkörper noch andere Interessen, die nicht in direkter aber doch im weiteren Sinne die Qualität der Lehre bestimmen und deren Bestand für die Zukunft sichert, nämlich die wissenschaftliche Forschung und die Förderung und Heranbildung des akademischen Nachwuchs. Ihre Möglichkeiten, Ziele zu formulieren und auf bereits bestehende Einfluss zu nehmen, sind innerhalb der gesetzlich vorgeschriebenen Bestimmungen durch diverse Organe nicht zuletzt innerhalb des akademischen Senates wahrnehmbar. Die Steuerung der zu setzenden Maßnahmen und die Kontrolle der Wirkung durch den Lehrkörper entspricht der bisher stattfindenden traditionell internen Qualitätskontrolle im Universitätsbereich, die jedoch nicht alle Bereiche universitärer Aufgabenerfüllung in gleicher Weise betreffen und in der Regel keinem objektiven Maßstab unterliegen (so werden z.B. Ein- bzw. Weiterbestellungen eher nach Forschungsqualifikationen denn nach Lehrqualifikation getätigt).

Die **Studentenschaft** als zweite Interessensgruppe besitzt nur bedingt Möglichkeiten (als Vertreter der ÖH in verschiedenen Gremien) ihre Vorstellungen und Wünsche ,soweit vorhanden, hinsichtlich der Lehre zu äußern. Dieses Ungleichgewicht, sich in das Thema einzubringen, erscheint aufgrund einer sachlichen Unreife auch teilweise berechtigt, denn zuerst müssen die Studenten bevor sie am Prozess des Erkennens teilhaben können, die Grundlagen des Faches methodisch und dem Wesen nach ergründen. Jedoch betrifft die Qualität der Lehre nicht nur inhaltliche Aspekte. Dem Wunsch nach einer methodisch und didaktisch wertvollen Darbietung und damit zu einer Steigerung der formalen Qualität der Lehre muss auf breiterer Basis nachgegangen werden und eine Differenzierung des Mitspracherechts der Studenten überlegt werden.

Die **Wirtschaft** als weiterer Interessensvertreter ist darauf Bedacht ihren Bedarf für hochqualifizierte Arbeitskräfte mit fundiertem Fachwissen möglichst schnell sicherzustellen. Während das Spektrum der speziellen Kenntnisse und Fähigkeiten, welche die Arbeitswelt in den unterschiedlichen Berufsbildern fordert, in informellen Gremien z. B. Fakultätentag oder Wirtschaftskammer noch relativ problemlos

abgesteckt werden kann, komplettieren Schlagworte wie praxisorientiert, flexibel, Selbständigkeit bei Problemlösung, Fähigkeit zu analytischem Denken die Vorstellung, welche von Akademikern erwartet wird. Der Unterschied zwischen den Anforderungen, die hier beispielhaft einmal mit Fachkenntnissen und zum anderen mit den oben erwähnten Schlagwörtern genannt wurden, liegt, es sei hier der Klarheit wegen nochmals erwähnt, in der zeitlichen Beständigkeit der Absichten. So müssen Kenntnisse, die gelehrt werden, einer ständigen Prüfung auf Aktualität bezüglich des wissenschaftlichen Fortschrittes geprüft werden, und die Entscheidung darüber getroffen werden, ob sich das Berufsprofil geändert hat, und damit andere Schwerpunkte in der Lehre gesetzt werden müssen. Ihre Zielsetzung unterliegt einer zeitlichen Änderbarkeit und ihre Wirksamkeit kann zumindest auf der Grundlage des Vorhandenseins bestimmter Lehrinhalte festgestellt werden, wohingegen Selbständigkeit sowie analytisches Denken etc. dauerhaft und gemeinhin als erstrebenswerte Eigenschaften erkannt werden und dadurch ihnen ein eher **grundsätzlicher** Charakter anhaftet, deren Prüfbarkeit nicht gegeben oder aber nur mit einem verhältnismäßig großen Aufwand zu bewerkstelligen ist. Die Grenzen sind nicht immer einfach zu ziehen und verschwimmen auch zuweilen, sind jedoch für das Funktionieren eines Qualitätssicherungssystems Voraussetzung.
Steuerung durch den Arbeitsmarkt ist disziplinspezifisch unterschiedlich stark ausgeprägt, je nach Dringlichkeit und Bedeutung des erkannten Arbeitskräftemangels.
Ein weiteres Instrument der Einflussnahme manifestiert sich über die Ressourcenverteilung. Die Industrie finanziert Forschungs- und Entwicklungsprojekte und gratifiziert damit Institute, deren Forschungs- und Ausbildungs-Output sie als adäquat ansieht. Im Rahmen solcher Projekte werden zudem Diplomarbeiten geschrieben, Doktoranden und wissenschaftliche Mitarbeiter finanziert, die auch Lehraufgaben übernehmen.
Der **Staat** als vierte Interessensgruppe sieht seine Aufgabe darin einerseits, mangelnder Kommunikation zwischen Arbeitsmarkt und Universität ausgleichend entgegenzuwirken und andererseits seinerseits Aktivitäten und innovatorische Impulse zu fördern, die eine gedeihliche Weiterentwicklung der Gesellschaft versprechen. Die bekannte und aus vergangener Sicht als kritisch zu beurteilende Prioritätensetzung des Staates, die Tore der Universitäten für jedermann zu öffnen, der über die nötige Berechtigung verfügte, um einem Studium nachzugehen, mag für sich genommen eine verständliche Überlegung für eine Investition in den Standort und die Zukunft Österreichs gesehen werden, zumal Österreich ein Land knapper Ressourcen ist, und seine Wettbewerbsfähigkeit und wirtschaftliche Schlagkraft sich vornehmlich auf das technologische „Know-how“, das heißt eine gezielte Förderung des menschlichen Kapitals des Wissens, gründet. Kritisch bleibt aber anzumerken, dass die eigenen Ziele und die der Wirtschaft, nämlich Qualität der Ausbildung zu fördern und in geeigneter Quantität möglichst schnell dem Markt zur Verfügung zu stellen, durch den Massenandrang auf die Universitäten torpediert wurden. Ob dies nun durch Versäumnisse geschah, geeignete und der Situation entsprechende kapazitätserweiternde Maßnahmen in baulicher sowie personeller Hinsicht durchzuführen oder aber als ein vorübergehend zu akzeptierender Zustand höherer Belastung gesehen wurde, dessen zeitliches Ausmaß fehleingeschätzt wurde, bleibt anderen zu beurteilen. Die Auswirkungen solchen Handels sind mannigfaltig und spiegeln sich in unterschiedlichen Bereichen wider:

- Der erhöhten Lehrverpflichtung wurde mit einer Erhöhung des Lehrdeputats entsprochen, was sich negativ auf den Bereich Forschung auswirkte, wodurch vor allem die Qualifikationschancen des wissenschaftlichen Nachwuchses in

unvertretbarer Weise beeinträchtigt werden und damit die Weiterentwicklung ganzer Fächer gefährdet wird.

- Es entstehen Engpässe bei der Bereitstellung von Mitteln wie Praktikumsplätze, Laborplätze, Bücher etc., die fast zwangsläufig zu einer Studienzeitverlängerung führen.
- Die Möglichkeit, das Lehrangebot in der Raum- und Zeitdimension zu organisieren, werden durch die Summe der Lehrdeputate, die Zahl und Kapazität der Räume und durch die Ausstattung begrenzt. Die Verteilung der Veranstaltungen auf Räume und Zeiten kann die Nachfrage der Studierenden beeinflussen und zu ungleichmäßiger Belastung der Lehrenden und Veranstaltungen führen. Überfüllung von, Wartezeiten und Wartelisten für Lehrveranstaltungen, die zu Engpässen im Studienablauf führen können, sind die Folge.
- Vorlesungen werden nicht mehr nach inhaltlichen als vielmehr nach zeitlichen oder platztechnischen Gründen ausgesucht.
- Große Hörerzahlen verhindern nicht selten das selbständige, analytische Arbeiten der Studenten.

Die Steuerung und Lenkung der eigenen Zielformulierungen werden über die bürokratische Hierarchie von Beamten und Rektoren nach unten weitergegeben. Es bestehen des weiteren die Möglichkeiten über die Studienordnung, die Prüfungsordnung, sowie über Leistungsvereinbarungen Einfluss auf bestimmte Inhalte zu nehmen. Vor allem die im Gesetz geregelten Bestimmungen der Studien- wie auch der Prüfungsordnung geben dem Staat die Möglichkeit minimal-einheitliche Standards in Studium und Prüfungswesen auf sich differenzierenden Universitäten festzuschreiben.
Als letzte Interessenspartei sei noch die **Gesellschaft** oder die Öffentlichkeit genannt. Ihr Interesse, dass die Universität für eine gedeihliche Weiterentwicklung der Gesellschaft zu sorgen hat, ist ein schwer zu kontrollierendes und unterliegt weitestgehend dem Vertrauensgrundsatz und dem Ethos der Professorenschaft. Dennoch beschränkt sich ihre Kontrollwirksamkeit nicht nur auf Vertrauen. Kritische mediale Berichterstattung über die Effizienz in der Verwendung von Ressourcen und über die Effektivität in der Erreichung von gesetzten Zielen ist ein probates Mittel das Funktionieren öffentlicher Einrichtungen zu gewährleisten.
Diese fünf Interessensgruppen müssen in der Folge ihre Wünsche und Absichten in einem Plenum vorbringen und einer daraus sich entwickelnden, übergeordneten, gemeinsamen Zielvorstellung ein- bzw. unterordnen, wobei die Gesellschaft über gesetzlich festgelegte Grundsätze und Ziele indirekt bei der Entstehung des Zielvorgabeprozesses eingebunden ist, die sozusagen als Grundkonstrukt für die Feinformulierungen angesehen werden kann, innerhalb derer sich Konkretisierung abspielen kann:

„Ziele:

§ 1. Die Universitäten sind berufen, der wissenschaftlichen Forschung und Lehre, der Entwicklung und der Erschließung der Künste sowie der Lehre der Kunst zu dienen und hiedurch auch verantwortlich zur Lösung der Probleme des Menschen sowie zur gedeihlichen Entwicklung der Gesellschaft und der natürlichen Umwelt beizutragen. Universitäten sind Bildungseinrichtungen des öffentlichen Rechts, die in Forschung und in forschungsgeleiteter akademischer Lehre auf die Hervorbringung neuer wissenschaftlicher Erkenntnisse sowie auf

die Erschließung neuer Zugänge zu den Künsten ausgerichtet sind. Im gemeinsamen Wirken von Lehrenden und Studierenden wird in einer aufgeklärten Wissensgesellschaft das Streben nach Bildung und Autonomie des Individuums durch Wissenschaft vollzogen. Die Förderung des wissenschaftlichen Nachwuchses geht mit der Erarbeitung von Fähigkeiten und Qualifikationen sowohl im Bereich der wissenschaftlichen und künstlerischen Inhalte als auch im Bereich der methodischen Fertigkeiten mit dem Ziel einher, zur Bewältigung der gesellschaftlichen Herausforderungen in einer sich wandelnden humanen und geschlechtergerechten Gesellschaft beizutragen. Um den sich ständig wandelnden Erfordernissen organisatorisch, studien- und personalrechtlich Rechnung zu tragen, konstituieren sich die Universitäten und ihre Organe in größtmöglicher Autonomie und Selbstverwaltung.

§ 2. Die leitenden Grundsätze für die Universitäten bei der Erfüllung ihrer Aufgaben sind:

1. Freiheit der Wissenschaften und ihrer Lehre (Art. 17 des Staatsgrundgesetzes über die allgemeinen Rechte der Staatsbürger, RGBl. Nr. 142/1867) und Freiheit des wissenschaftlichen und des künstlerischen Schaffens, der Vermittlung von Kunst und ihrer Lehre (Art. 17a des Staatsgrundgesetzes über die allgemeinen Rechte der Staatsbürger);
2. Verbindung von Forschung und Lehre, Verbindung der Entwicklung und Erschließung der Künste und ihrer Lehre sowie Verbindung von Wissenschaft und Kunst;
3. Vielfalt wissenschaftlicher und künstlerischer Theorien, Methoden und Lehrmeinungen;
4. Lernfreiheit;
5. Berücksichtigung der Erfordernisse der Berufszugänge;
6. Mitsprache der Studierenden, insbesondere bei Studienangelegenheiten, bei der Qualitätssicherung der Lehre und der Verwendung der Studienbeiträge;
7. nationale und internationale Mobilität der Studierenden, der Absolventinnen und Absolventen sowie des wissenschaftlichen und künstlerischen Universitätspersonals;
8. Zusammenwirken der Universitätsangehörigen;
9. Gleichstellung von Frauen und Männern;
10. soziale Chancengleichheit;
11. besondere Berücksichtigung der Erfordernisse von behinderten Menschen;
12. Wirtschaftlichkeit, Sparsamkeit und Zweckmäßigkeit der Gebarung.“

Studentenschaft und Wirtschaft werden oft als Kunden der Universität angesehen je nachdem, ob man sich auf den Standpunkt von Universität als Produktions- oder Dienstleistungsunternehmen stellt. Einerseits erwerben die Studenten das Wissen und stellen in diesem Fall die Anwender oder Kunden dar, andererseits kann man

auch die Unternehmen, welche die Absolventen einer Universität einstellen, als Kunden bezeichnen. In diesem Falle stellen die Universitäten die Produktionsunternehmen dar und die Absolventen wären die kennzeichnenden Produkte.

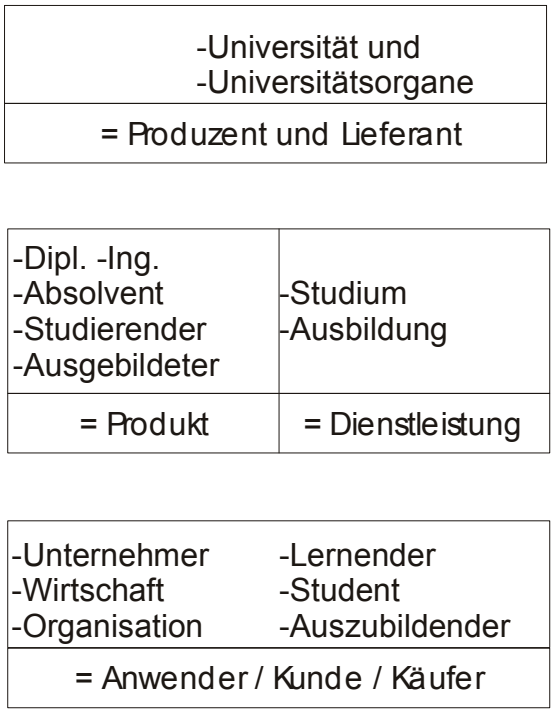


Abbildung 1: Die Universität als Produzent bzw. Lieferant

Die Feststellung der Absolventen, die mit sich eine Fülle von Qualitätsmerkmalen verbinden, auf Erfüllung der gesetzten Anforderungen beinhaltet als Produkt 2 unterschiedliche Arten der Daten zur Prüfung:

- Objekt bzw. Produktdaten,
- Methoden bzw. Prozessdaten.

Hier scheint zum klareren Verständnis ein Vergleich mit der Werkstoffkunde angebracht. Auch hier werden Probestäbe der verschiedenen Stähle und Werkstoffe unterschiedlichen Prüfungen Kerbschlagbiegeversuch, Zugversuch usw. unterzogen. Es werden Mindeststandards festgelegt, die gleichermaßen Abnahmekriterien der Werkstoffe darstellen, bei derer Unter- bzw. Überschreitung ein Mangel an Qualität der Probestäbe festgestellt werden kann. Obgleich die Werkstoffe nach festgelegten einheitlichen Prozessen erzeugt bzw. vergütet werden, besitzen die Proben doch unterschiedliche Festigkeitswerte und voneinander in gewissem Rahmen abweichende Eigenschaften. Dies kann auf Vorgänge innerhalb der Prozessführung zurückgeführt werden, deren Beherrschbarkeit nicht vollständig oder gar nicht gegeben ist und zum zweiten auch zum Teil zufälligen auf jeden Fall aber statistischen Charakter besitzen. So werden z. B. die Eigenschaften der Metalle durch zahlreiche Größen bestimmt z. B: Ausrichtung und Größe der Kristallite, Kristallisationsfehler, Diffusionsfähigkeit von Legierungselementen bei der Veredelung der Stoffe usw., auf die zwar durch geeignete Prozessführung Einfluss genommen werden kann (z. B. kleinere Größe der Kristallite durch raschere Erstarrung, Feinkornhärtung) deren konkrete Ausbildung und Ausformung jedoch sich in einem bestimmten Spektrum bewegen kann. Grundsätzlich lassen sich also Wirkzusammenhänge auf bestimmte Größen und Eigenschaften der Werkstoffe von gewissen Prozessabläufen durch Versuche und Messungen bestätigen. Für die

Prüfung und Qualitätssicherung von Werkstoffen bedeutet dies wie oben schon angedeutet zweierlei:

- Finden von geeigneten Prozessen, um gewünschte Resultate zu erlangen und Überprüfung der richtigen Handhabung und Wirksamkeit des Prozesses durch Kontrolle und Dokumentation der **Prozessdaten** bzw. Verteilung der Produktdaten.
- Feststellen, ob die erwünschten **Produkteigenschaften** und Anforderungen bei jeder Probe erfüllt werden, da trotz geeigneter Maßnahmen und beherrschter Prozesse statistische Abweichungen von den erwarteten Werten auftreten.

Diese Analogie soll einen ersten Überblick über die zu erwartenden Probleme und adäquate Vorgehensweise bei der Beurteilung und Prüfung der Lehre vermitteln, denn wie im vorher erwähnten Fall stellt sich eine ähnliche Situation bei der Evaluierung der Lehre und Qualitätssicherstellung bei den Studenten bzw. Absolventen dar.
Der Prozess des Lernens und Lehrens vollzieht sich in einem offenen System, einem Regelkreissystem, innerhalb dessen durch Setzen von koregierenden Maßnahmen in Verbindung mit paralleler ständiger Überwachung der Studenten auf Basis ihres Wissenstandes die Qualität der Ausbildung gewährleistet werden soll. Zu Prüfen sind dabei vor allem, die Objekt- bzw. Produktdaten, die Aufschluss über die Erfüllung der gestellten Anforderungen geben und die Qualität der Produkte gewährleisten, da eine alleinige Beherrschung und Prüfung des Lehrprozesses keine Sicherheit für eine zufriedenstellende Ausprägung der qualitätsrelevant angesehenen Produkteigenschaften bergen. Ein Artikel über „die Qualität an der Universität“, der von der Fakultät Austauschbau und Messtechnik von der Technischen Universität Wien verfasst wurde, hält jenen Sachverhalt folgendermaßen fest:[1]
„It is presupposed that lectures are the basis of the education at the university. In addition to the selection of the studying material and the way of lecturing, the mediation of the studying material is most important for the quality of a teaching process. The mediation of the studying material is determined by the disposition of the lecturer. The verification of this process, which is only based on the results of examinations, includes also the disposition of the students and therefore this way is not correct. If for instance the course of lectures is conscientiously prepared and excellently presented and the pertinent text book is easily accessible but the student is badly prepared because of various reasons, which are not influenced by the teaching process, then the realized correction of the process on the basis of these results would have been false.
On the other hand an insufficiently prepared and inferiorly presented course of lectures can hardly lead to an appropriate improvement of the knowledge level. If at the examination a considerable difference between the expected knowledge level and the real one is detected a correction is necessary.”
Dabei ist zum Unterschied zu der Prozessprüfung nur wichtig, dass die Produkteigenschaften als qualitätsrelevant eingestuft werden und prüfbar im obig genannten Sinn sind. Eine Abweichung von festgeschriebenen Werten ist aufgrund der nur bedingten Einflussnahme durch die Prozessführung auf die Endqualität möglich.
Beispiele für wünschenswerte **Produkteigenschaften** für das Produkt Absolvent können unter anderem folgende sein:

- Fachwissen und technisches Verständnis den Anforderungen entsprechend

[1] vgl. [1]

- schnelle Auffassungsgabe
- Fähigkeit zur Problemlösung
- Teamfähigkeit
- theoretisches Grundwissen
- praktische Erfahrung
- Flexibilität und Anpassungsfähigkeit
- physische und psychische Belastbarkeit der Probanden
- kurze Dauer des Studiums

Des Weiteren stellt sich auch das Problem, dass von vornherein keine objektiven Maßstäbe für die verschiedenen Kriterien aufliegen, sondern sich diese erst über Einigung der Interessensgruppen intersubjektiv herauskristallisieren und über das Wissen der möglichen Ausprägungen bei einem unveränderten Prozess gefunden bzw. festgeschrieben werden können. So wäre es beispielsweise nicht sinnvoll Abnahmekriterien bei einer neu entwickelten Legierung zu vereinbaren ohne grundlegende Kenntnis über die denkbar erreichbaren Werte.
Mögen auch die erwünschten Zielvorstellungen an einen Absolventen noch so umfangreich und in seiner Summe doch eher unklar definiert sein, so besteht doch der klarer geäußerte Hauptauftrag einer berufsbefähigenden Ausbildung für die Universität, und diesem zumindest auf inhaltlicher Basis nachzukommen. Dass bedeutet aber den Prozess der Lehre:

- **informell, inhaltlich** passend abzustimmen,
- in einer **formal** geeigneten Art und Weise zu präsentieren und
- diesen Prozess in zweckmäßige **Rahmenbedingungen** eingebetet anzubieten.

Diese Schlagworte stellen die Prozessdaten im engeren Sinne, bei bloßer Konzentration auf die Wissensvermittlung dar. Folgende Überlegung muss Leitgedanke des Prozesses Lehre sein:
Wir studieren für die Qualifikation und nicht für die Prüfung. Die Prüfung ist ein Instrument und kein Zweck. Das bedeutet, dass die Prüfungen klar in diagnostische Lernkontrollen mit nachbereitender Korrekturmöglichkeit von Studienverhalten, Lernvorgängen und Kommunikationsabläufe zwischen Lehrenden und Studierenden als erstens **Prozessüberwachungsinstrument** und zweitens laufbahnentscheidende Prüfungen als **Objektkontrollmittel** unterschieden werden sollen. Die letzteren sollen entscheidende Qualifikationsetappen unter den Anforderungen der vorher festgelegten „Vertragsbedingung“ sein.
Die informell, inhaltliche Komponente der Lehre bildet die Schnittstelle zwischen Objekt als Student (Produkt) und Ausbildungsprozess. Anders als bei den Objektdaten unterliegen die Prozessdaten einer Möglichkeit der direkten Einflussnahme, wobei das inhaltlich, informelle Element größtenteils von Professorenschaft, gesetzlichen Bestimmungen und Wirtschaft beeinflusst wird, auf das Formale durch den Lehrkörper und Studentenschaft eingewirkt bzw. beurteilt werden sollte und die Rahmenbedingungen vornehmlich vom Lehrkörper wie auch vom Staat generiert werden.
Die Feststellung der inhaltlichen Sachdienlichkeit erstens zur Berufsbefähigung als auch zum verständlichen Wissenserwerb kann nur über Experten vorgenommen werden, wobei die Studierbarkeit stark von der inhaltlichen Abstimmung des Lehrangebots geprägt ist. Des weiteren wäre eine ständige Überprüfung der Aktualität des Lehrangebots wünschenswert. Dazu: „...Die Lehrbeauftragten legen beim internen Antrag für die Erteilung einer Lehrveranstaltung die Auswertung der Lehrveranstaltungsmatrix des letzten Semesters vor. Gleichzeitig muss die Auswertung der befragten Experten aus Industrie und Wirtschaft sowie der Nachweis

der Einbindung neuer wissenschaftlicher Erkenntnisse vorgelegt werden."[2] Die Voraussetzung der direkten (sowohl internen, im Sinne von universitären Interessensgruppen, als auch externen) Einflussnahme auf Prozessdaten mit überprüfbaren Kriterien bilden die Grundlage für Qualitätskontrolle, -sicherung und Beherrschung des Ausbildungsprozesses.
Auf die Möglichkeiten der Indikatorisierung der Prozess- und Objektdaten und deren Grenzen soll im nächsten Kapitel näher eingegangen werden.

3. Aspekte der Qualität und deren Indikatoren:

3.1. Konzepte der Qualitätskontrolle:

Die Frage nach einer Notwendigkeit der Qualitätskontrolle der Lehre wurde seit je her einstimmig auf allen Ebenen für richtig, wichtig und essentiell angesehen. Jedoch war man sich über die Art und Weise der Durchführung dieser Überprüfung nicht immer einig. Manche Gruppen tendierten mehr zu einer Evaluation auf Basis des sogenannten Instruments „peer-review", einer Kontrolle durch eine externe Expertenschaft, dessen Hauptaugenmerk vermehrt im Erkennen von Schwachstellen innerhalb der Prozessführung im Bereich Studium und Lehre liegt. Dabei wurden nicht nur organisatorisch, strukturelle Seiten des Ausbildungsprozesses, sondern auch finanzpolitische Entscheidungen betrachtet. Auf ausreichende personelle wie auch gerätetechnische Ausstattung, auf qualitative Befriedigung der Ansprüche einer zweckmäßigen, den Anforderungen entsprechende Möglichkeit des Wissenserwerbs vor allem hinsichtlich des erwünschten, wirtschaftlichen Ausbildungsprofils und der Durchführbarkeit des Studiums für die Lernenden in Regelstudienzeit wurde verstärkt Wert gelegt. Die Folge war und ist, dass es sich aus vorstellbaren Gründen um ein zeitlich, wie auch finanziell kostspieliges Verfahren der Qualitätsprüfung handelt.
Ein anderer Ansatzpunkt der Qualitätskontrolle spiegelt sich in der Idee der Indikatoren oder „performance indicators" wider.[3] Sicherlich nicht so geeignet wie das „peer-review" Ausbildung in seiner qualitativen Gesamtheit zu erkennen, bietet es allerdings die Vorteile einer rascheren Evaluation, einer Unabhängigkeit des Ergebnisses von der Zusammensetzung und Auswahl der Expertenschaft und den einfacheren Überblick über den Stand und die Entwicklung von Systemen einzufangen und darzustellen, da eine vergleichsweise raschere Auswertung ermöglicht wird.
Nach Kells gibt es zwei unterschiedliche Arten, sich der Aufgabe der Qualitätsmessung mittels performance indicators zu nähern. Die erste beschreibt die performance indicators als normativen Wertmaßstab, als Gegenüberstellung eines zu erreichenden Ziel- bzw. Sollwertes mit dem realiter erreichten Istwert. Eine solche Definition impliziert bereits das durch die Verwendung der verschiedenen Indizes indirekte Vorhandensein eines sich generierenden Wertemaßstabes für die Qualität der Lehre, der stark von den möglichen Ausprägungen der Indikatoren und deren vermutete wechselseitige Wirkung auf das zu messende Qualitätsziel abhängt, um in weiterer Folge für die Qualität „sinnvolle" Sollwerte und Toleranzgrenzen festzusetzen. Das soll heißen:
Die verschiedenen Indikatoren werden per Beschluss der diversen Interessensgruppen als Subkriterien für die Qualität von Studium und Lehre definiert, ein angezeigtes Abweichen des Indikators vom Sollwert über die Toleranzgrenze

[2] Vgl. [12]
[3] Vgl. [2]

hinaus hätte definitionsgemäß prozesskoregierende Maßnahmen zur Folge, da dies als eine Abeichung von festgesetzten Qualitätsmerkmalen interpretiert wird, obgleich ein wirklicher Wirkzusammenhang mit der Qualität von Studium und Lehre in vielen Fällen nur vermutet und nicht nachgewiesen wird.
Die andere Methode begnügt sich mit Indikatoren, denen grundsätzlich ein Zusammenhang auf das Qualitätsziel zugeschrieben wird, diese jedoch in ihrer Größe und letztendlichen Wirkung auf Bildung nicht in vollem Maße bekannt sind. Diese Indikatoren fungieren als eine Art Merkposten, deren Aussagekraft in der mehr abgeschwächten Form des Hinweises oder des möglichen Erklärungsversuches von Entwicklungen und Zuständen gesehen wird und deren Verwendung sich in der Regel auf eine zeitliche Analyse der Fortschreibung dieser Indikatoren bezieht.
In dem Artikel „Methodische Zugänge zur Qualitätsbestimmung von Lehre und Studium“ von Heiner Treinen[4] wird dieser Sachverhalt zur Nutzung der Indikatoren in einer mehr pragmatisch und empirischen Weise folgenderweise festgehalten: „...Dem angedeutetem Dilemma versucht man auf Seiten der Sozialindikatorenforscher zu entgehen, indem eine pragmatische Version an die Stelle des überaus anspruchvollen und gescheiterten normativen Ansatzes gewählt wurde. Die daraus hergeleitete Veränderung des Forschungsprogramms bezog sich auf den Verzicht eines modelltheoretisch begründeten Indikatorensystems. Statt dessen wurden die zugrundeliegenden Indikatoren als Eckwerte für entsprechende Problembereiche betrachtet; das Ziel der nunmehrigen Forschung bestand in einer Dauerbeobachtung, wobei anhand der gleichbleibenden Indizien Trendentwicklungen ablesbar sein sollten, wenn auch ohne die ursprünglich gedachte Möglichkeit, auf dieser Grundlage ein theoretisches Modell empirisch interpretierbar zu machen.“
Im folgenden soll nun immer der empirisch, pragmatische Indikatoreinsatz verstanden werden, da nach Treinen: „Das Programm der Dauerbeobachtung von Qualität der Lehre anhand von Indikatoren ohne vorgegebenes Modell ist nur dann sinnvoll, wenn ein Nachweis darüber vorliegt, dass der zugrundeliegende Realitätsbereich einen empirisch nachvollziehbaren Zusammenhang aufweist;...“gilt und dieser eben nur über Dauerbeobachtung verifiziert werden kann.
Oder anders ausgedrückt, die Bezugsnorm der Indikatoren soll der „ipsative“ und in weiterer Folge, wenn möglich, der „kriteriumsbezogene Maßstab“ nach Ackeren und Hovestadt sein.[5]
Über die Probleme die sich aus dem sozialen Maßstab ergeben können wird an anderer Stelle in Kapitel 4 gesondert eingegangen.
Der Klarheit wegen sollen die drei Bezugsnormen nochmals kurz erläutert werden:

1. Der ipsative Maßstab

Die zu einem Gegenstand der Berichterstattung für eine Region mitgeteilten Befunde, eine Gruppe derartiger Befunde oder der Bericht insgesamt werden mit früheren Befunden für die gleiche Region verglichen und auf dem Hintergrund dieses Vergleiches bewertet. Dieser ipsative Maßstab kann in der Regel für einen großen Teil referierter Gegenstandsbereiche angeboten werden, da derartige Zeitreihendaten zumeist für viele Bereiche vorliegen.

2. Der soziale (vergleichsgruppenbezogene) Maßstab

Die zu einem Gegenstand der Berichterstattung für eine Region mitgeteilten Indikatoren oder eine Gruppe derartiger Befunde können mit den entsprechenden Befunden anderer Regionen verglichen werden.

3. Der kriteriumsbezogene Maßstab

[4] Vgl. [3]
[5] Vgl. [4]

Die zu einem Gegenstand des Indikatorensystems für eine Region mitgeteilten Befunde, eine Gruppe derartiger Befunde oder der Bericht insgesamt werden Zielsetzungen gegenübergestellt, die in Gesetzen, Richtlinien, Vereinbarungen oder in politischen Absichtserklärungen formuliert sind. Die mit diesem Maßstab gewählte Bezugsnorm eignet sich für die Bildungsberichterstattung besonders, weil sie auf qualitative und quantitativ erfassbare Gegenstandsbereiche gleichermaßen anwendbar ist und weil sie – mit Blick auf politische Absichtserklärungen – den Gedanken der Rechenschaftslegung, den die Politik den einzelnen Bildungseinrichtungen nahe zu bringen versucht, auf die Programmatik der Politik selbst anwendet und damit zugleich deren Glaubwürdigkeit festigen kann.

3.2. Besonderheiten, Stärken und Schwächen der Evaluierung mittels Indikatoren:

Indikatoren beziehen sich auf statistische Erhebungen, die zeitlich regelmäßig, periodisch stattfinden sollten, um gleichermaßen einen aktuellen Zustand des Systems einfangen als auch Entwicklungstendenzen vorhersagen zu können. Außerdem soll Ihnen im Hinblick auf das zu erreichende Qualitätsziel eine Handlungsrelevanz im bildungspolitischem Sinne anhaften, deren Relevanz sich nicht nur aus Ihrer Größe selbst, sondern auch durch das wechselseitige Zusammenwirken und Effekte anderer Indikatoren auf das Qualitätsziel bestimmt werden soll.
Indikatoren sollten also damit zumindest die folgenden wichtigen Eigenschaften besitzen:[6] Sie

- basieren auf regelmäßiger, periodischer Erhebung und sind somit aktuell,
- beruhen auf ausreichend großen Stichprobengrößen,
- liefern gültige, zuverlässige und objektive Informationen,
- vermitteln ein Bild aktueller oder möglicher Probleme,
- lassen Querverbindungen untereinander zu bzw. verweisen gezielt auf (mögliche) Zusammenhänge,
- besitzen mit Blick auf bestimmte Ziele und politische Einflussmöglichkeiten Relevanz,
- zeigen Änderungen im Zeitverlauf an.

Fitz-Gibbon / Tymms halten die Eigenschaften die ein Indikator haben sollte durch folgende Definition fest: „Ein Indikator umfasst regelmäßig gesammelte Informationen, um die Leistungsfähigkeit eines Systems zu bestimmen. […]
Der wichtigste Aspekt eines Indikatoren-Systems ist seine Reaktivität: Der Einfluss auf das System."
Bildungspolitisch relevante Indikatoren sollen also nicht in erster Linie Aufschluss über die Leistungsbereitschaft und –vermögen der Studenten als Individualeigenschaft geben, sondern Auskunft über die „Lage" und Anhaltspunkte über die Verteilung dieser Eigenschaften in aggregierter Form als Eigenschaft des zu untersuchenden Systems widerspiegeln und des weiteren Zusammenhänge wie diese Fähigkeiten mit wichtigen wirtschaftlichen und das Bildungssystem betreffenden Variablen verknüpft sind, aufdecken.
Obwohl durch die Verwendung von gemeinsam gefundenen Indikatoren eine zunehmende „Objektivierung" des Messprozesses für die Qualität von Studium und

[6] Vgl. [4]

Lehre im Vergleich zum peer-review vordergründig stattgefunden hat, bleibt doch das subjektiv geartete Problem der Auswahl und Präferenz der einzelnen Indikatoren, die oft mehr durch die Zugänglichkeit der Daten als durch die vermutete Aussagekraft und Relevanz bestimmt wird. Genau jene gewählten Indikatoren stellen zugleich aber das Konglomerat der zu beschreibenden Qualität als auch deren Grenzen dar, innerhalb derer sich die so definierte Qualität bewegen kann. Inwieweit nun diese Art von Qualität nun wirklich das Bild der Realität mit ihren Wechselwirkungen „genügend hinreichend" abbilden kann und was denn schon „genügend hinreichend" sein soll, kann vermutlich nicht geklärt werden und muss sich wohl eines gemeinsamen Findungs- und Einigungsprozesses bemühen, der sich wiederum über die Zeit gesehen als veränderbar darstellen kann. Dazu Treinen: „Qualität des Lebens" ist – genauso wenig wie die Qualität der Lehre – nicht auf wenige Dimensionen rückführbar und nicht als **statisch festlegbares** konzeptuelles System beschreibbar."
Ein weiterer Punkt, der die Sensibilität beim Finden der für entscheidend gehaltenen Indikatoren aufzeigt, besteht in der Tatsache, dass sowohl Methoden als auch Indikatoren eine zeitliche Konstanz ,nicht in ihrer Größe, aber hinsichtlich ihrer Verwendung, aufweisen sollten, um eine empirische Wirksamkeit bezüglich der Qualitätsmessgröße(n) aufzeigen zu können. Rein aus diesen Überlegungen ist vor einer überschnellen, unreflektierten Wahl von Indikatoren, welcher der Komplexität des Messzieles weder im Umfang noch in seiner Dimension gerecht werden können, zu warnen.
Die mittels der Indikatoren zu untersuchenden Einheiten und Institutionen zeigen automatisch ein Bestreben in den gemessenen Größen zu überzeugen. Dies wiederum kann aber zu einer einseitigen, verzerrten Betrachtung der Qualitätsaspekte und im extremen Fall zu einer völligen Außerachtlassung für die Lehrqualität ebenso wichtigen Teilaufgaben und Maßnahmen führen. Ein Versteifen auf einige, wenige Qualitätsindikatoren und deren übermäßige Betonung bzw. nicht Beachtung kann nicht über das grundlegende Problem der sowohl einseitigen Auswahl der Indikatoren für die Qualitätsbeschreibung hinwegtäuschen noch die leider verständliche Gefahr, jene Indikatoren bevorzugt zu behandeln, deren Ergebnisse gut bis hervorragend ausfallen, mindern.
Das bloße Unterstreichen der vordergründigen Qualitätserbringung durch den Verweis auf die positiven Ergebnisse der Indikatoren beinhaltet noch kein Hinweis auf die Wirksamkeit der Indikatoren auf die Erfüllung des zu messenden Qualitätsziels. Nur wenn die Indikatoren empirischen Zusammenhang mit dem Qualitätsziel aufweisen und somit ein Effekt auf das Qualitätsziel vermutet werden kann, gewinnen die Indikatoren Aussagekraft (außer die Indikatoren werden per se als Zielgröße definiert).

3.3. Grundsätzliche Begriffsbestimmungen und Gütekriterien in der Testanalyse:

Basis jeder Auswertung sind Daten, deren Qualität unter anderem von der Art des Messvorgangs beeinflusst wird und auch von der Eigenschaft des Merkmals selbst determiniert wird. So werden Daten mehr augenscheinlicher oder physikalischer Natur z.B. die Körpergröße eines Menschen einfacher zu messen sein als Angaben über Personen mehr soziologischer Ausrichtung wie z. B. Intelligenz, Motivation oder Gesundheitszustand. Die Klassifizierung der erhobenen Daten nach ihrer Qualität erfolgt formal durch die Zuordnung zu den 4 unterschiedlichen Skalenniveaus:[7]

[7] Vgl. [6]

- Nominalskala
- Ordinalskala
- Intervallskala
- Ratioskala

Das Skalenniveau bedingt nicht nur den Informationsgehalt der Daten, sondern ist auch entscheidend für die Operationen, die mit den Daten vollzogen werden können und damit in letzter Folge auch für die Verwendbarkeit von statistischen Modellen.
Es sollen kurz die verschiedenen Skalentypen beschrieben werden:
Die Nominalskala stellt die einfachste Art der Messung dar. Sie begnügt sich mit einer primitiven Klassifizierung qualitativer Merkmale (z.B. Geschlecht, Farbe,...). Merkmalsausprägungen der Klassen werden für die weitere Auswertung mit Zahlen kodiert (z.B. 0...männlich, 1...weiblich), deren Zuordnung willkürlich, aber einheitlich und eindeutig zu erfolgen hat. Es leuchtet ein, dass arithmetische Operationen mit solchen Größen (Addition, Subtraktion, Multiplikation, Division) schon aufgrund der Willkürlichkeit der Zuordnung nicht sinnig wären und eine Auswertung der Daten sich lediglich auf eine Häufigkeitsermittlung der verschiedenen Merkmalsausprägungen beschränken kann.
Die Ordinalskala stellt das nächst höhere Zahlenniveau dar. Die Ordinalskala erlaubt die Aufstellung einer Rangordnung. Die Untersuchungsobjekte können immer nur in eine Rangordnung gebracht werden (z.B. Produkt A ist besser als Produkt B). Die Rangwerte 1.,2. oder 3. sagen nichts über die Abstände zwischen den Objekten aus, es enthält somit keine Information um wie viel das Produkt A besser als Produkt B ist.
Das wiederum nächst höhere Messniveau stellt die Intervallskala dar. Diese zeichnet sich durch gleich große Skalenabschnitte aus. Ein typisches Beispiel ist die Celsiusskala zur Temperaturmessung, bei der der Abstand zwischen Gefrier- und Siedepunkt des Wassers in 100 gleichgroße Abschnitte eingeteilt wurde. Bei intervallskalierten Größen besitzt auch die Differenz im Gegensatz zu nominal- und ordinalskalierten Werten zwischen den Daten Informationsgehalt. Intervallskalierte Daten ermöglichen die arithmetischen Operationen der Addition und Subtraktion. Zulässige statistische Maße sind z. B. der Mittelwert und die Varianz.
Die Ratio- oder Verhältnisskala stellt das höchste Zahlenniveau dar. Sie unterscheidet sich von der Intervallskala, dass zusätzlich noch ein natürlicher Nullpunkt existiert, der sich für das zu messende Merkmal in dem Sinne von „nicht vorhanden“ interpretieren lässt. Bei Werten aus dieser Skala besitzt nicht nur die Differenz, sondern mit der Festlegung des Nullpunktes auch das Verhältnis bzw. der Quotient der Daten Informationsgehalt. Ratioskalierte Daten erlauben alle arithmetischen Operationen.
Nominal- und Ordinalskala bezeichnet man auch als kategoriale und Intervall- und Ratioskala als metrische Skalen. Je höher das Skalenniveau, umso höher ist auch der Informationsgehalt der betreffenden Daten und desto mehr Rechenoperationen und statistische Maße lassen sich auf die Daten anwenden. Es ist immer möglich Daten von einem höheren Niveau auf ein niedrigeres zu transformieren nicht aber umgekehrt. Dies kann nützlich sein um Ergebnisse einfacher und klarer zu präsentieren oder die Analyse zu vereinfachen. So werden z. B. häufig Einkommens- oder Preisklassen gebildet. Mit der Transformation der Daten auf ein niedrigeres Zahlenniveau geht natürlich stets ein Informationsverlust der Daten einher.
Losgelöst von der Frage nach der Qualität der Daten müssen solche zuerst einmal erhoben werden. Verfahren, die zur Untersuchung von Persönlichkeitsmerkmalen

dienen und damit jene Eigenschaften einer Quantifizierung der individuellen Ausprägungen zugänglich machen werden als Tests bezeichnet.[8]
Tests als Messinstrument individueller Persönlichkeitsmerkmale müssen, da die Qualität der Daten auch vom richtigen Einsatz des Messinstruments abhängen, gewissen Mindeststandards und Forderungen genügen, die im Folgenden noch kurz erläutert werden sollen.
Ein Test soll als Hauptgütekriterien 3 Forderungen erfüllen:

1. er soll objektiv,
2. er soll zuverlässig
3. er soll gültig sein.

Ad 1: Unter Objektivität versteht man den Grad, in dem die Ergebnisse unabhängig vom Untersucher sind. Ein Test wäre demnach vollkommen objektiv, wenn verschiedene Untersucher bei demselben Probanden zu gleichen Ergebnissen gelangten. Man spricht deshalb auch von „interpersoneller Übereinstimmung“ der Untersucher. Als Maß für die interpersonelle Unabhängigkeit könnte der durchschnittliche Korrelationskoeffizient zwischen den durch verschiedene Untersucher an einer Stichprobe von Probanden erhoben Testbefunden gelten.
Ad 2: Zuverlässigkeit eines Testes ist ein Maß für die Genauigkeit des Testes. Es gibt an, wie exakt den Probanden eine Maßzahl zugeordnet wird, gleichgültig ob der Test auch dieses Merkmal auch zu messen beansprucht (welche eine Frage der Gültigkeit ist). Ein Test wäre demzufolge vollkommen zuverlässig, wenn die mit seiner Hilfe ermittelten Werte, den Probanden genau, das heißt fehlerfrei, beschreiben bzw. auf der Testskala lokalisieren. Der Grad der Zuverlässigkeit wird durch einen Reliabilitätskoeffizienten bestimmt, der angibt, in welchem Maße unter gleichen Bedingungen gewonnene Messwerte über ein und denselben Probanden übereinstimmen, in welchem Maße also das Testergebnis reproduzierbar ist. Dabei bestehen mehrere Möglichkeiten der methodischen Zugänge zur Bestimmung der Zuverlässigkeit eines Testes:

1. Die Paralleltest-Reliabilität: Sie wird in der Weise bestimmt, dass 2 einander streng vergleichbare Teste (Parallelteste) von einer Stichprobe von Probanden absolviert und deren Ergebnisse miteinander korreliert werden.
2. Die Retestreliabilität: Diese wird mit der Testwiederholungsmethode bestimmt. Man gibt ein und denselben Test einer Stichprobe von Probanden 2x vor und ermittelt die Korrelation der beiden Ergebnisreihen.
3. Die innere Konsistenz eines Testes:
    a) Nach der Methode der Testhalbierung. Die Halbierungsreliabilität oder –konsistenz errechnet man nach folgendem Prinzip: Ein Test wird einer Gruppe von Probanden vorgelegt, und zwar nur ein einziges Mal. Dann wird der Test in 2 gleichwertige Hälften geteilt (gesplittet) und das Testergebnis eines Einzelnen für jede Testhälfte gesondert ermittelt. Schließlich werden die Testergebnisse der beiden Hälften korreliert und der (für den halbierten Test geltende) Reliabilitätskoeffizient so aufgewertet, dass er für den ganzen Test Geltung beanspruchen kann (Halbierungs- oder Split-Half-Konsistenzkoeffizient).

[8] Vgl. [7].

b) Nach der Methode der Konsistenzanalyse: Hier handelt es sich darum, die Elemente eines Tests als multiple halbierte Testteile aufzufassen, und die Zuverlässigkeit über bestimmte Kennwerte dieser Testelemente (Aufgabenschwierigkeit und Trennschärfestatistiken) auf indirektem Wege zu ermitteln, und nicht wie oben durch Korrelation von Testwertepaaren.

Alle hier beschriebenen Arten der Zuverlässigkeitsermittlung würden in aller Regel zu unterschiedlichen Werten für die Zuverlässigkeit führen, wobei freilich bestimmte Gesetzmäßigkeiten zwischen den verschieden wirkenden Koeffizienten bestehen.

	Frage 1	Frage 2	Frage 3	Frage 4	Frage 5	Frage 6	Frage 7	Frage 8	Frage 9	Frage 10	usw.
Proband 1											
Proband 2											
Proband 3	1. Testhälftenergebnis					2. Testhälftenergebnis					
Proband 4	Korrelation über alle Probanden der Testhälftenergebnisse										
Proband 5											
Proband 6											
Proband 7											
Proband 8											
Proband 9											
Proband 10											
usw.											

Abbildung 2: Innere Konsistenz nach der Methode der Testhalbierung, Veranschaulichung der Aggregation.

Ad 3: Es soll hier der Kürze wegen und da nur für diese Art der Validität im allgemeinen ein Koeffizient berechenbar ist auf die **kriterienbezogene Gültigkeit** eingegangen werden. Dieser Koeffizient berechnet sich, indem die Testergebnisse der Stichprobe von Probanden mit einem sogenannten Außenkriterium korreliert werden. Dieses Außenkriterium spiegelt dabei in direkter oder indirekter Weise das Persönlichkeitsmerkmal wider, das es zu erfassen gilt und wird unabhängig vom Test erhoben. Dabei wird das Kriterium als ausreichend gültig angenommen und zwar in dem Maße, in dem es zuverlässig also reproduzierbar ist. Die so berechnete Validität hängt im Wesentlichen von 3 Faktoren ab:

- von dem Grad dessen, was an Gemeinsamkeit durch den Test und das Kriterium erfasst und oft als „Zulänglichkeit“ des Testes bezeichnet wird,
- von der Zuverlässigkeit des Testes und
- von der Zuverlässigkeit des Kriteriums.

3.3.1 Güte- und Unterscheidungskriterien der Testaufgaben:

Bevor wir uns der Ermittlung der einzelnen Kennwerte zuwenden, sei darauf hingewiesen, dass zur Berechnung von einzelnen Kennwerten die erreichte Punkteanzahl der Probanden im Test benötigt wird, der sogenannte Rohwert. Die Rohwertermittlung geschieht am einfachsten und häufigsten, dass jedem Probanden für jede richtige Antwort ein Punkt zuerkannt wird. Der Rohwert **X** eines Probanden soll also wie folgt definiert sein:

$$X = R$$

R = Alle richtigen Antworten eines Probanden.

Gleichung 1

Es leuchtet ein, dass wenn gewisse Gütekriterien für den Test verlangt werden, diese im gewissen Umfang auch für die Aufgaben als Elemente des Testes im „Kleinen“ zu

gelten haben, da sie ja zur Entstehung des Testwertes und dessen Güte beitragen. Diese Kriterien für die Aufgaben sind

1. Schwierigkeit,
2. Zuverlässigkeit,
3. Gültigkeit,
4. Trennschärfe,
5. Objektivität,
6. Homogenität oder Heterogenität.

3.3.1.1. Schwierigkeit der Aufgabe:

Sollte der Test insgesamt, notwendigerweise eine etwa mittlere Schwierigkeit auf die bezogene Untersuchungsgruppe haben, so können die unterschiedlichen Aufgaben durchaus sowohl sehr leicht als auch sehr schwer von den Probanden zu lösen sein. Teste oder Prüfungen wie sie auf Universitäten abgehalten werden, haben in der Regel eher den Charakter eines **Niveautestes**. Der Schwierigkeitsindex einer Aufgabe ist nun definiert durch das Verhältnis von richtig beantworteten zu der Anzahl der Probanden, die diese Aufgabe bearbeitet haben. Der Einfachheit halber soll für die weiteren Berechnungen angenommen werden, dass die Zahl der bearbeiteten Aufgaben mit der Zahl der Probanden übereinstimmt, dass also kein Proband aus Zeitmangel eine Aufgabe nicht bearbeiten konnte und somit reine Niveautestverhältnisse vorliegen[9].Auch der Zufallseinfluss wird vernachlässigt, so dass sich der höchstmögliche Punktwert (Rohwert) des Testes der einzelnen Probanden sich nur um die Anzahl der Fehllösungen und Auslassungen vermindert. Damit ergibt sich:

$$p = \frac{N_R}{N}$$

N_R = Anzahl der Probanden, die die Aufgabe richtig beantwortet haben,
N = Gesamtzahl der Probanden.

Gleichung 2

	Aufgabe 1	Aufgabe 2	Aufgabe 3	...m	
Proband 1	0	1	1		$X1 = \sum_j^m a1j$
Proband 2	1	0	1		$X2 = \sum_j^m a2j$
Proband 3	1	0	1		Rohwert
n...					
	$p1 = \sum_i^n \frac{ai1}{n}$	Aufgabenschwierigkeit			

Abbildung 3: Veranschaulichung von Rohwert, Aufgabenschwierigkeit in der Datenmatrix und der Trennschärfenberechnung.

Besitzen die Aufgaben unterschiedlichen Schwierigkeitsindex, so ist bei der Gestaltung und Anordnung der Testaufgaben darauf zu achten, dass die Aufgaben beginnend bei der leichtesten nach steigender Schwierigkeit (kleiner werdendem Index) geordnet werden. Der größere Teil der Aufgaben sollte eine Schwierigkeit von 0.5 für die zu untersuchende Grundgesamtheit aufweisen, aber auch, wie bei Niveautesten gefordert wird, wenigstens über den ganzen Bereich des Persönlichkeitsmerkmals streuen (nämlich von p = 0.2 bis 0.8),da ansonsten die

[9] Zum Begriff des „reinen“ Niveautestes und zur Korrektur der Schwierigkeitsberechnung bei vorliegen von Mischformen, Niveau- versus Schnelligkeitstest, vgl. [7].

Differenzierungsfähigkeit in den Extrembereichen des Merkmals leidet. Im Test müssen soviel Aufgaben mit geringer Schwierigkeit enthalten sein, dass kein Proband punktlos ausgehen (X > 0) wird. Aufgaben, die einen Schwierigkeitsindex außerhalb des Bereiches 0.95 > p > .05 haben, sollten nach Möglichkeit nicht im Test verweilen. Eine gleichmäßige Differenzierungsfähigkeit erlangt man, wenn die Verteilung der Häufigkeiten der Aufgaben über die Schwierigkeitsindizes ungefähr normal verteilt ist,[10]dass heißt weniger Aufgaben extreme Schwierigkeitsindizes aufweisen und mehrere Aufgaben einen Index um 0.5.

3.3.1.2. Zuverlässigkeit der Aufgabe:

Eine Aufgabe gilt als zuverlässig wenn ihre Beantwortung bei Wiederholung in einem angemessen Zeitintervall in derselben Weise erfolgt wie bei der erstmaligen Vorgabe, vorausgesetzt, dass sich das Persönlichkeitsmerkmal selbst während dieser Zeit in keiner Weise verändert hat. Je mehr zuverlässige Aufgaben ein Test enthält umso zuverlässiger wird auch der Test als Messinstrument, die Zuverlässigkeit des Testes kann also durch Hinzufügen von zuverlässigen Aufgaben (Testverlängerung) verbessert werden.

3.3.1.3. Gültigkeit der Aufgabe:

Eine Aufgabe gilt dann als gültig, wenn sie von dem Probanden mit starker Merkmalsausprägung häufiger im Sinne der Erwartung beantwortet wird als von dem Probanden mit der geringeren Ausprägung. Da die Gültigkeit des universitären Testes zur Wissensabprüfung als gegeben angesehen werden kann, reduziert sich die Validierung der Aufgabe am Testrohwert des Probanden und damit auf die Ermittlung der Trennschärfe.[11]

3.3.1.4. Trennschärfe der Aufgabe:

Der Trennschärfenkoeffizient einer Aufgabe ist gleich dem Korrelationskoeffizienten zwischen der Aufgabenantwort (der richtigen oder falschen) und dem Rohwert des Probanden berechnet über alle Probanden. Ein hoher Trennschärfenkoeffizient besagt also, dass die entsprechende Aufgabe „gute“ von „schlechten“ Probanden deutlich unterscheidet, indem gute Probanden die Aufgabe meist richtig und die Schlechten diese meist falsch beantworten oder auslassen. Ein Trennschärfekoeffizient um 0 bedeutet, dass die Aufgabe sowohl von schlechten und von guten gleich häufig beantwortet werden. Solche Aufgaben sind ungeeignet und sollten aus dem Test ausgeschieden werden. Es soll ebenfalls von dem Fall ausgegangen werden, wie auch in Zukunft bei jeder hinkünftigen Berechnung; dass vollständige Aufgabendarbietung vorliegt. Die Berechnung erfolgt nun z.B. für den Trennschärfekoeffizienten für die Aufgabe 1 als Korrelation von den Aufgabenwerten der Aufgabe 1 (dichotome Werte = 0 oder 1 in Spalte 1) mit den entsprechenden Werten der Rohwertspalte. Veranschaulicht wird dieser Vorgang in Abbildung 3. Korreliert werden also für die Berechnung der Trennschärfe für Aufgabe 1 (r_{1t})die gelb hinterlegten Spalten über alle Probanden. Dieser so berechnete Wert entspricht der punktbiserialen Korrelation, die gleichzeitig algebraisch identisch dem Produkt-Moment-Koeffizienten entspricht. Eine alternative Berechnung als biserialer Koeffizient, was dann erforderlich würde, wenn die Aufgabe z.B.: zu 20 oder 30% gelöst werden kann, was erst im Nachhinein zu 0 dichotomisiert wird, der Dichotomie also in Wirklichkeit vielleicht eher sogar eine normalverteilte Variable zugrunde liegt,

[10] Vgl. [7] Seite 138.

[11] Eine mögliche Art der Berechnung eines Validitätskoeffizienten werden wir unter 3.6 kennen lernen, Die so gefundene reduzierte Prädiktorenmatrix...

und deshalb ein biserialer Koeffizient besser geeignet wäre, wird aufgrund der Tatsache, dass der biseriale Koeffizient ± 1.0 über bzw. unterschreiten kann und erhebliche Überschätzungen zeigen kann, wenn p stark von 0.5 abweicht, wird nicht erwogen.[12] Algebraisch berechnet sich die Trennschärfe der Aufgabe i nach folgender Formel:

$$_{pbis}r_{it} = \frac{\overline{X}_R - \overline{X}}{s_x} * \sqrt{\frac{p}{q}}$$

wobei p = N_R/N und q = 1-p ist.
In dieser Formel bedeuten

$\overline{X}$ = arithmetisches Mittel aller Testrohwerte,
$\overline{X}_R$ = arithmetisches Mittel der Testrohwerte von denjenigen Probanden, die die Aufgabe richtig beantwortet haben,
s_x = Standardabweichung der Testrohwerte aller Probanden[13],
N = Anzahl aller Probanden
N_R = Anzahl derjenigen Probanden, die die Aufgabe richtig beantwortet haben.

Gleichung 3

Dieser so berechnete Kennwert sollte zumindest, wenn sonst kein oberer Grenzwert vereinbart wurde, signifikant von 0 verschieden sein. Die Signifikanzüberprüfung erfolgt durch folgenden Test:[14]

$$t = \frac{_{pbis}r_{it}}{\sqrt{\frac{1-_{pbis}r_{it}^{\ 2}}{N-2}}}$$

Gleichung 4

Der so ermittelte t-Wert ist mit N-2 Freiheitsgraden versehen und wird mit dem auf einem bestimmten α-Niveau erwartetem einseitigen $t_{crit.}$ verglichen, da ja Aufgaben mit einer negativen Trennschärfe sowieso nicht im Test verbleiben dürfen. Ist $t > t_{crit.}$ dann ist der Trennschärfekoeffizient signifikant von 0 verschieden. Die Trennschärfe von Testaufgaben wird im wesentlichen durch 2 Faktoren verändert:

1. durch die Scheinkorrelation gegenüber dem Testrohwert als Analysenkriterium
2. durch die Stellung der Aufgabe innerhalb des Testes

Ad 1: Der Trennschärfekoeffizient berechnet sich als Korrelation der dichotomen Aufgabenwerte mit dem jeweiligen Rohwert über die Probanden. Nun erhält man aber auf diese Weise einen zu hohen Wert und zwar deshalb, weil jede einzelne Aufgabe auch ein Teil des Testrohwertes ist, mit dem diese Aufgabe korreliert wird. Eine exakte Berechnung wäre dann möglich, wenn für die Rohwertberechnung jene Aufgabe nicht berücksichtigt wird, für welche man die Trennschärfe bestimmt. Ein solches Vorgehen ist jedoch im Vergleich zu seiner praktischen Bedeutung eine unzumutbare Mehrbelastung. Die Teil-Ganzheits-Scheinkorrelation spielt überhaupt nur dann eine gewisse Rolle, wenn ein Test nur wenig Aufgaben von mittlerer

[12] Vgl. [7]
[13] Es handelt sich hierbei um den nicht erwartungstreuen Maximum Likelihood Schätzer der Standardabweichung $\sqrt{\frac{\sum\left(X-\ddot{X}\right)^2}{N}}$.
[14] Vgl. [8]

Schwierigkeit hat, die untereinander nicht oder nur minimal korrelieren. In solchen Fällen wird eine Korrektur des Trennschärfekoeffizienten empfohlen:

$$r_{i(t-i)} = \frac{r_{it} * s_x - \sqrt{p*q}}{\sqrt{s_x{}^2 + p*q - 2*r_{it}*s_x*\sqrt{p*q}}}$$

Gleichung 5

Bei heterogenen Tests sollte der Trennschärfekoeffizient ein Maximum erreichen, bei Homogenen sollten jene Aufgaben im Test belassen werden, die einen hohen Selektionskennwert besitzen.[15]Dieser Selektionskennwert S ist nun nach Lienert so beschaffen, dass Aufgaben, die eine extreme Schwierigkeit und damit einhergehend auch eine geringere Trennschärfe besitzen, nicht ungebührlich benachteiligt werden. Denn wie aus Gleichung 3 ersichtlich und durch Abbildung 4 graphisch veranschaulicht wird, besteht ein Zusammenhang zwischen Trennschärfe und Aufgabenschwierigkeit.

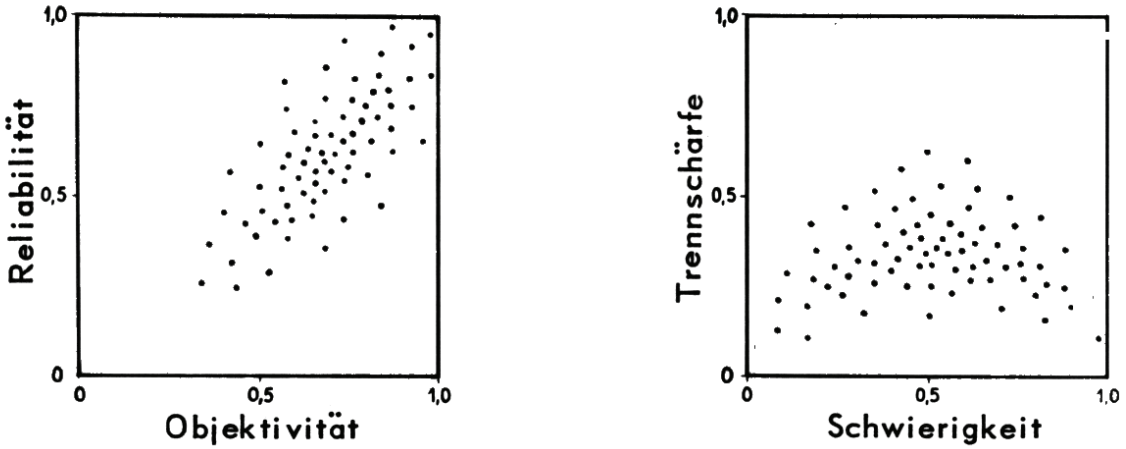


Abbildung 4 Zusammenhang Trennschärfe und Schwierigkeit.

Dieser Selektionswert berechnet sich nun durch die Relativierung des Trennschärfenkoeffizienten an der Standardabweichung der Aufgabe, die wie der Trennschärfekoeffizient ebenso nach den Extremen hin abnimmt.

$$S_{it} = \frac{r_{it}}{2*\sqrt{p*q}}$$

Gleichung 6

Dieser S-Kennwert erlaubt eine Selektion nach dem Prinzip, dass die Aufgaben mit dem niedrigen S-Wert ausgeschieden werden, ohne befürchten zu müssen, dass auf diese Weise zu viele Aufgaben mit extremer Schwierigkeit verloren gehen und damit die Differenzierungsfähigkeit des Testes eingebüßt wird.
Schwierigkeit und Homogenität des Testes stellen ebenfalls keine unabhängigen Größen dar. Stellt man sich den Fall vor, dass alle Aufgaben perfekt miteinander korrelieren (r_{ij} =1, i ≠ j; i,j = 1..m) und damit eine höchstmögliche Homogenität des Testes vorliegt, so werden entweder alle Aufgaben vom Proband beantwortet oder nicht. Eine vollkommene Korrelation kann nur bei gleicher Aufgabenschwierigkeit bestehen, wenn eine Aufgabe häufiger als eine andere beantwortet würde, ist eine Korrelation von r_{ij} = 1 nicht möglich. Je größer die Schwierigkeitsunterschiede sind, umso geringer werden die Korrelationskoeffizienten. Hier kann gesehen werden,

[15] Vgl. [7]

dass nicht nur die inhaltliche Homogenität, sondern auch die Schwierigkeitsunterschiedlichkeit für die Höhe der Korrelation relevant ist. Wenn also inhaltlich vollkommen homogen erscheinende Aufgaben nicht gemäß der Erwartung hoch, sondern nur relativ niedrig miteinander korrelieren, so kann bereits daraus geschlossen werden, dass die Aufgaben dieses homogenen Tests von sehr unterschiedlicher Schwierigkeit sein müssen. Es bleibt also festzuhalten, dass die beste Differenzierungsfähigkeit und damit Zuverlässigkeitsbedingungen eines Testes bestehen, wenn der Test entweder aus unkorrelierten (heterogenen) Aufgaben gleicher Schwierigkeit oder mäßig positiv korrelierten (homogenen) Aufgaben unterschiedlicher Schwierigkeit besteht.

3.3.1.5. Objektivität der Aufgabe:

Eine Aufgabe ist dann objektiv, wenn sie von verschiedenen Testbeurteilern übereinstimmend als richtig oder falsch, als kennzeichnend für das Vorhandensein oder Fehlen des untersuchten Persönlichkeitsmerkmals, bzw. dessen Ausprägungsgrad gewertet wird. Ein Beispiel für eine absolut objektive Aufgabe stellen Rechenaufgaben dar, deren Richtigkeit bzw. Unrichtigkeit durch das Ergebnis quasi definitorisch festgelegt werden und nicht durch interpretatorische Mängel der Urteiler die eigentliche und wahre Objektivität der Aufgabe noch verzerrt werden kann. Solche Verzerrungen der Objektivität liegen natürlich nur für Aufgabetypen vor, die einer Beurteilung der Richtigkeit der Aufgaben bedürfen. Die individuelle Urteilerübereinstimmung oder interpersonelle Übereinstimmung wird als Korrelation (Zuverlässigkeitsmaß) über die Aufgabenwerte (0...Aufgabe nicht, 1...Aufgabe gelöst) ermittelt und bildet ein Maß für die Objektivität der Aufgabenbewertungen der verschiedenen Beurteiler im engeren Sinne. Sie gibt an, wie gut man von einer Beurteilung auf eine andere schließen kann. Die Urteilerübereinstimmung von 2 Urteilern errechnet sich als Mittelung der Korrelation des Proband i über alle i = 1..m Probanden, wobei die Korrelationen zuvor noch Fischer Z transformiert werden sollen.[16]

$$Z = \frac{1}{2} * \ln\left(\frac{1+r}{1-r}\right)$$

Gleichung 7

Beurteilen mehr als 2 nämlich n die verschiedenen Aufgaben, so berechnet sich die individuelle Urteilerübereinstimmung als Mittelung über alle

$$\frac{n*(n-1)}{2}$$

Gleichung 8

Kombinationen der Fischer Z transformierten Urteilübereinstimmungen der verschiedenen Urteiler.
Die Rücktransformation des gemittelten Z Wertes erfolgt nach Gleichung 9 zum entscheidenden Korrelationsmaß für die individuelle Urteilerübereinstimmung.

$$r = \frac{e^{2*Z} - 1}{e^{2*Z} + 1}$$

Gleichung 9

[16] Wenn Korrelationen die Werte -1 oder 1 annehmen , werden sie natürlich nicht transformiert, sondern die Urteilerübereinstimmung berechnet sich als Mittelung der nicht transformierten Korrelationen.

Natürlich werden bei dieser Berechnung für die Objektivität dieselben durch die Urteiler bewerteten Testleistungen der Probanden herangezogen, um die Objektivität hinsichtlich der Testdurchführung zu gewährleisten. Im Folgenden soll nun die Berechnung der Übereinstimmung von 2 Urteilern exemplarisch durchgeführt werden und Gedankenansätze dazu erläutert werden.
In unserem exemplarischen Beispiel, vergleiche Abbildung 6, wurden von 2 Urteilern Aufgaben bewertet, die unterschiedliche Schwierigkeitsindizes aufweisen. Die Blöcke mit verschiedenem Schwierigkeitsindex, hier mit 0.9, 0.5 und 0.1 angenommen, besitzen jeweils 10 Aufgaben. Es ist einleuchtend, dass der Schwierigkeitsindex sowohl durch den Beurteiler beeinflusst wird als auch von Proband zu Proband unterschiedliche Werte annimmt. So wird Proband i mehr oder weniger Aufgaben richtig lösen als Proband j (Aufgaben aus derselben durchschnittlichen Schwierigkeitsklasse) oder Proband i wird von Urteiler I besser beurteilt (hat mehr Aufgaben richtig) als von Beurteiler II. Die „wahre“ Aufgabenschwierigkeit, die im Beispiel mit 0.9, 0.5, und 0.1 angenommen wurden, soll über alle Probanden und Urteiler als Mittelwert geschätzt werden.

$$p_k = \frac{1}{m * n} * \sum_{i}^{m} \sum_{j}^{n} a_{ij}$$

p_k = Schwierigkeitsindex der Aufgabe k
n = Anzahl der Probanden
m = Anzahl der Beurteiler
a_{ij} = nimmt 1 oder 0 als Wert an (richtig / falsch).

Gleichung 10

		Aufgabe k
Urteiler 1	Proband 1	1
	Proband 2	0
	Proband 3	1
	Proband 4	1
	⋮	
	n	
Urteiler 2	Proband 1	1
	Proband 2	0
	Proband 3	0
	Proband 4	1
	⋮	
	n	
⋮		
m		

Abbildung 5 Berechnung der „wahren“ Aufgabenschwierigkeit der Aufgabe k bei mehreren Urteilern

Die Urteilerübereinstimmung soll jedoch nicht als einfache Korrelation ohne Differenzierung der Aufgaben über alle Aufgaben berechnet werden. Die Aufgaben unterscheiden sich ja, wie auch in unserem Beispiel angenommen in ihrer Schwierigkeit. Um diesem Umstand Rechnung zu tragen, müssen die Aufgabenwerte (0 oder 1) bezüglich der jeweiligen Aufgabenschwierigkeit residualisiert werden. Würde nur die Korrelation über die 0/1 Werte ermittelt, wie im obersten Block angedeutet, würde eine „Fehlkombination“ der 0/1 Werte egal in welcher Schwierigkeitsklasse mit einem jeweils für die 2 Urteiler über alle Klassen ermittelten Schwierigkeitsgrad von 0.5 und damit mit einer über die Maßen großen, geometrisch gemittelten Varianz 0.25 (beide Urteiler besitzen über alle Klassen im obersten Block einen Schwierigkeitsindex von 0.5) abgewertet werden. Es leuchtet aber ein ,dass

eine Fehlkombination bei einem Schwierigkeitsindex von 0.5 (die durchschnittliche Chance des Probanden sich zu irren beträgt 50% und auch die Aufgabe jetzt bei einem knappen Ergebnis die Aufgabe gerade noch als richtig oder schon falsch zu bewerten gewinnt an Schwierigkeit) bei gleicher Sorgfaltspflicht der Urteiler häufiger anzutreffen sein wird als bei einem Schwierigkeitsindex von 0.9. Eine Fehlkombination muss daher in den Schwierigkeitsklassen 0.9, 0.1 schwerer „wiegen“ als in der 0.5er Klasse.
Diesen Sachverhalt spiegelt der oberste Block wider: 2 Fehlkombinationen in Klasse 0.9 ergeben sogar eine negative Urteilerübereinstimmung (für die Klassen getrennt berechnet), während die gleiche Anzahl an Fehlbewertungen in der 0.5 Klasse noch eine Urteilerübereinstimmung von 0.6 ausweist. Ebenso darf, wie am Beispiel für die Klasse 0.9 dargestellt, bei extremen Aufgabenschwierigkeiten durch die extremere Bewertung der richtigen Zuordnungen bzw. Fehlzuordnungen durch die kleiner geometrisch gemittelte Varianz im Fall des Auseinanderklaffens der über die jeweiligen Urteiler berechneten Aufgabenschwierigkeiten p1 bzw. p2, eine bessere Übereinkunft erwartet werden, wenn die über die jeweiligen Urteiler berechneten Aufgabenschwierigkeiten p1 bzw. p2 für Proband 1 übereinstimmend gleich beurteilt wurden und somit ein reines Zuordnungsproblem (in unserem Beispiel die gelb gekennzeichneten Felder im oberen Block) besteht, als im 2. Fall (Proband 2) der systematisch falschen Beurteilung (gelbe Felder im Block 2), da auch die Zahl der zufälligen und damit erwarteten Übereinstimmungen mit dem Auseinanderklaffen der Aufgabenschwierigkeitseinschätzungen sinkt, sofern beide Schwierigkeitsschätzung von 0,5 verschieden sind. Zur Verdeutlichung des eben Gesagten soll Abbildung 7 und die dazu gehörenden Erläuterungen am Ende dieses Kapitels dienen.
Der Schätzwert für die Aufgabenschwierigkeit ergibt sich in beiden Fällen zu.

$$\mathrm{p}_{\mathrm{mittel}} = \frac{\mathrm{p}1 + p2}{2} = \frac{0.9 + 0.9}{2} = \frac{1 + 0.8}{2} = 0.9$$

Gleichung 11

Die Aufgabenwerte der Probanden werden bezüglich dieser ermittelten Aufgabenschwierigkeit residualisiert (jeweils untere Zeile in den Blöcken) und über diese Werte die Korrelationen berechnet. Die niedrigere Urteilerübereinstimmung für den 2. Fall errechnet sich nun, obwohl sich in beiden Fällen die gleiche Kovarianz ergibt, aus der Tatsache, dass in die Korrelationsberechnung nicht die arithmetisch gemittelte Varianz, berechnet über die Residuen der jeweiligen Urteiler eingeht,

$$\sigma^2{}_{ar.-mittel} = \frac{\sigma 1^2 + \sigma 2^2}{2}$$

Gleichung 12

wobei $\sigma 1^2 = (0.1^2+0.1^2+\cdots+0.1^2+0.1^2)/10$ und $\sigma 2^2 = (0.1^2+0.1^2+\cdots+(-0.9)^2+(-0.9)^2/10$ ist, die der Varianz von

$$\sigma_{(pmittel)}{}^2 = p_{mittel} * (1 - p_{mittel}) = 0.09$$

Gleichung 13

entspricht, sondern die geometrisch gemittelte Varianz.

$$\sigma^2_{geo.-mittel} = \sqrt{\sigma 1^2 * \sigma 2^2} = 0.041$$

Gleichung 14

Somit werden die 2 Fehlzuordnungen im 1. Fall mit 0.09 im 2. mit nur 0.041 abgemindert, denn es gilt stets, dass die geometrisch gemittelte Varianz ≤ der

arithmetisch gemittelten Varianz entspricht. In Abbildung 6 (Fortsetzung 1) wird ein Proband (2. Block) mit einem anderen Probanden (3. Block) bei einer Aufgabenschwierigkeit von 0.5 „verglichen“. Der Erstere wird systematisch von beiden Urteilern unterschiedlich bewertet (p1 = 1, p2 = 0) woraus eine Korrelation von -1 folgt. Im 3. Block liegen sowohl die separat geschätzten Aufgabenschwierigkeiten (p1 = 0.7, p2 = 0.3) besser beieinander und die Übereinstimmungen zu nicht Übereinstimmungen verhalten sich wie 6:4, was in einer besseren Übereinstimmungskorrelation von 0.2 resultiert. Aus dem oben gesagten wird ersichtlich, warum für die Berechnung der Urteilerübereinstimmung somit die bezüglich der Aufgabenschwierigkeit residualisierten Werte der Probanden herangezogen werden müssen. Mit dieser Prozedur wird gleichermaßen die Unterschiedlichkeit der Aufgaben aus der Berechnung herausgefiltert und ermöglicht nunmehr eine Korrelationsberechnung über die verschiedenen Aufgabenklassen. So wird die Übereinstimmung für Proband 1 und Proband 2 ohne eine solche Berichtigung auf 0.6 bzw. 0.27 berechnet, eine residualisierte Übereinstimmung ergibt sich zu 0.3 bzw. -0.64, was zum einen die unterschiedliche „Schwere“ der Zuordnungsfehler in den unterschiedlichen Klassen berücksichtigt als auch durch die gebräuchliche Verwendung der geometrisch gemittelten Varianz in der Korrelationsrechnung die Unterschiedlichkeit der separat von Urteiler I und II geschätzten Aufgabenschwierigkeiten beachtet.

Aufgabenschwierigkeit
0,9
Aufgabe1 Aufgabe2 Aufgabe3 Aufgabe4 Aufgabe5 Aufgabe6 Aufgabe7 Aufgabe8 Aufgabe9 Aufgabe10
Urteiler I 1 1 1 1 1 1 1 1 0
Urteiler II 1 1 1 1 1 1 1 0 1
Urteiler I 0,1 0,1 0,1 0,1 0,1 0,1 0,1 0,1 0,1 -0,9
Urteiler II 0,1 0,1 0,1 0,1 0,1 0,1 0,1 0,1 -0,9 0,1
Korrelation-> -0,111111111 0,09 <- geometrisch gemittelte Varianz
Kovarianz-> -0,01
Proband 1
Aufgabenschwierigkeit
0,9
Aufgabe1 Aufgabe2 Aufgabe3 Aufgabe4 Aufgabe5 Aufgabe6 Aufgabe7 Aufgabe8 Aufgabe9 Aufgabe10
Urteiler I 1 1 1 1 1 1 1 1 1
Urteiler II 1 1 1 1 1 1 1 1 0 0
Urteiler I 0,1 0,1 0,1 0,1 0,1 0,1 0,1 0,1 0,1 0,1
Urteiler II 0,1 0,1 0,1 0,1 0,1 0,1 0,1 0,1 -0,9 -0,9
Korrelation-> -0,242535625 0,041231056 <- geometrisch gemittelte Varianz
Kovarianz-> -0,01
Proband 2
m... usw.

Abbildung 6

Abbildung 6 (Fortsetzung 1)

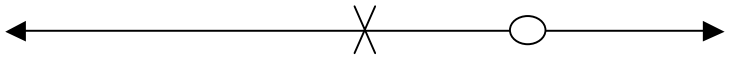
Aufgabenschwierigkeit 0,1										
Aufgabe21	Aufgabe22	Aufgabe23	Aufgabe24	Aufgabe25	Aufgabe26	Aufgabe27	Aufgabe28	Aufgabe29	Aufgabe30	
1	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	1	0,6 1).
0,9	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	
-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	0,9	0,302325581 2).

Aufgabenschwierigkeit 0,1										
Aufgabe21	Aufgabe22	Aufgabe23	Aufgabe24	Aufgabe25	Aufgabe26	Aufgabe27	Aufgabe28	Aufgabe29	Aufgabe30	
1	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	1	0,26984127 1).
0,9	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	
-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	0,9	-0,639064442 2).

Abbildung 6 (Fortsetzung 2) [17]

[17] Der mit 1). gekennzeichnete Wert enthält die Korrelation über alle Werte ohne Rücksichtnahme auf die Aufgabenschwierigkeit.
Der mit 2). Gekennzeichnete Wert enthält die „korrekte“ Berechnung der Urteilerübereinstimmung als Korrelation über alle Aufgaben mit Berücksichtigung der Aufgabenschwierigkeit.

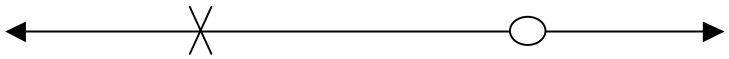
		Kombinationsmöglichkeiten:																	
		[0,0,0,0]	[1,0,0,0]				[1,1,0,0]						[1,1,1,0]				[1,1,1,1]	
fixer Tupel:	1	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	1	
	1	0	0	1	0	0	0	0	1	1	1	0	1	1	0	1	1	
	1	0	0	0	1	0	0	1	0	1	0	1	1	0	1	1	1	
	0	0	0	0	0	1	1	0	0	0	1	1	0	1	1	1	1	
richtige 1-Zuordnung		0	1	1	1	0	1	2	2	2	1	1	3	2	2	2	3	
richtige 0-Zuordnung		1	1	1	1	0	0	1	1	1	0	0	1	0	0	0	0	
total R		1	2	2	2	0	1	3	3	3	1	1	4	2	2	2	3	
Erwartungswert 1:		0	0,75				1,5						2,25				3	
Erwartungswert 0:		1	0,75				0,5						0,25				0	
Erwartungswert R:		1	1,5				2						2,5				3	
Fehlzuordnungen:		3	2,5				2						1,5				1	
Schwierigkeit:		0	0,25				p2=0,5						p1=0,75				1	



Fehlzuordnungen: + Fehlzuordnungen: -

$$\frac{9}{6} * \left(1 - \overline{p}\right)^2 + \overline{p}^2 * \frac{3}{6} - (1 - \overline{p}) * \overline{p} * 2 = \min \qquad \overline{p} = \frac{p1 + p2}{2} = 0,625$$

		Kombinationsmöglichkeiten:																	
		[0,0,0,0]	[1,0,0,0]				[1,1,0,0]						[1,1,1,0]				[1,1,1,1]	
fixer Tupel:	1	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	1	
	1	0	0	1	0	0	0	0	1	1	1	0	1	1	0	1	1	
	1	0	0	0	1	0	0	1	0	1	0	1	1	0	1	1	1	
	0	0	0	0	0	1	1	0	0	0	1	1	0	1	1	1	1	
richtige 1-Zuordnung		0	1	1	1	0	1	2	2	2	1	1	3	2	2	2	3	
richtige 0-Zuordnung		1	1	1	1	0	0	1	1	1	0	0	1	0	0	0	0	
total R		1	2	2	2	0	1	3	3	3	1	1	4	2	2	2	3	
Erwartungswert 1:		0	0,75				1,5						2,25				3	
Erwartungswert 0:		1	0,75				0,5						0,25				0	
Erwartungswert R:		1	1,5				2						2,5				3	
Fehlzuordnungen:		3	2,5				2						1,5				1	
Schwierigkeit:		0	p2=0,25				0,5						p1=0,75				1	



Fehlzuordnungen: + Fehlzuordnungen: -

$$\frac{3}{4} * \left(1 - \overline{p}\right)^2 + \overline{p}^2 * \frac{3}{4} - (1 - \overline{p}) * \overline{p} * \frac{10}{4} = \min \qquad \overline{p} = \frac{p1 + p2}{2} = 0,5$$

Abbildung 7

Die nun folgenden Überlegungen sollen beispielhaft an einem 4-Tupel, mit 0en und 1en belegt, wie in Abbildung 7 geschehen, angedeutet werden. Es soll der fix gewählte Tupel [1,1,1,0] mit allen möglichen Kombinationen des 4-Tupels gepaart werden und die richtigen 1- als auch 0-Zuordnungen jeweils festgehalten werden. Hieraus lassen sich pro Kombination mit dem fixierten Tupel die erwarteten richtigen (1 als auch 0 Zuordnungen) Zuordnungen und erwarteten Fehlzuordnungen berechnen. Wie in Abbildung 7 oben geschehen, wird am Beispiel der Kombination des [1,1,0,0]- Tupel (mit X in der Abb. gekennzeichnet) mit dem fix gewählten Tupel (mit 0 in der Abb. gekennzeichnet) die verschiedenen Rechenoperationen erläutert. Es ergeben sich für diese Kombination 9 richtige 1-Zuordnungen und 3 richtige 0-

Zuordnungen. Werden diese Zahlen noch durch die 6 Variationsmöglichkeiten dividiert (der [1,1,0,0] Tupel besitzt 6 Variationsmöglichkeiten) folgt der **Erwartungswert** der richtigen 1-Zuordnungen zu 1,5, der richtigen 0-Zuordnungen zu 0,5 und der Fehlzuordnungen zu 2.
Werden fixer und variabler Tupel vertauscht erhält man auf analoge Weise (diesmal mit 4 Variationsmöglichkeiten des nun variablen [1,1,1,0] Tupels) dieselben Erwartungswerte für richtige 1-Zuordnung, 0-Zuordnung und Fehlzuordnung.
Dass heißt wir erwarten uns bei diesen zwei 4-Tupeln bei zufälliger Anordnung im Schnitt 1,5 richtige 1-Zuordnungen, 0,5 richtige 0-Zuordnungen und insgesamt 2 Fehlzuordnungen.
Allgemein erhalten wir obige **Erwartungswerte** der gegenseitigen Zuordnungen bei **Unabhängigkeit** der zwei Tupel, dies wurde im obigen Beispiel so angenommen, da ja der fixe Tupel mit allen Variationsmöglichkeiten des variablen Tupels kombiniert, dass heißt alle Variationen des nicht fixierten [1,1,0,0] Tupels treten unabhängig vom fixen Tupel mit gleicher Wahrscheinlichkeit auf, dass heißt keine Variation ist ausgezeichnet, indem die Wahrscheinlichkeit des Auftretens einer 1 oder 0 des fixen Tupels, mit der Wahrscheinlichkeit des Auftretens einer 1 oder 0 des variablen Tupels und der **Tupel-Länge** multipliziert wird.

	fixer Tupel: [1,1,1,0]	variabler Tupel: [1,1,0,0]	Erwartungswert:
	1	1	p1*p2*N=1,5
	1	0	p1*(1-p2)*N=1,5
	0	1	(1-p1)*p2*N=0,5
	0	0	(1-p1)*(1-p2)*N=0,5
Schwierigkeit:	p1=0,75	p2=0,5	
Tupellänge:	N=4		

Abbildung 8
Die Kovarianz für unabhängige Tupel lautet allgemein[18],

$$(1-p1)*(1-p2)*p1*p2+p1*p2*(1-p1)*(1-p2)$$
$$-(1-p1)*p2*p1*(1-p2)-(1-p2)*p1*p2*(1-p1)=0$$

wobei p1 und p2 die Wahrscheinlichkeit des Auftretens einer 1 in den unterschiedlichen Tupeln und dementsprechend (1-p1), (1-p2) die Wahrscheinlichkeit des Auftretens einer 0 angibt.

Gleichung 15

Wie leicht zu sehen ist, ist die Kovarianz für unabhängige Tupel für beliebige Werte für p1 und p2 gleich 0. Setzten wir für die Wahrscheinlichkeiten nun die Erwartungswerte aus unserem Beispiel ein, ersichtlich in **Abbildung** 8, die sich mit p1 = 0,75 und p2 = 0,5 ergeben haben, so folgt:

$$\frac{3}{2}*(1-p1)*(1-p2)+\frac{1}{2}*p1*p2-\frac{3}{2}*(1-p1)*p2-\frac{1}{2}*(1-p2)*p1=0$$

Gleichung 16.

Gleichung 16 besitzt aufgrund der größeren Anzahl von Variablen zur Anzahl von Gleichungen einen unendlichen Lösungsraum mit den Lösungen L1 = {p1 = 3/4, p2 = p2} und L2 = {p1 = p1, p2 = 1/2}. Setzten wir noch $p1=p2=\overline{p}$ so folgt der linke Teil der zu minimierenden Gleichung unterhalb des Diagramms in Abbildung 7 mit den Nullstellen L = {3/4,1/2} für den quadratische Ausdruck. Da die Erwartungswerte in

[18] Allgemein lautet die Kovarianz für zwei stochastische Variablen, x und y: E[(x-E(x))*(y-E(y))].

unserem Beispiel durch die unabhängige Kombination der Tupel mit den Schwierigkeiten p1 = 0,75 und p2 = 0,5 erhalten wurden, müssen dies auch die 2 Nullstellen dieses Ausdrucks sein. Der quadratische Ausdruck liefert an der Stelle des Mittelwerts

$$\bar{p} = \frac{p1 + p2}{2}$$

Gleichung 17

ein Minimum.
Die Kovarianz berechnet über die verschiedenen Aufgaben, die zuvor noch bezüglich der gemeinsam gefundenen Aufgabenschwierigkeit ($\bar{p}$) residualisiert wurden, bilden für nicht systematische, zufällige Zusammenhänge ein Minimum, das umso größer ausfällt je weiter die Werte p1 und p2 auseinanderliegen, dass heißt je unterschiedlicher die 2 Urteiler die Aufgabenschwierigkeit beurteilt haben.

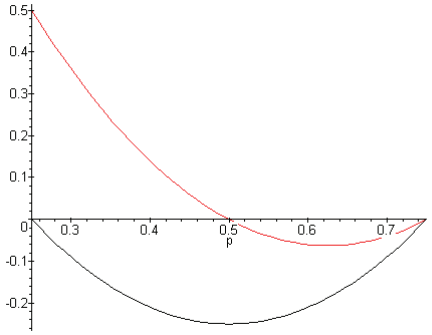


Abbildung 9 Die rote Funktion bedeutet die obere zu minimierende Funktion in Abbildung 7 mit p2 = 0,5 und p1 = 0,75. Sie ist gleichbedeutend mit der Kovarianz der unabhängigen Tupel, die noch mit der Tupellänge multipliziert wurde. Wie man sieht, ist das Minimum der roten Funktion kleiner als der schwarz gezeichneten Funktion, deren p-Werte weiter auseinander liegen (p1 = 0,75, p2 = 0,25).

Bei reiner Zufälligkeit der Zuordnungen fällt also die Bewertung der Auswertungsübereinstimmung bei gleicher Schwierigkeitsbeurteilung der Aufgaben (p1 = p2) der Urteiler „humaner“ (nämlich 0) aus als bei unterschiedlicher Auffassung über die Aufgabenschwierigkeit (negative Objektivität).
Generell wird die Tendenz zu systematischen Zusammenhängen (richtige Zuordnungen über die der reinen Zufälligkeit hinaus) für extremer werdende Aufgabenschwierigkeiten (je mehr p1 und p2 von 0,5 abweicht, z.B. p1 = p2 = 0,8) positiver beurteilt als bei einer Schwierigkeitseinschätzung nahe 0,5. Dies leuchtet unmittelbar ein, da es nahe p1 = p2 = 0,5 (entspricht einer Varianz von 0,25) es zu falschen Zuordnungen mit einer Irrtumswahrscheinlichkeit von 50% kommen kann, dass heißt es wahrscheinlicher ist positive oder negative Kovarianzen zu erhalten, obwohl es eigentlich keinen Zusammenhang gibt. Diese Irrtumswahrscheinlichkeit ist bei extremeren Aufgabenschwierigkeiten geringer, deshalb werden auch systematische Zusammenhänge (Kovarianz≠0) stärker durch die kleinere Varianz von 0,16 in der Korrelationsberechnung bei p1 = p2 = 0,8 bewertet. Ähnliches gilt auch für unterschiedliche Schwierigkeitseinschätzungen. Werden zufällig erwartete Zuordnungen bei verschiedener Schwierigkeitseinschätzung schlechter bewertet, so wird bei Vorliegen einer systematischen Übereinstimmung diese bei verschiedenen p1 und p2 durch die kleinere Varianz besser bewertet, da sie nicht in dieser Höhe erwartet werden. Dies soll noch in Abbildung 10 verdeutlicht werden. Es nehmen ja, wie aus Abbildung 7 ersichtlich wird, mit größer werdender Unterschiedlichkeit der

Schwierigkeitseinschätzungen, die Zahl der Fehlzuordnungen bei zufälliger Anordnung der 2 Tupel im Vergleich zu der Anzahl an richtigen Zuordnungen zu.

1	Urteiler I:	Urteiler II:	residualisierte Werte:	
	1	1	0,2	0,2
	1	1	0,2	0,2
	1	1	0,2	0,2
	1	1	0,2	0,2
	1	1	0,2	0,2
	1	1	0,2	0,2
	1	1	0,2	0,2
	0	1	-0,8	0,2
	0	1	-0,8	0,2
	0	0	-0,8	-0,8
p1:	0,7			
p2:		0,9		
(p1+p2)/2:	0,8			
geo.-gem. Varianz:			0,14832397	
Kovarianz:			0,6	
Korrelation:			0,404519917	
8 Richtige-2Fehler:				

2	Urteiler I:	Urteiler II:	residualisierte Werte:	
	1	1	0,2	0,2
	1	1	0,2	0,2
	1	1	0,2	0,2
	1	1	0,2	0,2
	1	1	0,2	0,2
	1	1	0,2	0,2
	1	1	0,2	0,2
	1	0	0,2	-0,8
	0	1	-0,8	0,2
	0	0	-0,8	-0,8
p1:	0,8			
p2:		0,8		
(p1+p2)/2:	0,8			
geo.-gem. Varianz:			0,16	
Kovarianz:			0,6	
Korrelation:			0,375	
8 Richtige-2Fehler:				

Abbildung 10

Es folgt eine größere Korrelation und damit Objektivität der Aufgabenauswertung der 2 Urteiler für das mit 1 gekennzeichnete Beispiel in Abbildung 10, was zuerst überraschen mag, obwohl im Gegensatz zu Beispiel 2 die Aufgabenschwierigkeit unterschiedlich geschätzt wurden. Im Beispiel 1 würde mit **Erwartungswerten** von 6,3 für richtige 1-Zuordnungen, von 0,3 für richtige 0-Zuordnungen und 3,4 Fehlzuordnungen von den 10 möglichen Anordnungen eine negative Korrelation der bezüglich $\overline{p}$ residualisierten Werte folgen, was auf die unterschiedlich Schwierigkeitseinschätzung zurückzuführen ist. Diese wäre im 2. Beispiel für 6,4 erwartet, richtige 1-Zuordnungen, 0,4 erwartete, richtige 0-Zuordnungen und 3,2 Fehlzuordnungen, was ebenfalls auf eine Unabhängigkeit der Tupel schließen lässt, gleich 0, da die Schwierigkeit der Aufgaben gleich eingeschätzt wurden. Die tatsächliche Übereinstimmung von 8 richtigen und 2 Fehlzuordnungen lassen aber das 1. Beispiel besser erscheinen, da sie mehr von den zufällig erwarteten Anordnungen abweicht und dies bei auseinanderliegenden p1 und p2 nicht erwartet wird.
Resultieren negative Korrelationen oder solche gleich 0 kann von keiner Objektivität der Aufgaben gesprochen werden, da diese eine gleichsinnige Aufgabenauswertung der verschiedenen Urteiler voraussetzt und dies daher einer Korrelation >0 entsprechen muss.

3.3.1.6. Homogenität oder Heterogenität der Aufgabe:

Die Aufgabe eines Testes können mehr oder weniger homogen und nicht selten auch heterogen sein. Homogenität bedeutet dabei inhaltliche Einheitlichkeit bei gleichzeitiger Unabhängigkeit der Aufgaben voneinander. Universitäre Leistungsteste werden zur größeren Zahl einen mehr heterogenen Charakter besitzen. Der Begriff der Homogenität oder Heterogenität des Testes ist von entscheidender testtheoretischer Bedeutung auch welche Koeffizienten überhaupt treffender Weise berechnet werden sollen und welche Bedeutung ihnen schlussendlich zukommt vgl. dazu auch die Bemerkungen über die Zuverlässigkeitsbestimmungen unter 3.3. Es leuchtet ein, dass eine Zuverlässigkeitsbestimmung des Testes z.B. über die Testhalbierungsmethode nur Sinn macht, wenn sich beide Testhälften genügend ähneln. Auf diese Problematik wird aber noch einmal unter 3.3.2 gesondert eingegangen. Es bestehen sowohl zwischen Homogenität und Aufgabenschwierigkeit, wie unter 3.3.1.4 schon angedeutet wurde, als auch zwischen Homogenität und Trennschärfe wechselseitige Beziehungen. Daher soll für Teste folgende Forderung bestehen:
Für einen heterogenen Test ist der Anspruch zu erheben, dass für jede Aufgabe bei einem Maximum an Trennschärfe ein Minimum an Homogenität besitzen soll. Homogene Teste sollen maximal trennscharfe und zugleich maximal homogene Aufgaben enthalten.
Die Homogenität einer Aufgabe wird bei Lienert, wie folgt definiert: „Die Homogenität einer Aufgabe ist der mittlere Grad, mit dem sie mit allen übrigen Testaufgaben korreliert."

$$r_{i-n} = \frac{\sum r_{ij}}{n-1}$$

wobei i≠j und n die Anzahl der Testaufgaben bedeutet.

Gleichung 18

Das arithmetische Mittel aller Aufgabenkorrelationen kann man unter der Annahme, dass sich die Aufgabenkorrelationen größenordnungsmäßig gleichen, nach Richardson durch die Beziehung:

$$\overline{r_{ij}} = {}_{pbis}\overline{r}^{2}_{it}$$

abgeschätzt werden, wobei ${}_{pbis}\overline{r}^{2}_{it}$ das Quadrat des arithmetischen Mittels aller Trennschärfekoeffizienten, ausgedrückt als punktbiseriale Korrelationskoeffizienten, darstellt.

Gleichung 19

3.3.2 Besonderheiten und Bestimmung der Zuverlässigkeit des Testes:

Wie schon unter 3.3 erwähnt wurde gibt es mehrer Berechnungsarten zur Bestimmung der Zuverlässigkeit, die im Normalfall auch von der Testart (z.B. heterogen versus homogen) abhängen. Grundsätzlich gilt, dass interindividuelle Unterschiede in jeder Testleistung beobachtbar sind, deren Größe in der Varianz der Testrohwerte X zum Ausdruck kommen. Diese Varianz erklärt nun zum einen Teil die „wahre" Unterschiedlichkeiten des Persönlichkeitsmerkmals, die dadurch zustande kommt, dass das Merkmal bei den einzelnen Probanden tatsächlich in verschiedenem Grade vorhanden ist. Dieser wahre Varianzanteil sei in weiterer Folge mit s_{∞}^2 bezeichnet. Wäre es möglich, das Merkmal unabhängig von allen inneren und äußeren Bedingungen zu messen, so würde sich für eine definierte Stichprobe stets dieselbe „wahre" Varianz ergeben. Aufgrund allgemeiner

experimenteller Erfahrung weiß man jedoch, dass jede Messung auch Fehlerquellen besitzt und genau diese verursachen einen Fehleranteil bzw. Fehlervarianz, die sich ebenfalls in der Gesamtvarianz manifestiert. Die Fehlervarianz sei mit s_e^2 benannt. Es gilt nun die einfache Beziehung, dass sich die Gesamtvarianz aus der „wahren" Varianz und der Fehlervarianz zusammensetzt.

$$s_x^{\,2} = s_\infty^{\,2} + s_e^{\,2}$$

Gleichung 20

Nun ist der Reliabilitätskoeffizient einer Messung ganz allgemein bestimmt durch das Verhältnis der „wahren" Varianz zur Gesamtvarianz.

$$r_{tt} = \frac{s_\infty^{\,2}}{s_x^{\,2}}$$

Oder anders formuliert:

$$r_{tt} = 1 - \frac{s_e^{\,2}}{s_x^{\,2}}$$

Gleichung 21

Die Zuverlässigkeit nimmt also mit zunehmender Fehlervarianz ab. Die Störquellen einer Messung sind mannigfach. 2 Gruppen von Störfaktoren sollen aber besonders erwähnt werden.

1. Die Ungenauigkeit des Testes als Messinstrument bzw. seine mangelnde Konsistenz, bezogen auf die Homogenität der Aufgaben und die Aufgabenreihung, gegebenenfalls auch auf die Eindeutigkeit der Aufgabenbewertung (Objektivität).
2. Die Veränderlichkeit der Bedingung der Testdurchführung (z.B. Motivation, körperliches Befinden, situationsbedingte Umstände,...).

Entsprechend ist die Fehlervarianz zu zerlegen in eine Komponente zu Lasten der Konsistenz - $s_{e(consist)}^2$- und eine weitere Komponente zu Lasten der Durchführungsbedingungen – $s_{e(cond)}^2$ – und es gilt:

$$s_e^{\,2} = s_{e(consist)}^{\,2} + s_{e(cond)}^{\,2}$$

Gleichung 22

Will man ausschließlich die Genauigkeit des Testes als Messinstrument der Zuverlässigkeit zugrunde legen, so lässt man $s_{e(cond)}^2$ außer Acht. Reliabilitätskoeffizienten, die $s_{e(cond)}^2$ berücksichtigen, müssen auf solche Art ermittelt werden, dass sich nicht nur die Testkonsistenz sondern auch die Versuchsbedingungen ändern können. Dies ist am ehesten bei der Testwiederholung der Fall (Paralleltest- oder Retestreliabilität), während im Verlauf einer einmaligen Testgabe die Durchführungsbedingungen im allgemeinen als nahezu konstant angesehen werden können.
Da der Retestreliabilitätskoeffizient eine erneute Testdurchführung fordert und der Paralleltestkoeffizient eine Parallelform des Testes verlangt, die auf Äquivalenz überprüft werden muss und oft schwierig wenn nicht gar unmöglich zu finden ist, berechnen wir den Reliabilitätskoeffizienten nach der Methode der Konsistenzanalyse. Diese Vorgehensweise ist für heterogene Teste eigentlich nicht gestattet, da die Voraussetzungen die selben sein müssten wie bei Testhalbierungskoeffizienten. Die Konsistenzanalyse bildet nämlich eine Fortführung des Gedankens der gleichwertigen Testhalbierung des Testes, nämlich eine Unterteilung des Testes in m (m = Anzahl der Aufgaben im Test) vergleichbare Unterteste und damit wie Aufgaben vorhanden sind. Eine Vorraussetzung der

Konsistenzanalyse wäre damit die Homogenität der Aufgaben. Wir verzichten jedoch auf die strenge Einhaltung der Voraussetzung der Homogenität zugunsten des Vorteils der einfacheren Berechnung des Konsistenzkoeffizienten, der nur eine einzelne Darbietung des Testes verlangt. Bei zunehmender Heterogenität des Testes resultiert ein niedrigerer Konsistenzkoeffizient. Die 2. Vorraussetzung für eine eindeutige Interpretation des Koeffizienten, nämlich der Niveaucharakter des Testes kann bei universitären Testen als erfüllt angesehen werden. Somit lautet die Berechnungsvorschrift, wenn man von folgender Datenanordnung und Bezeichnung ausgeht:[19]

	Aufgabe 1	Aufgabe 2	Aufgabe 3	...m	$\sum r$
Proband 1	1	...	...		
Proband 2	0	...			
Proband 3	1				
n...					
$\sum c$					$\sum$

Abbildung 11 Datenmatrix zur Berechnung des Konsistenzkoeffizienten.

In Abbildung 11 haben die Symbole folgende Bedeutung:

X = Rohwert den ein Proband für die Beantwortung einer Frage erhält (0/1),
m = Anzahl der Aufgaben,
n = Anzahl der Probanden,
∑r = Zeilensummen,
∑c = Spaltensummen.

$$QuS_{total} = \sum X^2 - \frac{\left(\sum X\right)^2}{m * n}$$

$$QuS_{zw._den_Pbd.} = \frac{\sum\left(\sum r\right)^2}{m} - \frac{\left(\sum X\right)^2}{m * n}$$

$$QuS_{zw._den_Aufg.} = \frac{\sum\left(\sum c\right)^2}{n} - \frac{\left(\sum X\right)^2}{m * n}$$

$$QuS_{Rest} = QuS_{total} - QuS_{zw._den_Pbd.} - QuS_{zw._den_Aufg.}$$

$$Varianz_{zw._den_Pbd.} = \frac{QuS_{zw._den_Pbd.}}{n-1}$$

$$Rest\,varianz = \frac{QuS_{Rest}}{(n-1)*(m-1)}$$

und schließlich der Konsistenzkoeffizient zu

$$r_{tt} = 1 - \frac{Rest\,varianz}{Varianz_{zw._den_Pbd.}}.$$

Gleichung 23

Diese Berechnung berücksichtigt einmal die Unterschiedlichkeiten sowohl der Aufgaben als auch die der Probanden. Es wird die Fehler- oder Restvarianz in Beziehung zur Varianz $_{zw.\ den.\ Pbd.}$ gestellt, wobei wie Anfangs diskutiert wurde, eine

[19] Vgl. : [7] S 230 ff.

Unterschiedlichkeit zwischen den Aufgaben, die bei vollkommener Homogenität (gleiche Schwierigkeitsindizes der verschiedenen Aufgaben) der Aufgaben auch nicht existieren würde, nicht zu Lasten der Restvarianz gehen darf, gewissermaßen bestehende Schwierigkeitsunterschiede aus der Restvarianz heraus residualisiert werden. Das Problem der Heterogenität der Aufgaben wird durch diese Berechnung freilich ebenso wenig gelöst wie die zufällige Bedingungsvariation bei Parallel- oder Retestkoeffizienten. Ferner fällt auf, dass nicht wie in Gleichung 21 die Fehlervarianz in Bezug zur Varianz der Rohwerte gesetzt wird, die in diesem Falle abzüglich der Unterschiedlichkeit der Aufgaben einer Quadratsumme

$$QuS_{tot*} = QuS_{tot} - QuS_{zw._den\_Aufg.} = QuS_{zw._den\_Pbd.} + QuS_{\mathrm{Re}st}$$

Gleichung 24

entspräche, die noch durch die (n-1)*m Freiheitsgrade dividiert werden müsste, sondern die Fehlervarianz in Bezug zur „Varianz zwischen den Probanden“ gesetzt wird. Wir berechnen somit den Konsistenzkoeffizienten nach der von Lienert bzw. Hoyt angegebenen Vorschrift nach Gleichung 23 und weisen darauf hin, dass wenn man in Gleichung 27 diesen so ermittelte Konsistenzkoeffizient und für (n'/n) = 1/n einsetzt, wobei n für die Anzahl der Aufgaben des Testes steht, die mittlere **Aufgabenkorrelation** folgt. Zur Veranschaulichung soll folgendes Zahlenbeispiel dienen:

	Aufgabe 1	Aufgabe 2	Aufgabe 3	Aufgabe 4	Aufgabe 5	Aufgabe 6
Proband 1	1	1	0	1	1	1
Proband 2	0	0	1	0	1	0
Proband 3	1	1	1	1	1	1
Proband 4	0	0	0	0	0	0
Proband 5	1	1	1	1	1	1
Proband 6	0	0	0	0	0	0
Proband 7	1	1	1	1	1	1
Proband 8	0	0	0	0	0	0
Proband 9	1	0	0	0	0	0
Proband 10	1	1	1	1	1	0
Schwierigkeit	0,6	0,5	0,5	0,5	0,6	0,4

Abbildung 12

In diesem Beispiel errechnet sich der Konsistenzkoeffizient nach Gleichung 23 aufgerundet zu 0,94. Die mittlere Aufgabenkorrelation erhalten wir, indem wir den Konsistenzkoeffizienten des 6-fach verkürzten Testes (entspricht der mittleren Aufgabenkorrelation) für n'/n = 1/6 entsprechend unseres Beispiels (6 Aufgaben) mittels der **Spearmen_Brown** Beziehung nach Gleichung 27 berechnen. Die so aus dem Konsistenzkoeffizienten abgeleitete mittlere Aufgabenkorrelation berechnet sich abgerundet zu 0,71.
Die über alle 15 Kombinationen der Aufgaben gemittelten Korrelationen ergeben sich aufgerundet zur mittleren Aufgabenkorrelation von 0,71 (exakt 0,70990929), während die Aufgabenkorrelation nach Gleichung 19 aufgerundet auf 0,76 geschätzt wird. Wie aus dem Beispiel ersichtlich wird, liegt die aus der internen Konsistenz berechnete Aufgabenkorrelation wesentlich besser als die nach Gleichung 19 geschätzte Aufgabenkorrelation, was auf die Verletzung der annähernd gleichen Korrelationen, was eine Vorraussetzung für Gleichung 19 darstellt, zurückzuführen ist. Denn die Korrelationen bewegen sich zwischen den jeweils aufgerundeten Extremen 0,41 (Korrelation der Aufgabe z.B. [1,3]) und 0,82 (z.B. [1,2]). Daher erscheint die Berechnung der mittleren Aufgabenkorrelation aus dem Konsistenzkoeffizienten gegenüber Verletzungen von Voraussetzungen robuster. Je weiter die Schwierigkeitsindizes der Aufgaben auseinander liegen, umso ungenauer wird auch

die Schätzung der mittleren Aufgabenkorrelation über den Konsistenzkoeffizienten nach Gleichung 23. Um dennoch die exakte Aufgabenkorrelation zu erhalten müssten die Werte der Aufgaben bezüglich der jeweiligen Aufgabenschwierigkeit p_i normalisiert werden, wie für Proband j angedeutet wurde.

	Aufgabe i
Proband 1	1
Proband 2	0
	⋮
Proband j	$\frac{a_{ij} - p_i}{\sqrt{p_i * (1 - p_i)}}$
⋮	
Schwierigkeit	pi

Abbildung 13

Damit vereinfacht sich Gleichung 23, da die „Quadratsumme zwischen den Aufgaben“ 0 wird. Die Restquadratsumme bekommt dann die einfache Form:

$$QuS_{\mathrm{Re}\,st} = QuS_{total} - QuS_{zw._den\_Pbd.}$$

Gleichung 25

Die Restvarianz berechnet sich indem obige Restquadratsumme durch die aufgrund der Normalisierung verminderten Freiheitsgrade (m-1)*(n-2) dividiert wird. Da jedoch auch die Freiheitsgrade der Quadratsumme zwischen den Probanden auf (n-2) sinkt, bleibt das Verhältnis der Freiheitsgrade $df_{Rest}/df_{zw\_den\_Pbd}$ = (m-1)*(n-2)/(n-2) und das Verhältnis $df_{Rest}/df_{zw\_den\_Pbd}$ = (n-1)*(m-1)/(n-1) nach Gleichung 23 unverändert, sodass zur Berechnung des Konsistenzkoeffizienten unverändert Gleichung 23 benützt werden darf. Wird aus diesem Konsistenzkoeffizienten nun die mittlere Aufgabenkorrelation nach Spearmen-Brown berechnet erhalten wir das exakte Ergebnis.
Die Größe der Zuverlässigkeit des Testes hängt auch von der Streuung der Testrohwerte ab. Wenn wir annehmen, dass der Messfehler s_e nach Gleichung 21 in einer Stichprobe mit der kleineren Varianz s_x^2,die zur Ermittlung der Zuverlässigkeit untersucht wurde, derselbe ist wie in einer Stichprobe, die die größere Varianz S_x^2 besitzt aber die zu testende Grundgesamtheit besser repräsentiert als die vielleicht soziologisch zu homogen zusammengesetzte Gruppe bei der Ermittlung des Reliabilitätskoeffizienten, so kann durch umformen der Beziehungen für s_x^2 bzw. r_{tt} und S_x^2 bzw. R_{tt} in Gleichung 21 eine Korrektur des Reliabilitätskoeffizienten entsprechend den vorliegenden Varianzverhältnissen gefunden werden.

$$R_{tt} = 1 - \frac{s_x^{\,2}(1 - r_{tt})}{S_x^{\,2}}$$

Gleichung 26

Zuletzt sollen noch Möglichkeiten aufgezählt werden mit deren Hilfe die Zuverlässigkeit eines Testes beeinflusst werden können.

Dazu gehören:

- Testverlängerung,
- Streuungsbereich der Aufgabenschwierigkeit,
- Aufgabenkorrelation und die
- Objektivität der Testauswertung.

Der wirksamste Faktor stellt die **Testverlängerung** dar. Sind die neu hinzukommenden Aufgaben von der gleichen Art wie die bereits vorhandenen, so kann man die **Spearman-Brown**-Beziehung verwenden, um vorauszusagen, um wie viel die Zuverlässigkeit des Testes ansteigen wird, wenn man eine bestimmte Anzahl von Aufgaben hinzunimmt. Natürlich muss kritisch angemerkt werden, dass ein Test nur in gewissem Umfang verlängert werden kann, da er ökonomisch und zeitlich zumutbar bleiben muss.

$$r'_{tt} = \frac{\frac{n'}{n} * r_{tt}}{1 + \left(\frac{n'}{n} - 1\right) * r_{tt}}$$

In dieser Formel bedeuten:
r'_{tt} = Voraussichtliche Zuverlässigkeit des veränderten Testes,
r_{tt} = Zuverlässigkeit des ursprünglichen Testes,
n = Anzahl der Aufgaben im ursprünglichen Test,
n' = Anzahl der Aufgaben im veränderten Test.

Gleichung 27

Eine Vorhersage des neuen Reliabilitätskoeffizienten bis zu einer Verdoppelung der Aufgabenzahl auch unter sehr ungünstigen Bedingungen z.B. bei Hinzunahme relativ heterogener Aufgaben liefert noch nach Holzinger und Clayton einigermaßen befriedigende Voraussagen.[20]
Je mehr Aufgaben mittlerer **Schwierigkeit** ein Test enthält, umso höher fällt seine Zuverlässigkeit unter sonst gleichen Bedingungen aus. Hat der Test z. B. eine große Streuung der Aufgabenschwierigkeit, so kann man also zur Zuverlässigkeitsverbesserung einen Teil der Aufgaben mit extremer Schwierigkeit durch solche mit mittlerer Schwierigkeit ersetzen.
Einzelne allzu hohe **Aufgabenkorrelationen** können zu einer Zuverlässigkeitsverminderung führen, indem sie durch eine Gleichbeantwortung der Aufgaben indirekt zu einer Testverkürzung beitragen.
Es ist selbstverständlich, dass die **Objektivität der Testauswertung** eine grundsätzliche Voraussetzung für die Zuverlässigkeit des Testes darstellt. Eine Verbesserung der Auswertungsobjektivität der Aufgaben kann durch eine Präzisierung der Auswertungsinstruktion erreicht werden.
Für die Beurteilung individueller Differenzen wird ein Reliabilitätskoeffizient von r_{tt} = 0.7 bei Lienert als gerade noch ausreichend angegeben.

3.4. Ansätze zur Qualitätsmessung von Studium und Lehre:

Wie bereits oben kurz angedeutet wurde, besteht das vornehmliche Ziel von Qualitätsmanagement Prozesse zu beherrschen und in ihrer **Wirksamkeit** zu verbessern. Das bedeutet in weiterer Konsequenz den Kernbereich der Lehre auf festgesetzte Qualitätsstandards hinsichtlich erstens **informell, inhaltlicher,** zweitens **formaler** Aspekte und drittens **geeigneter Rahmenbedingungen** zu überprüfen.

[20] Vgl.: [7]

Unter Rahmenbedingungen sollen in Zukunft Variablen besser Faktoren verstanden werden, die in direktem Bezug zum Bildungssystem stehen, dass heißt, strukturelle und organisatorische Bedingungen und Gegebenheiten unter denen Studium und Lehre vollzogen werden.
Die informell, inhaltlich Komponente der Lehre bildet die Schnittstelle zwischen Student als Objekt und Ausbildungsprozess. Anders als bei den Objektdaten unterliegen die Prozessdaten einer Möglichkeit der direkten Einflussnahme, wobei das **inhaltlich, informelle** Element größtenteils von Professorenschaft, gesetzlichen Bestimmungen und Wirtschaft beeinflusst wird, auf das **Formale** durch den Lehrkörper und Studentenschaft eingewirkt bzw. beurteilt werden sollte und die **Rahmenbedingungen** vornehmlich vom Lehrkörper wie auch vom Staat generiert werden.
Die Feststellung der **inhaltlichen** Sachdienlichkeit erstens zur Berufsbefähigung als auch zum verständlichen Wissenserwerb kann am Besten über Experten vorgenommen werden, wobei die Studierbarkeit stark von der inhaltliche Abstimmung des Lehrangebots geprägt ist. Die Expertenschaft sollte sich dabei sowohl aus Professoren der jeweiligen Fachrichtung als auch aus Mitgliedern der Wirtschaft konstituieren, um einerseits sicherzustellen, dass Lehrpläne erstellt werden, die sich am letzten Stand der Forschung orientieren, und andererseits Wünsche und Bedürfnisse des Marktes Eingang und Gehör im jeweiligen Ausbildungsprogramm finden. Eine zusätzliche Möglichkeit vermeintliche, inhaltliche Schwächen der Ausbildung zugänglich und transparent zu machen, besteht durch einen Vergleich des Anforderungs- mit dem Kompetenzprofils, das über eine Befragung der bereits im Beruf stehenden Absolventen (siehe Kapitel 5) ermittelt werden kann. Diese Einschätzungen der Profile durch die Absolventen können, wenn der Aufwand der Untersuchung es erlaubt, durch Beurteilungen der Arbeitgeber oder Vorgesetzten ergänzt oder validiert werden. Dabei werden 2 Aspekte untersucht. Einmal ob und wie sich studentische Leistungen im späteren Beruf (Stellung, Aufgabenbereich, Verdienst,...) niederschlagen und zweitens welche Bereiche im Lehrplan einer verstärkten Behandlung bedürfen, um den Anforderungen der Absolventen in der späteren Berufswelt gerecht zu werden.
Die Überprüfung des **formalen** Aspekts der Lehre, ob das Wissen in qualitativ zufriedenstellender Weise transportiert wurde und wird, kann mit dem mittlerweile sehr routinierten und viel untersuchten Instrument der „Lehrevaluation“, eines Bewertungsverfahrens durch Studenten mit der Möglichkeit der Erweiterung durch Fremdhörer bewerkstelligt werden.
Zuletzt sollen aber auch Wirksamkeiten von **Rahmenbedingungen**, ohne deren Vorhandensein zweifelsohne keine Lehre stattfinden könnte, untersucht werden und Möglichkeiten der Aufdeckung von empirischen Zusammenhängen mit Kriteriumsvariablen erläutert werden.

3.4.1 Die Lehrevaluation, ein Instrument der Erhebung der Vortragsqualität als formalem Aspekt der Lehre als auch ein Mittel zur Erfassung nicht objektiv messbarer Daten im Sinne von Bewertungen und Urteilen:

Es kann auf die Erhebungsart mittels Fragebögen, obwohl sie vielleicht mehr dem subjektiven Charakter unterliegt als ein standardisierter Test, bei einer ganzheitlichen Evaluation der Lehre nicht verzichtet werden, da verschiedene, relevante Aspekte der Lehre unter anderem der Formale (Vortragsqualität) nur über eine Bewertung zugänglich gemacht werden können, mithin keine objektiven Messkriterien vorliegen, und eine zunehmende Objektivierung nur über eine intersubjektive Mittelung verschiedener Beurteilungen gewonnen werden kann. Ziel der Lehrevaluation soll

eine Bewertung über die formale Vortragsqualität sein, die durch zahlreiche Faktoren wie fachliche Kompetenz des Vortragenden, rhetorische Fähigkeiten, Didaktik und Strukturierung des Vortrages..., beeinflusst wird. Es soll in Anlehnung zu anderen Untersuchungen und Evaluationen die Vortagsqualität über mehrer Dimensionen erfasst werden und nicht nur eine generelle, nebulöse Zufriedenheitseinschätzung über den Vortrag der Studenten erhoben werden. Dies ermöglicht auch einen, gezielter auf Schwächen und Mängel, die durch die mehrdimensionale Befragung aufgedeckt werden können, einzugehen und koregierende Maßnahmen eben in jenen unzureichend bewerteten Bereichen zu setzen. Es sollen also mehrdimensionale Lehrevaluationsbögen zum Einsatz kommen und generiert werden, die sich sowohl an den Anforderungen der Lehrpraxis orientieren als auch eine Feedbackfunktion für die Lehrenden darstellen. Abgefragt werden dabei Dimensionen der Lehrqualität, die unmittelbar mit dem Dozenten in Verbindung stehen, als auch solche Rahmenbedingungsvariablen der Veranstaltung, die nur oder leichter über eine Bewertung zugänglich zu machen sind (z.B. Ausstattung mit Ressourcen / Lernmaterial zufriedenstellend?). Ein Vergleich der Wirksamkeit der Rahmenbedingungen über persönliche Einschätzungen mit gewissen Kriteriumsvariablen ist oft von größerer Aussagekraft als mit objektiv gewonnen Rahmenbedingungsdaten. So scheint es z. B. einleuchtend, dass der objektiv erfassbare Bücherbestand der Universitätsbibliothek, dessen Bestand über die Jahre beinahe gleichsam konstant und damit seine Variabilität nahezu null bedeutet, beinahe keine Kovarianz und damit einen vernachlässigbaren Zusammenhang mit der Kriteriumsvariablen Lernfortschritt aufzeigen würde (Veränderung des totalen Bücherbestandes resultiert nicht zwingend in einer Veränderung von knappen Ressourcen), wohl aber eine Beziehung zwischen Reservierungen von bestimmten Lernbüchern und Lernfortschritt oder -misserfolg aufgezeigt werden könnte, die aber zumeist wegen der Einfachheit der Erfassung über die Zufriedenheitseinschätzung über die Ressourcenversorgung des Studenten gewonnen wird.
Ein nicht unwesentlicher Punkt bei Evaluationen stellt auch der Einfluss von urteilsverzerrenden Größen der Beurteilung, sogenannten Biasvariablen, dar, die ebenfalls zur Überprüfung der Gültigkeit studentischer Urteile über die Vortragsqualität durch die Evaluationsfragebögen erhoben werden müssen.
Weiter ist eine Ermittlung studentischer Merkmale und Eigenheiten, die ebenfalls zum Lernerfolg des Prüflings beitragen, für eine korrekte Auswertung der wechselseitigen Beziehungen zwischen den diversen Größen unerlässlich. Es wäre falsch die studentischen Eigenheiten und Merkmale als Biasvariablen zu bezeichnen, da diese wirklich die Qualität und die Art und Weise des Lehrvortrages beeinflussen und verändern können und nicht wie eben die Biasvariablen „nur“ die Beurteilungen der Studenten verändern und verfälschen, wobei eine theoretische Wechselwirkung der Biasvariablen mit Lehrverhaltensskalen ausgeschlossen wird. So kann bei einer Gruppe, die über ein sehr unterschiedliches Leistungspotential besitzt, es Schwierigkeiten bereiten ein für alle angemessenes Anforderungsniveau zu wählen, dass heißt aber das studentische Merkmale tatsächlich den Unterricht in einem gewissen Umfang mitgestalten. Häufig sind auch die Skalen nicht immer eindeutigen Dimensionen zuzuordnen. So kann z.B. die Skala Thema/Interesse am Fach sowohl als Bias oder als Rahmenbedingung angesehen werden, da eine Evaluation zumeist Mitte oder am Ende des Semesters stattfindet, und für eine eindeutige Trennung in einen Faktor Bias und Rahmenbedingung ein Vorinteresse am Thema, bevor die Unterweisung stattfindet, erhoben werden müsste. So aber ist der Bewertung nicht eindeutig zu entnehmen, inwieweit sich im Interesse auch schon ein Vorinteresse widerspiegelt, dass heißt inwieweit das Vorinteresse die Bewertung der Variablen

Interesse an der Vorlesung beeinflusst hat und welcher Anteil der Bewertung Interesse am Fach auf die Lehrkompetenz, sprich Vermittlung des Stoffes, zurückzuführen ist. Die Korrelation zwischen zwei Größen, in unserem Beispiel die Größen Interesse am Fach und der Lehrkompetenz, kann durch mehrere Kausalmodelle, –beziehungen zustande kommen.

- X_1 ist die Ursache für X_2,
- X_2 ist die Ursache für X_1,
- X_1 und X_2 beeinflussen sich wechselseitig
- X_1 und X_2 werden von einer dritten oder weiteren Variablen beeinflusst.

Anhand der Korrelation alleine kann nicht entschieden werden, welches der oben genannten Modelle zutrifft.
Während eine Korrelation zwischen den Größen Interesse am Fach und Lehrkompetenz, ob nun das Interesse am Fach vorwiegend durch die hervorragende Lehrkompetenz geweckt wurde oder aber durch das von vornherein höhere Interesse am Fach die Lehrperson dazu veranlasst wurde, den Unterricht dem höheren Interesse anzupassen, die „theoretisch“ erwartete Wechselwirkung zwischen diesen zwei Größen im Sinne einer modelltheoretischen Annahme bestätigen kann aber nicht muss, so sollte zumindest doch eine Korrelation zwischen diesen zwei Größen bestehen , wenn zwischen diesen Größen eine Abhängigkeit bestehen soll, da die Korrelation eine notwendige, wenn auch nicht hinreichende Bedingung für eine Kausalhypothese zwischen den zwei Variablen darstellt.
Im Gegensatz dazu wäre es wünschenswert, wenn sogenannte Biasvariablen, Variablen, die modelltheoretisch keinen Einfluss auf die „wahre“ Qualität des Unterrichts haben sollten, einen eher kleinen bis gar keinen verzerrenden Effekt auf die studentischen Beurteilungen der Vortragsqualität haben, da dies die Gültigkeit studentischer Evaluation untermauert. Sie stellen damit verzerrende Merkmale der studentischen Veranstaltungswahrnehmung und Fragebogenbeantwortung dar, die nicht mit den tatsächlichen, wirklichen Geschehnissen einer Veranstaltung zusammenhängen. Lässt sich nun die Annahme eines solchen verzerrenden Effektes durch Korrelationen um 0 zwischen den Biasgrößen und den diversen Skalen und Kriteriumsvariablen aber nicht entkräften, müssen sie mit in die Beurteilungen einfließen und gegebenenfalls das Urteil um den verzerrenden Effekt bereinigt werden. Manche möglichen Verzerrungseffekte können durch eine richtig terminierte Lehrevaluationserhebung ausgeschaltet beziehungsweise kontrolliert werden. So sollten zum Beispiel um eine Beeinflussung der Evaluationsbewertung durch die Notengebung zu verhindern, die Erhebung stets vor der Durchführung einer Prüfung stattfinden. Auf diese Weise lassen etwaige Korrelationen zwischen den Noten und Veranstaltungsbewertungen eher den Schluss einer Validitätshypothese zu, da in diesem Fall es wahrscheinlicher ist, dass die hohe Lerneffektivität und Vortragsqualität Grund für die guten Prüfungsleistungen sind, als im anderen Fall, wo gute Noten ungerechtfertigter Weise auch die Beurteilung milder ausfallen lassen können oder von guten Noten implizit auch automatisch auf gute Lehre geschlossen wird, was beides für einen Bias, eine Urteilsverzerrung sprechen würde. Der Katalog möglicher **Biasvariablen** ist mannigfach und können wie schon oben angedeutet nicht immer streng von anderen Urteilsdimensionen getrennt werden. Es sollen hier exemplarisch ein paar aufgezählt werden, wobei die Liste keinen Anspruch auf Vollständigkeit erhebt und durch weitere, für wichtig empfundene Verzerrungsgrößen ergänzt werden soll.

- Geschlecht der Studierenden,
- Vorinteresse am Fach (entweder bei Eingangstest erhoben oder retrospektive Einschätzung der Studenten),

- Popularität des Dozenten,
- Relevanz des Faches für späteren Beruf,
- Attraktivität der Lehrkraft,
- Motivation,
- Sympathie der Lehrkraft
- Gerüchte über Dozenten oder Fach,
- Pflichtfach versus Interessensfach (Besuchsgrund, Prüfungsrelevanz des Faches).

Bezugnehmend auf schon existierende Fragebögen und Erfahrungen anderer empirischer Studien[21] in der Evaluationspraxis, deren gefundene Dimensionen/ Skalen sich meist auf Studierenden- und Lehrendenbefragung und auf lehrtheoretische, didaktischen Annahmen sich stützten, werden folgende Dimensionen für eine Befragung mittels Fragebögen vorgeschlagen.

- **Persönlichkeit des Dozenten**: humorvoll, freundlich, gerecht, offen, hilfsbereit, flexibel, ansprechbar, kreativ selbstbewusst, souverän, spontan, ernsthaft, nahbar, sozial kompetent, vielseitig.
- **Lehrkompetenz**: Es soll die allgemeine Fähigkeit zu lehren gemessen werden. Diese äußert sich z.B. in einer guten Vorbereitung, Präsentation des Stoffes, rhetorische Fähigkeiten (deutliche, laute, klare Aussprache, freie Vortragsweise, verständliche, einfache Sprache; angemessenes Sprechtempo, gute Ausdrucksweise; lebendiger, anregender Sprechstil, flüssiger Sprechstil, redesicher) oder Kompliziertes erklären, auf wichtiges hinweisen zu können, Querverbindungen aufzuzeigen, gutes Zeitmanagement zu besitzen und die Beherrschung der verschiedenen und Auswahl der adäquaten Präsentationstechniken, Anschaulichkeit und graphische Darstellungen so weit möglich in den Unterricht einfließen zu lassen,...
- **Strukturierung des Lehrstoffes /der Veranstaltung**: Eine klare und verständliche Gliederung des Stoffes stellt eine der offensichtlich wichtigsten Bedingungen guter Lehre dar. Eine logisch aufgebautes und nachvollziehbares Konzept sowohl hinsichtlich des Vortrages, des Ablaufes als auch der Hilfsmaterialien (Skript) erleichtern die Aufnahme des Stoffes und den späteren Abruf des Erlernten.
- **Anwendungsmöglichkeiten**: Vermittlung, wie der gelernte Stoff in der Praxis angewendet wird und in welchen Bereichen das erlernte Verwendung findet. Zu diesem Aspekt zählt vor allem die beispielhafte Behandlung des Stoffes, durch das die Relevanz des Gelernten aufgezeigt und ein Theorie-Praxis Bezug hergestellt wird.
- **Themeninteresse**: Nachträgliche Einschätzung der Bedeutsamkeit und informativen Brisanz des Inhaltes (wie interessant war der Vortrag rückblickend für mich ?). Im Gegensatz zum Vorinteresse wird eine Beeinflussung des Themeninteresses mit der Vortragsqualität postuliert. Ein Vergleich der absoluten Bewertungen dieser Dimension über verschiedene Veranstaltungsthemen und Dozenten ist nicht zulässig, da der Lehrkörper nur begrenzt auf die Thematik eingehen kann und diese auch als vorgegebene Rahmenbedingung aufgefasst werden muss (teils vom Lehrplan vorgeschrieben).
- **Engagement und Enthusiasmus am Lehren**: Engagement und Enthusiasmus des Vortragenden steigert die Motivation der Studenten, sich mit dem Thema eingehender zu beschäftigen, bewirken ein größeres

[21] Vgl. [9],[10],[11].

Interesse im Kurs und steigert die Anstrengungsbereitschaft der Studenten. Zu dieser Skala kann auch die Bereitschaft des Dozenten gezählt werden, inwieweit dieser Studenten ausreichend betreut und sie mit Rückmeldungen versorgt. Feedback und Betreuung steigert das Zufriedenheitsgefühl der Hörer und sichert die Qualität studentischer Beiträge und damit der Ausbildungsqualität.

- **Leitung und Handhabung des Unterrichts**: Studentische Diskussion bedürfen einerseits der Förderung durch Schaffen einer angenehmen Atmosphäre und aktiven Einbezug durch an die Hörerschaft gerichtete Fragen andererseits benötigt sie auch eine Lenkung und gezieltes Eingreifen des Dozenten um ein Abschweifen der Diskussion zu vermeiden.
- **kognitive Fähigkeiten**: intellektuelle und kognitive Fähigkeiten sollten gesondert über einen extra dafür vorgesehenen Eingangstest erhoben werden, da damit eine bessere Vergleichbarkeit der erzielten Noten, Leistungen gegeben ist und eine homogenere Be-und Auswertung derselben mehr gesichert scheint. Ansonsten muss auf die Schulabschlussnote in einem für das Studium/Fach repräsentativen Fach zurückgegriffen werden. Ein weiterer Vorteil einer gesonderten Überprüfung des Wissenstandes der Studenten besteht darin, dass der Dozent sein Lehrprogramm in gewissem Umfang dem Wissenstand der Studenten anpassen kann und damit eine optimale, effektive Wissensvermittlung ermöglicht.
- **Fleiß**: Fleiß als Merkmal der individuellen Arbeitshaltung ist entscheidend für den Lernerfolg von Studenten mitverantwortlich. Dieser ist nicht unabhängig von äußeren Bedingungen wie die Existenz einer Prüfung und die Höhe der Anforderung. Damit Fleiß nicht zu einer subjektiven Einschätzung auf einer an sich beliebigen Skala degradiert- haben sie das Gefühl das sie fleißig sind? 1 für trifft nicht zu bis z.B. 7 für trifft vollkommen zu -und somit absolute Niveauunterschiede ohne Bedeutung sind, da jede Relation zwischen den einzelnen Einschätzungen fehlt, sollte stets nach der investierten Lernzeit (absolute oder aber besser pro Woche oder pro Tag, da genauer einschätzbar) gefragt werden und diese dann entsprechend der benützten Anzahl von Kategorien skaliert werden. Eine mögliche Einteilung der Skala Fleiß könnte wie folgt aussehen:

Lernzeit:	Skala:
weniger bis max. 45 min./Tag	1
mehr als 45 min. bis max. 1Std. 30 min./Tag	2
mehr als 1 Std. 30 min. bis max. 2 Std.15 min./Tag	3
mehr als 2 Std. 15 min. bis max. 3 Std./Tag	4
mehr als 3 Std. bis max. 3 Std. 45 min./Tag	5
mehr als 3 Std. 45 min. bis max. 4 Std. 30 min./Tag	6
mehr als 4 Std. 30 min./Tag	7

Abbildung 14

- **Belastbarkeit/Stressbewältigung**: Diese Größe könnte quasi als Ersatzgröße, Prädiktor für die unterschiedliche Nervosität der Probanden bei Prüfungsleistungen dienen, die die Ergebnisse der Leistungsüberprüfung verzerren können. Die Nervosität kann aufgrund der Tatsache, da Lehrveranstaltungsevaluationen prinzipiell vor Prüfungsablegung stattfinden sollten, nicht direkt erhoben werden.[22]

[22] Vergleiche Ausführungen weiter oben.

- **Anforderung**: Die Höhe der Anforderung stellt ein entscheidendes Merkmal sowohl der Charakterisierung des Vortrages hinsichtlich Stoffumfang, Schwierigkeit und Tempo des vorgetragenen Stoffes als auch im Zusammenhang mit den kognitiven Fähigkeiten der Studenten dar. Im Gegensatz zu Fleiß und kognitiven Fähigkeiten, die Aspekte der Persönlichkeit der Studenten einbeziehen, sollen auf der Variablen Anforderung nur Veranstaltungsmerkmale attribuiert werden, sprich Veranstaltungsunterschiede angezeigt werden.
- **Prüfung**: Wie bereits erwähnt, stellt die Tatsache einer abschließenden Prüfung zum Scheinerwerb eine immanente Wirkung auf das Lernverhalten der Studenten dar. So zeichnet sich die Strategie des Wissenserwerb bei Vorhandensein einer Abschlussprüfung durch hohe Effektivität in dem Sinne aus , dass möglichst viel „prüfbares“ Wissen pro Zeiteinheit gepaukt wird, wohingegen Breite, Tiefe und Diskussion des Erlernten in den Hintergrund tritt. Je schwieriger die Prüfung erwartet wird, umso deutlicher wird sich dieser Effekt zeigen.
- **Ressourcenausstattung**: Die Frage nach der Ressourcenausstattung gibt ebenfalls entscheidende Hinweise, ob und in welchem Umfang sie erfolgt ist und in weiterer Folge inwieweit sie für das Scheitern oder Gelingen des Lernerfolgs ausschlaggebend war. Hier empfiehlt es sich neben der bloßen Bewertung, da die Ressourcenausstattung mehrere unterschiedliche Bereiche tangieren kann (Skripten, Lehrbücher, Labor- oder Hörsaalplätze, Erreichbarkeit der Assistenten /Professoren,...), auch offene Antwortmöglichkeiten zu ermöglichen, um den Studenten im Falle einer ungenügend oder mangelhaft empfundenen Darbietung die Möglichkeit zu geben, ihre Aussage auf die entscheidenden Bereiche zu präzisieren und eine prompte Reaktion seitens der Universität/ Institutes zu ermöglichen.
- **Zeitpunkt und –Dauer der Veranstaltung**: Nicht nur die Zeitdauer einer Veranstaltung ist von entscheidender Bedeutung für die Studenten den Stoff aufnehmen und verarbeiten zu können, sondern auch zu welchem Zeitpunkt die Veranstaltung angesetzt wird. Da die benötigte, absolute Zeitdauer durch die Stoffmenge, die im Semester vermittelt werden soll, festgesetzt ist, kann nur eine der Größen, Zeitdauer oder Anzahl der Veranstaltungen pro Woche in gewissen Grenzen gewählt werden. Grundsätzlich „frei“ wählbar ist dagegen der Zeitpunkt, zu dem die Veranstaltung festgesetzt werden soll. Hierzu sollte folgende Faustregel beachtet werden: Erfolgt die Unterweisung in längeren Lehreinheiten, sollten diese eher in der Früh, sprich morgens unterrichtet werden, da zu diesem Zeitpunkt die Aufmerksamkeit der Studenten noch höher ist. Kürzere Einheiten können auch noch am späteren Nachmittag oder abends gelehrt werden, wobei natürlich noch die Schwierigkeit besteht die Lehreinheiten der verschiedenen Fächer gegenseitig räumlich und zeitlich im Semester kollisionsfrei unterzubringen. Auf jeden Fall sollte aber sowohl die Zeitdauer als auch der Zeitpunkt der Veranstaltung zu Auswertungszwecken erhoben werden.

Zur Zeichenerklärung: die mit

- gekennzeichneten Skalen gehören zur Oberkategorie *Dozent*, die mit
- gekennzeichneten Skalen gehören zur Oberkategorie *Student* und die mit
- aufgezählten Skalen werden der Kategorie *Rahmenbedingung* zugeschrieben.

Die Dozentenkategorie enthält Skalen/ Dimensionen, die primär Eigenschaften und Merkmale der Lehrkraft beschreiben, die Oberkategorie der Studenten solche der Lernenden und schließlich beinhaltet die Rahmenbedingungskategorie Skalen, die

Veranstaltungsattribute skizzieren. Skalen, die ihrerseits keiner Beurteilung durch Studenten bedürfen, sondern direkt erhoben werden können, werden natürlich nicht in den Fragebogen aufgenommen (Zeitpunkt/-dauer der Veranstaltung, Prüfung,...). Unerlässlich ist dabei die **Erhebung der Teilnehmerzahlen** an den Veranstaltungen, um die Besucherzahl der Studenten an der jeweiligen Vorlesung zu erfassen, die in weiterer Folge als Richtlinie zur Bestimmung der geforderten Mindestrücklaufquote dient.[23]
Ein möglicher Rohentwurf für einen Fragebogen, der die oben genannten Skalen zur Bewertung durch die Studenten enthält, könnte wie folgt aussehen.

[23] Siehe Kapitel 3.5.

Dimension/Skala:	Items:	Bewertung (1 für trifft überhaupt nicht zu 7 für trifft vollkommen zu):
Persönlichkeit des Dozenten	1. Die Lehrerin/Lehrer war zu Schüler/innen freundlich.	1 2 3 4 5 6 7
	2. Die Lehrerin/Lehrer war allen Schüler/innen gegenüber gerecht.	1 2 3 4 5 6 7
	3. Der/die Lehrende war humorvoll.	1 2 3 4 5 6 7
Lehrkompetenz	4. Die eingesetzten Medien haben zum Verständnis der Inhalte beigetragen.	1 2 3 4 5 6 7
	5. Der/die Dozentin konnte Komplizietes verständlich machen.	1 2 3 4 5 6 7
	6. Der/die Dozentin wirkt gut vorbereitet.	1 2 3 4 5 6 7
	7. Der/die Lehrende spricht klar und verständlich.	1 2 3 4 5 6 7
	8. Der/Die Lehrende konnte adäquat auf die Fragen der Studierenden antworten.	1 2 3 4 5 6 7
Strukturierung des Lehrstoffes/der Veranstaltung	9. Der vermittelte Stoff besitzt eine inhaltlich, klare Gliederung.	1 2 3 4 5 6 7
	10. Die Veranstaltung wurde straff und zeitlich gut organisiert.	1 2 3 4 5 6 7
Anwendungsmöglichkeiten	11. Der Stoff wird mit Anwendungsbeispielen veranschaulicht.	1 2 3 4 5 6 7
	12. Die Bedeutung des Gelernten wird auch praktisch ersichtlich.	1 2 3 4 5 6 7
Themeninteresse	13. Das Thema/ der Vortrag war rückblickend interessant für mich.	1 2 3 4 5 6 7
Engagement und Enthusiasmus am Lehren	14. Dem/ der Dozentin ist es wichtig, dass Studenten etwas lernen.	1 2 3 4 5 6 7
	15. Der/Die Lehrende ging auch nach der Vorlesung auf Fragen ein.	1 2 3 4 5 6 7
	16. Die Lehrerin/Lehrer bereut auch außerhalb der Sprechstd. die Studenten gut.	1 2 3 4 5 6 7
Leitung und Handhabung des Unterrichts	17. Der/die Dozentin fördert Fragen und aktive Mitarbeit.	1 2 3 4 5 6 7
	18. Der/die Dozentin geht gut auf Beiträge ein.	1 2 3 4 5 6 7
Fleiß	19. Durchschnittliche Beschäftigung mit dem Stoff pro Tag:	
	weniger bis max. 45 min./Tag,	1
	mehr als 45 min. bis max. 1Std. 30 min./Tag,	2
	mehr als 1 Std. 30 min. bis max. 2 Std. 15 min./Tag,	3
	mehr als 2 Std. 15 min. bis max. 3 Std./Tag,	4
	mehr als 3 Std. bis max. 3 Std. 45 min./Tag,	5
	mehr als 3 Std. 45 min. bis max. 4 Std. 30 min./Tag,	6
	mehr als 4 Std. 30 min./Tag.	7

Abbildung 15 Vorschlag eines Lehrevaluationsfragebogens

Belastbarkeit/Stressbewältigung	20. In Prüfungssituationen reagiere ich normalerweise über die Maßen nervös.	1 2 3 4 5 6 7
	21. Es fällt mir schwer unter Druck zu arbeiten.	1 2 3 4 5 6 7
Anforderung	22. Das Tempo des Vortrages ist mir viel zu schnell.	1 2 3 4 5 6 7
	23. Ich habe immense Schwierigkeiten den Stoff zu verstehen.	1 2 3 4 5 6 7
	24. Der Umfang des zu lernenden Stoffes überfordert mich extrem.	1 2 3 4 5 6 7
Ressourcenausstattung	25. Die Ressourcenausstattung hat mich vollends zufrieden gestellt.	1 2 3 4 5 6 7
offene Antwortmöglichkeiten bei unzureichender Ressourcenausstattung:	1.)	
	2)	
	3.)	
kognitive Fähigkeiten*	26. Welche Schulabschlußnote hatten Sie im Fach Mathematik?	
	Mathematiknote:	
	27. Welche Schulabschlußnote hatten Sie im Fach Physik?	
	Physiknote:	
Biasgrößen:		
Geschlechtsverzerrung	28. Geschlecht des Studenten?	0 männlich 1 weiblich
Vorinteresse am Fach*	29. Wie groß war das Interesse Ihrer Einschätzung nach für das Fach vor ihrer Unterweisung?	1 2 3 4 5 6 7
Popularität/ Sympathie der Lehrkraft	30. Der/die Lehrende hat einen guten Ruf.	1 2 3 4 5 6 7
	31. Mir war das Auftreten der/des Lehrenden sympathisch.	1 2 3 4 5 6 7
Relevanz des Faches für späteren Beruf	32. Die Inhalte scheinen mir für meinen späteren Beruf irrelevant.	1 2 3 4 5 6 7
Attraktivität der Lehrkraft	33. Die Lehrkraft wirkt auf mich sehr attraktiv.	1 2 3 4 5 6 7
Motivation	34. Es fällt mir überaus leicht mich für etwas zu motivieren.	1 2 3 4 5 6 7
Pflichtfach versus Interessensfach	35. Die Vorlesung ist eine Pflichtveranstaltung.	0 ja 1 nein

Abbildung 15 (Fortsetzung)[24]

[24] Die mit * gekennzeichneten Skalen müssen hier erhoben werden, wenn auf einen Eingangstest verzichtet wird.

Zum einen sollten möglichst klare und extreme Formulierungen in den verschiedenen Items Verwendung finden, sie sollten verhaltensnah und für die entsprechenden Skalen typisch bzw. charakteristisch sowie leicht verständlich (keine verneinten Fragen) und gut zu interpretieren sein. Extreme Formulierungen helfen den Urteilern die zu bewertenden Skalen zwischen den Extremen einzuordnen und unterstützen somit indirekt die Differenzierungsfähigkeit und Aussagekraft der Schätzung (Streuung der Antworten wird durch Extremformulierung der Fragestellung gefördert!!!). Weiter müssen die zur Beschreibung einer Skala verwendeten Items gleichgerichtet formuliert werden, dass heißt, wenn das Vorhandensein des zu messenden Merkmals einer Skala durch das erste Befragungsitem eine hohe Bewertung nach sich zieht, muss dies natürlich auch für die weiteren Items gelten. Als Antwortformat für die Items wird einheitlich eine siebenstufige Skala mit den Endpolen völlig unzutreffend (1) bis völlig zutreffend (7) vorgeschlagen. Eine Bewertungsskalenbreite bis zu maximal 7 hat sich in vergangenen Studien als optimal herausgestellt, einen breiteren Umfang der Bewertungsskala (über 7 Stufen) ist aufgrund fehlender Erfahrung und Differenzierungsfähigkeit seitens der Beurteiler nicht anzuraten.[25].

Es wird ausdrücklich darauf hingewiesen, dass dieser Vorschlag eines Fragebogens zur Ermittlung der Vortragsqualität sich ausschließlich auf die Unterrichtsform der Vorlesung bezieht, da ein Fragebogen zur Erfassung von anderen Unterrichtsformen (Seminare, Praktika,...) aufgrund der Verschiedenartigkeit der Veranstaltungen andere und mehr auf die Eigenart der jeweiligen Veranstaltung eingehende modulare Fragebogenteile benötigt und eine gewissenhafte Konstruktion solcher Zusatzmodule ein Wissen über die konkreten Ausformungen und Gestaltungen der Veranstaltung bedarf. Außerdem soll noch kritisch bemerkt werden, dass Evaluationen von Seminaren, die zum überwiegenden Teil Referatscharakter besitzen und somit durch studentische Beiträge überwiegend gekennzeichnet sind, bei bloßem Vorliegen einer Bewertung im Sinne eines Gesamteindrucks eher Gefahr laufen, die Qualität studentischer Referate in ihrer Gesamtheit zu beurteilen als dozentenbezogene Aspekte, die sich in diesem Fall auf die Leitung der anschließenden Diskussion, kritische Bemerkungen über die Darbietung und inhaltliche Anregungen und Verbesserungen beschränken.

Neben der didaktischen und sozialen Kompetenz ist wahrscheinlich die Fachkompetenz des Dozenten eine der bedeutendsten Voraussetzungen für das Zustandekommen qualitativ hochwertigen Unterrichts. Diese kann jedoch nicht durch die Studenten beurteilt werden. Wenn sie das könnten, besuchten sie aufgrund der Redundanz die falsche Veranstaltung. Die fachliche, inhaltliche Dimension muss wie bereits weiter oben erwähnt wurde, durch eine Peer-Evaluation bestätigt und deren aktueller Bezug von einer Expertengruppe gesichert werden.

Zur Generierung von neuen oder weiterer Items oder Skalen, die auf die jeweilige Veranstaltung zugeschnitten und den speziellen Verhältnissen des Vortrages angepasst sind, ist es zweckmäßig Lehrkräfte als auch Studenten über weitere Merkmale und Kennzeichen guten Unterrichts zu befragen. Diese Personen verfügen in der Regel über ein breites Wissen über die Hochschulpraxis. Außerdem können solche Befragungen zumeist direkte Verbesserungsvorschläge für die Lehre beinhalten, Anregungen zur Änderung bereits bestehender Fragebögen liefern und durch die gegenseitige Befragung beider teilnehmenden Gruppen, Vortragender als auch Rezipienten, Eigenschaften intersubjektiver Wichtigkeit und Bedeutung gefunden werden, die eine einseitige Betrachtung des Phänomens „guter“ Lehre

[25] Vgl. [7]

ausschließt und eine breitere Akzeptanz der so gefundenen Eigenschaften sichert. Eine Differenzierung der Fragestellung nach Kennzeichen einer guten Vorlesung und solcher eines guten Dozenten, müssen nicht unbedingt unterschieden werden, da in vergangenen Untersuchungen kaum nennenswerte Unterschiede in den genannten Merkmalen aufschienen. Vorwiegend wurden in beiden Fällen Merkmale der didaktischen Fertigkeit, Persönlichkeitsmerkmale, Themenwichtigkeit und Interaktionsbedingungen angeführt.[26]
Besonders wichtig erscheint bei Durchführung einer Erst- oder Vorversion des Fragebogens nicht nur die verschiedenen Skalen des Fragebogens beziehungsweise die unterschiedlichen Items bezüglich ihrer Testgütekriterien zu überprüfen[27] und auf die mitunter speziellen Verhältnisse anzupassen, sondern auch die Skalen sprich Dimensionen bezüglich der Relevanz für die Bewertung der Vortragsqualität einzuschätzen. Die gemeinsame Einschätzung sowohl von Professorenschaft als auch Studentenschaft, welche Aspekte als wichtig empfunden werden, soll Ausgangspunkt für eine quantitativ gestaffelte Zuordnung der verschiedenen Items zu den entsprechenden Skalen sein, um der gemeinsam vermuteten, unterschiedlichen Bedeutung auch quantitativ Wirkung zu verleihen. Es wird für die unterschiedliche Gewichtung der diversen Dimensionen eine Bewertung vorgeschlagen, deren Gehalt für die vermutete Qualität der Lehre ansteigend mit 1 (am unwichtigsten) bis 4 (am wichtigsten) nummeriert werden soll, dabei dürfen auch gleiche Gewichtungen für unterschiedliche Skalen mehrmals vergeben werden, um die Möglichkeit zu haben, Skalen auch gleiche Bedeutung zuzuerkennen. Für eine Bewertung der Skalen ist es unerlässlich, die Skalen (Lehrkompetenz, Strukturierung,...), wie auch weiter oben geschehen, in ihrer Art und in ihrem Wesen zu beschreiben, damit ein klareres Bild für die Beurteilung und Vergabe für Gewichtungen entsteht, was denn eigentlich mit ihr beschrieben und in weiterer Folge auch gemessen werden soll.
Eine Erhebung der Gewichtung der Dimensionen könnte mit folgender Tabelle geschehen:

	Dimension	**Gewichtung**
1	Persönlichkeit des Dozenten	2
2	Lehrkompetenz	4
3	Strukturierung des Lehrstoffes/der Veranstaltung	3
4	Anwendungsmöglichkeiten	2
i	Themeninteresse	g_i
⋮		⋮

Abbildung 16

Die so erhaltenen Skalengewichte werden über alle Urteiler gemittelt und gerundet. Eine Richtlinie für die zu verwendende Itemanzahl pro Skala kann nun nach der Vorschrift errechnet werden:

$$\frac{\overline{g}_i * N}{\sum \overline{g}_i}$$

Wobei die so erhaltene Zahl noch gerundet werden muss.
In dieser Formel bedeuten:
$\overline{g}_i$ = über alle Urteiler gemitteltes und gerundetes Gewicht der Skala i,

[26] Vgl. [9]
[27] siehe 3.4.2.2, 3.4.2.5

$\sum \bar{g}_i$ = Summe der gemittelten Gewichte über alle Skalen,
N = Ungefährer Umfang sprich gedachte Anzahl der im Evaluationsbogen gestellten Fragen.

Gleichung 28

Die tatsächliche Anzahl der im Evaluationsbogen schlussendlich befindlichen Fragen, können aufgrund der Rundungen der nach Gleichung 28 gefundenen Zahlen und der Biasitems noch leicht variieren. Außerdem soll noch einmal betont werden, dass es sich hierbei lediglich um eine Richtlinie, also um einen Vorschlag handelt, wie die Items auf die diversen Skalen zu verteilen sind. Die endgültige Verteilung kann von diesem Vorschlag durchaus abweichen, da nicht alle Dimensionen die vorgeschlagene Anzahl an Items zur klaren Identifizierung verlangen und der vielgerühmte Hausverstand sich nicht an straffe und starre Berechnungsmethoden binden soll, sondern diese lediglich als Hilfestellung zu sehen sind.
Die obere Grenze der Gewichtungsskala, hier von 1 bis 4 vorgeschlagen, wird durch die Anzahl der Itemfragen im Fragebogen und durch die Anzahl der Dimensionen bestimmt. Setzen wir in Gleichung 28 $\bar{g}_i$ = 1 so folgt:

$$\frac{N}{\sum \bar{g}_i} = \frac{N}{\mathrm{E}(\bar{g}_i) * D} \geq 1$$

wobei D die Anzahl der verschiedenen Dimensionen und $\mathrm{E}(\bar{g}_i)$ den Erwartungswert oder den Mittelwert der $\bar{g}_i$ über die Dimensionen bedeutet.

Gleichung 29

Gleichung 29 sichert auch der niedrigst gewichteten Dimension, dass dieser zumindest ein Befragungsitem zugesprochen werden kann.
Evaluationsfragebögen zur Vorlesungsqualität sollten aus Gründen der Ökonomie und Zumutbarkeit einen Umfangsbereich von nicht viel mehr als ca. 30-45 Einzelfragen haben. Wie bereits in anderen Untersuchungen und auch die Überlegungen hier zu den verschiedenen Dimensionalitäten der Unterrichtsbewertung zeigen, ergibt sich eine Größenordnung von ca. 8-15 für die diversen Skalen. Eine allzu geringe Anzahl der Dimensionalität birgt wie schon angedeutet die Gefahr einer „schwammigen“ Gesamtbeurteilung des Unterrichts, die nicht nur den Blick auf die verschiedenartigen Bereiche und die unterschiedliche Bewertung verschleiert, sondern auch ein entsprechendes Handeln in den als kritisch empfundenen Bereichen verhindert, da sie in der Regel nicht als solche erkannt werden.
Werden in Gleichung 29 die Werte eingesetzt die sich vorab hier in der Vorversion ergeben, nämlich D = 11 und eine angenommene Anzahl der Befragungsitems von N = 35 für die verschiedenen Dimensionen (Bias ist nicht in N berücksichtigt), so ergibt sich $\mathrm{E}(\bar{g}_i)$ ≤ 3.1, wenn obige Gleichung erfüllt sein soll. Wir runden den Wert zu 3 ab. Diese Zahl $\mathrm{E}(\bar{g}_i)$ = 3 würde dem Mittelwert einer von 1..5 skalierten, symmetrischen Verteilung entsprechen, die sich allgemein für eine Skala von 1.. g_{os} zu

$$\frac{g_{os}+1}{2}$$

berechnen lässt. g_{os} bezeichnet dabei die obere Schranke der Skala.

Gleichung 30

Da wir aber über die wirkliche Verteilung der Gewichtungen nichts wissen und diese in der Regel nicht symmetrisch sein wird, wird der Mittelwert der Gewichtungsverteilung $\mathrm{E}(\bar{g}_i) \neq \frac{g_{os}+1}{2}$, der mittleren Gewichtung, sein.
Jedoch wird der Unterschied bei nicht symmetrischer Verteilung der Gewichtungen umso kleiner ausfallen, je kleiner die Skalenbreite und damit g_{os} gewählt wurde. Wir setzten deshalb die obere Schranke mit 4 fest.
Man könnte nun freilich dagegen halten, dass mit einer Skalierung der Gewichtung von lediglich 1 bis 4 nur eine ungenügende Differenzierung der Dimensionalitäten erfolgen kann und fast allen Dimensionen daher annähernd dieselbe Itemanzahl zugeordnet werden muss. Dem lässt sich allerdings entgegnen, dass die Wahl einer breiteren Skalierung den Effekt einer Gleichverteilung der Befragungsitems auf die Dimensionen von vornherein nicht verhindert (extreme Gewichtungen –sehr niedrige oder sehr hohe- werden vielmals „ungerechtfertigter Weise“ überhaupt nicht vergeben, Tendenz zum Mittelmaß der Bewertung) und weniger wichtig empfundene Dimensionen bei breiterer Skalierung eher Gefahr laufen bei gleichbleibender Länge des Fragebogens kein Befragungsitem zu erlangen und daher keine Möglichkeit mehr haben ihre wirkliche Relevanz oder nicht Relevanz empirisch zu belegen.

3.4.2 Richtlinien und Methodik der Auswertung von Evaluationsfragebögen:

Bei studentischen Lehrevaluationen von Vorlesungen an Universitäten liegen pro Dozent immer mehrere Beurteilungen von den Hörern vor. Dabei werden die verschiedenartigen Skalen sowohl innerhalb einer Veranstaltung als auch über die unterschiedlichen Veranstaltungen von den Studierenden verschieden bewertet. Dies liegt zum einen in der Natur der Sache , dass die einen eher kritisch die anderen günstiger urteilen, zum anderen in der vielleicht auch unterschiedlichen Performance in den verschiedenen Veranstaltungen. Ersteres spricht eher gegen die Fähigkeit der Hörer, ein den Dozenten zutreffend beschreibendes Urteil abzugeben, zweiteres würde unter Umständen für eine solche Fähigkeit sprechen, da sie Leistungsschwankungen in den diversen Unterrichtseinheiten aufzeigen könnte und somit sich die Lehrevaluation als ein sensibles, differenzierungsfähiges Instrument präsentiert. Es müssen also um die Aussagekraft von Lehrevaluationen mit der richtigen Relevanz und im richtigen Licht zu sehen, Untersuchungen und Überlegungen über die studentische Homogenität der Urteilsfindung angestellt werden.
Die Daten die in einem Evaluationsfragebogen über mehrere Veranstaltungen erhoben werden, haben im allgemeinen folgende Form und unten beispielhaft dargestellte, unterschiedliche Messniveaus bzw. Aggregationsniveaus.

			Urteiler 1,1	Urteiler 1,2		...Urteiler 1t	
Veranstaltung 1	Skala 1	Item 1	3	2		5	
		Item 2	5	2		4	
		...Item si	4	5		4	
	Skala 2	Item 1	5	4		3	
		Item 2	5	3		4	
		...Item si	4	1		3	
	Skala s...	Item 1	2	2		1	
		Item 2	4	3		4	
		...Item si	5	5		4	
	Urteilermittel:		4,1111111	3		3,5555556	
			Urteiler 2,1	Urteiler 2,2			...Urteiler 2t
Veranstaltung 2	Skala 1	Item 1	6	9			5
		Item 2	7	6			2
		...Item si	8	8			4
	Skala 2	Item 1	6	8			5
		Item 2	7	7			3
		...Item si	5	6			2
	Skala s...	Item 1	8	9			5
		Item 2	5	8			5
		...Item si	4	6			2
	Urteilermittel:		6,2222222	7,4444444			3,6666667

Veranstaltung r...

Abbildung 17 Darstellung der Auswertungsmatrix für Urteilerübereinstimmung, Interraterreliabilität und Skalenhomogenität eines Dozenten.

Die Auswertungen der Aussagekraft studentischer Urteile erfolgt also über eine im allgemeinen nicht quadratische Matrix. Die Matrix enthält Daten aus **r** Veranstaltungen eines Dozenten, die von **rt** Studenten pro Veranstaltung beurteilt wurden. Beurteilt werden **s** Dimensionen, die wiederum durch **is** Items beschrieben werden.

3.4.2.1. Berechnung der Interraterreliabilität und der Urteilerübereinstimmung über die Interraterreliabilität:

Betrachten wir zunächst die Berechnung der Urteilerübereinstimmung von 2 Urteilern über mehrere (3) Veranstaltungen.

			Urteiler 1,1	Urteiler 1,2
Veranst. 1	S 1	Item 1	3	5
		Item 2	5	3
		Item 3	4	5
	S 2	Item 1	2	2
		Item 2	2	1
	Urteilermittel:		3,2	3,2
			Urteiler 2,1	Urteiler 2,2
Veranst. 2	S 1	Item 1	6	5
		Item 2	5	6
		Item 3	7	5
	S 2	Item 1	4	3
		Item 2	5	5
	Urteilermittel:		5,4	4,8
			Urteiler 3,1	Urteiler 3,2
Veranst. 3	S 1	Item 1	6	5
		Item 2	5	4
		Item 3	6	5
	S 2	Item 1	4	3
		Item 2	3	3
	Urteilermittel:		4,8	4

Abbildung 18

Würde man die Urteilerübereinstimmung als Korrelation über die 2x15 Werte berechnen, so ergäbe sich ein „verfälschtes“ Maß der Übereinstimmung. Urteilermittelwerte der verschiedenen Veranstaltungen unterscheiden sich in aller Regel. Nicht nur, dass der Mittelwert Veranstaltungsspezifika festhält (die Leistungen können in Veranstaltung 1 besser gewesen sein als z.B. in Veranstaltung 2), sondern auch die Urteilsfähigkeit der Urteiler kann unterschiedlich sein (z.B. tagesverfassungsabhängig). Außerdem wird in den seltensten Fällen der Urteiler 1,1 die gleiche Person wie Urteiler 2,1 usw. sein und somit wird in der Regel ein anderer Beurteilungsmaßstab der einzelnen Studenten für auch gleiche Leistungen zum Einsatz kommen.[28] Es kann also nicht zweifelsfrei entschieden werden, ob der Unterschied im Urteilermittel1,1 im Vergleich zum Urteilermittel2,1 auf Unterschiedlichkeiten im Beurteilungsmaßstab oder in einem wirklichen qualitativen Unterschied der 2 Veranstaltungen zurückzuführen ist.
Um solche Inhomogenität sowohl der Beurteilungsstichprobe als auch der verschiedenen Veranstaltungen heraus zu filtern, müssen die Itemwerte bezüglich der Urteilermittelwerte residualisiert werden.
Eine Berechnung der Korrelation über die diversen Skalengrößen (Itemwerte einer Skala werden durch den entsprechenden Skalenmittelwert ersetzt z.B. die drei Itemwerte für Urteiler 1,1 der Skala 1 [5,5,4] in Veranstaltung 1 wird durch den Skalenmittelwert 4.6 von Skala 1 ersetzt [4.6,4.6,4.6])[29] wird aufgrund des Informationsverlustes der Kovariation der verschiedenen Itemgrößen in den entsprechenden Skalen von vornherein nicht erwogen, außer Umfang der Daten und Kapazitätsengpässe der Datenverarbeitung verlangen ein solches Vorgehen. Solcher Art berechnete Interraterreliabilitäten oder Urteilerübereinstimmungen (über die Skalenmittelwerte berechnet) weisen in der Regel größere Werte auf als die über die Rohdaten, da die Skalen oft treffender sprich übereinstimmender als Itemdaten beurteilt werden.

[28] Aus diesem Grund ist es auch nicht statthaft absolute Übereinstimmungen zu berechnen, da absolute Unterschiede in der Bewertungsskala von verschiedenen Urteilern keine Aussagekraft für sich beanspruchen können, sondern nur die gemeinsamen relativen Abweichungen vom jeweiligen Urteilermittelwert.
[29] Vgl. Abbildung 19.

Interraterreliabilitäten und Urteilerübereinstimmungen sollen über jene Skalen ermittelt werden, welche auch „objektiv“ beobachtbar und beurteilt werden können. Eine Berechnung der Urteilerübereinstimmung über Skalen, die sich größtenteils durch intersubjektive Unterschiedlichkeiten innerhalb der Studentenschaft auszeichnen, spiegeln nur die Gegensätzlichkeit dieser Merkmale innerhalb der Stichprobe wider und drücken nicht die Homogenität der Urteilsfindung der Studenten aus. So wäre es nicht sinnvoll Skalen überwiegend persönlicher Natur wie z.B. Fleiß, Belastbarkeit usw. in die Urteilerübereinstimmung mit einzubeziehen, da in der Regel von einer Verschiedenartigkeit dieser Merkmale innerhalb der Studentenschaft ausgegangen werden kann und diese nicht die Fähigkeit der Studentenschaft einer gemeinsamen, homogenen Beurteilung beobachtbarer Verhältnisse wiedergibt. Für die Berechnung der Interraterreliabilität und Urteilerübereinstimmung sollten somit nur **Daten aus den Skalen** verwendet werden, die unter 3.4.1 zur Beschreibung der Oberkategorien **Dozent** und **Rahmenbedingungen** erhoben werden.
Es soll grundsätzlich innerhalb und über die verschiedenen Veranstaltungen hinweg von einer **Homogenität der Varianz** bezüglich der verschiedenen Urteiler (Varianz von Urteiler 1,1 = Varianz von Urteiler 1,2 = Varianz von Urteiler 2,1,...) als auch von einer Varianzhomogenität bezüglich der Items innerhalb gleichartiger Skalen über die Veranstaltungen hinweg (Varianz von Item1 = Varianz von Item2 = Varianz von Item3 für Skala s1 in Veranstaltung 1 und Veranstaltung 2,...) ausgegangen werden können.[30] Eine Residualisierung der Daten bezüglich der s Skalen stellt ebenfalls eine nicht „korrekte“ Vorgehensweise dar, da die Übereinstimmung bzw. Kovariation über die entsprechenden Skalen hinweg ebenfalls entscheidend für die Gleichartigkeit der Urteilsfindung der 2 Urteiler ist und daher in die Berechnung der Urteilerübereinstimmung mit einfließen müssen. Außerdem würden Skalen, die nur durch eine Itemgröße beschrieben werden, durch eine Residualisierung bezüglich der Skala s automatisch zu 0 werden, daher keinen Beitrag zur Kovariation liefern und damit einen Informationsverlust bedeuten.
Wir berechnen die Interraterreliabilität in Analogie zu Gleichung 23 und in Übereinstimmung zum oben Gesagten, indem die residualisierten Itemdaten (Rohdaten) zur Berechnung herangezogen werden. Zur Veranschaulichung soll folgendes Beispiel dienen.

[30] Vgl. Abbildung 19.

			Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	
Veranst. 1	S 1	Item 1	5	4	6	
		Item 2	5	4	7	
		Item 3	4	3	5	
	S 2	Item 1	2	1	3	
		Item 2	2	2	4	
Urteilermittel:			3,6	2,8	5	
			Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4
Veranst. 2	S 1	Item 1	6	3	7	4
		Item 2	5	2	4	2
		Item 3	7	5	7	6
	S 2	Item 1	4	3	5	2
		Item 2	5	5	5	4
Urteilermittel:			5,4	3,6	5,6	3,6

								$\sum_i \left(\frac{A_i^2}{n_i} \right)$
			Urteiler 1,1	Urteiler 1,2	Urteiler 1,3			
Veranst. 1	S 1	Item 1	1,4	1,2	1			4,32
		Item 2	1,4	1,2	2			7,053333333
		Item 3	0,4	0,2	0			0,12
	S 2	Item 1	-1,6	-1,8	-2			9,72
		Item 2	-1,6	-0,8	-1			3,853333333
			Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4		
Veranst. 2	S 1	Item 1	0,6	-0,6	1,4	0,4		0,81
		Item 2	-0,4	-1,6	-1,6	-1,6		6,76
		Item 3	1,6	1,4	1,4	2,4		11,56
	S 2	Item 1	-1,4	-0,6	-0,6	-1,6		4,41
		Item 2	-0,4	1,4	-0,6	0,4		0,16

Abbildung 19 Beispiel zur Ermittlung der Interraterreliabilität und Urteilerübereinstimmung.

Obiges Beispiel enthält eine von der Form her typische Datenmatrix mit ungleicher Urteileranzahl über die verschiedenen Veranstaltungen.
Aufgrund der Charakteristik der Interraterreliabilität, die Interraterreliabilität stellt die Zuverlässigkeit der Beurteilerstichprobe einer definierten Größe dar, gibt es wenig Sinn den Konsistenzkoeffizienten über die verschiedenen Veranstaltungen hinweg mit verschieden großen Urteilerstichproben zu berechnen, da ja die Interraterreliabilität für unterschiedliche Beurteilungsstichprobengrößen unterschiedliche Werte annimmt und uns ja eigentlich die Interraterreliabilität für eine bestimmte Größe der Beurteilungsstichprobe z.B. 10 Beurteiler ($Irr_{(10)}$) interessiert. Wir berechnen also die Interraterreliabilitäten **getrennt für jede Veranstaltung**.
Da die Berechnung des Konsistenzkoeffizienten eine varianzanalytische Fortführung der Testhalbierungsmethode bedeutet, als deren Verallgemeinerung sie anzusehen ist, bedeutet die so errechnete Interraterreliabilität die **Zuverlässigkeit** der über die **Beurteilungsstichprobe** bestimmter Größe **gemittelten Itemeinschätzungen** für diese Veranstaltung. Sie ist ein Grad dafür, wie verlässlich die Beurteilungsmittel der Items einer bestimmten Stichprobengröße für eine bestimmte Veranstaltung eingeschätzt werden kann.
Die Interraterreliabilität nimmt mit zunehmender Größe der Stichprobe zu, dass heißt mit zunehmender Beurteilerzahl steigt der Grad der Objektivität der Itemeinschätzungen und damit auch der Veranstaltungseinschätzungen. Da in Evaluierungen die gemittelte Veranstaltungsbewertung die relevante Größe darstellt, bildet die Interraterreliabilität über den Konsistenzkoeffizienten das Ausschlag gebende Maß. Für eine ausreichende Einschätzung der Beurteilungsreliabilität wird bei Rindermann eine minimale Beurteilungsstichprobe der Größe von 10 bzw. 20 Urteilern pro Veranstaltung angegeben.[31] Bei besser besuchten Veranstaltungen nimmt die Repräsentativität der Einschätzungen der Kursteilnehmer weiter zu.
Zur Berechnung der Interraterreliabilität müssen folgende Schritte und Hilfsgrößen berechnet werden. Zuerst wird die obige Datenmatrix bezüglich der Urteilermittel

[31] Vgl.[9].

residualisiert und wir erhalten die in Abbildung 19 unten dargestellte Berechnungsmatrix.
Weiters dienen folgende Kennziffern zur Berechnung:

$$(1) = \frac{G^2}{N}$$ [32]

$$(2) = \sum_{i=1}^{p} \sum_{m=1}^{n_i} x^2{}_{im}$$

$$(3) = \sum_{i}^{p} \left(\frac{A_i{}^2}{n_i} \right)$$

$$p = \sum_{s} is$$

$$N = \sum_{i}^{p} n_i = rt * \sum_{s} is$$

In dieser Gleichung bedeuten:
G = die Summe über alle N Werte einer Veranstaltung,
A_i = die Zeilensumme über die in der Zeile i befindlichen n_i Werte der Veranstaltung,
p = die Anzahl der Zeilen der Matrix erhoben in einer Veranstaltung (in unserem Beispiel 5),
rt bedeutet die Anzahl der Urteiler in der Veranstaltung r,
is = die Anzahl der Items, die zur Beschreibung der Dimension s verwendet wird und schließlich bedeutet Ausdruck (2) die Summe über alle quadrierten Werte einer Veranstaltung.

Gleichung 31

Mit diesen Kennziffern kann nun die über die Veranstaltungen getrennt erfassten Interraterreliabilitäten nach folgender Art berechnet werden.

	Quadratsumme (QuS):	Freiheitsgrade (df):	Varianz:
Zw._den_Items:	(3)-(1)	$\left(\sum_{s} is - 1 \right)$	$\frac{Qus_{Zw\_den\_Items}}{df_{Zw\_den\_Items}}$
Fehler:	(2)-(3)	$\left(\sum_{s} is - 1 \right) * (rt - 1)$	$\frac{Qus_{Fehler}}{df_{Fehler}}$
Total:	(2)-(1)	$\left(\sum_{s} is - 1 \right) * rt$	$\frac{Qus_{Total}}{df_{Total}}$

Abbildung 20

Wir berechnen die Interraterreliabilitäten in Analogie zur Berechnung des Konsistenzkoeffizienten unter 3.3.2 nach Gleichung 23, wobei die Restvarianz nun der hier bezeichneten Fehlervarianz entspricht und die Varianz zwischen den Probanden der Varianz zwischen den Items hier.
Wir erhalten für unser Beispiel eine Interraterreliabilität für Veranstaltung 1 abgerundet zu $Irr_{(3)}$ = 0,98 und für Veranstaltung 2 abgerundet zu $Irr_{(4)}$ = 0,90. Die Werte in Klammern bedeuten die Urteileranzahl der jeweiligen Veranstaltung.
Die Urteilerübereinstimmungen, für jede Veranstaltung getrennt berechnet, ergeben sich jeweils aufgerundet zu $Üb_{[1]}$ = 0,96 und $Üb_{[2]}$ = 0,70.

[32] Ausdruck (1) berechnet sich aufgrund der residualisierten Daten zu 0.

Diese erhalten wir, indem wir die Korrelationen über alle 3 Kombinationen der 3 Urteiler ([1,2],[1,3],[2,3]) für Veranstaltung 1 berechnen und anschließend mitteln. Die gleiche Vorgehensweise für Veranstaltung 2 über alle 6 Kombinationen führt zu $Üb_{[2]}$. Werden diese getrennt berechneten Urteilerübereinstimmungen entsprechend ihrer Urteileranzahl gewichtet und gemittelt folgt die totale Urteilerübereinstimmung über die 2 Veranstaltungen aufgerundet zu:

$$\frac{\sum_{r=1}^{2} Üb_{[r]} * rt}{\sum_{r=1}^{2} rt} = \frac{Üb_{[1]} * 3 + Üb_{[2]} * 4}{7} = 0{,}81 = Üb_{total}$$

$Üb_{[1]}$ = Urteilerübereinstimmung für Veranstaltung 1,
$Üb_{[2]}$ = Urteilerübereinstimmung für Veranstaltung 2.

Gleichung 32

Die getrennt berechneten Urteilerübereinstimmungen müssen entsprechend ihrer Urteileranzahl gewichtet werden, da die Urteilerübereinstimmung für größere rt zufällige Übereinstimmungen eher ausschließt und repräsentativer für die wahre Urteilerübereinstimmung ist.
Diese Art der Berechnung ist allerdings für größere Urteileranzahl und mehrere Veranstaltungen nicht zumutbar.
Wir berechnen daher die Urteilerübereinstimmungen für die jeweilige Veranstaltung näherungsweise über die Interraterreliabilitäten. Werden die Interraterreliabilitäten mittels der **Spearman-Brown** Beziehung nach Gleichung 27 und für Veranstaltung 1 n'/n = 1/3 und für Veranstaltung 2 n'/n = 1/4 entsprechend ihrer Urteileranzahl eingesetzt ergeben sich die Urteilerübereinstimmungen näherungsweise zu $Üb_{[1]}$ = 0,95 aufgerundet und $Üb_{[2]}$ = 0,69 abgerundet. [33]
Die Ungenauigkeiten der Schätzungen der Urteilerübereinstimmung für Veranstaltung 1 und 2 über die Interraterreliabilitäten ist auf die Verletzung der Varianzhomogenität der Urteiler zurückzuführen. Würden die Daten zuvor noch normalisiert werden (residualisierte Daten werden noch durch die Standardabweichung der verschiedenen Urteiler dividiert), ergäben sich die exakt gleichen Urteilerübereinstimmungen $Üb_{[1]}$ und $Üb_{[2]}$ nach der unter Gleichung 32 beschriebenen Berechnungsvorschrift.
Das Freiheitsgradverhältnis $df_{zw\_den\_Items}/df_{fehler}$ = 1/(rt-1) wäre für normalisierte Größen das gleiche wie für die residualisierten Größen, da die Normalisierung der Daten jeweils nur den Faktor $\left(\sum_{s} is - 1\right)$ in Abbildung 20 um 1 vermindert.

Es müssten also zur exakten Berechnung der Urteilerübereinstimmungen nur die veränderten, „normalisierten" Quadratsummengrößen der jeweiligen Veranstaltung berechnet werden, diese dann entsprechend der Berechnungsvorschrift für die Interraterreliabilität nach Abbildung 20 und Gleichung 23 eingesetzt werden, um schlussendlich über die Gleichung 27 die exakten Urteilerübereinstimmungen für Veranstaltung 1 und 2 zu erhalten.
Eine gute Näherung für die totale Urteilerübereinstimmung nach Gleichung 32 über alle Veranstaltungen hinweg, erhalten wir auch wenn wir zur Berechnung die gesamte residualisierte Datenmatrix über alle Veranstaltungen heranziehen. Berechnen wir den Konsistenzkoeffizienten nach obiger Vorschrift mit den nun im Vergleich zu Gleichung 31 veränderten Kennziffern:

[33] Vgl. dazu die Ausführungen zur Berechnung der mittleren Aufgabenkorrelation aus dem Konsistenzkoeffizienten unter Punkt 3.3.2.

$$p = r * \sum_{s} is$$

$$N = \sum_{i}^{p} n_i = \sum_{r} rt * \sum_{s} is$$

wobei r = die Anzahl der Veranstaltungen und G in Kennziffer 1 nach Gleichung 31 die Summe über alle N Werte der gesamten Datenmatrix hinweg bedeutet

Gleichung 33

und den veränderten Freiheitsgraden,

	Freiheitsgrade (df):
Zw._den_Items:	$\left(\sum_{s} is - 1\right) * r$
Fehler:	$\left(\sum_{s} is - 1\right) * \left(\sum_{r} rt - r\right)$
Total:	$\left(\sum_{s} is - 1\right) * \sum_{r} rt$

Abbildung 21

so ergibt sich eine „Interraterreliabilität“ von $Irr_{(3,5)}$ = 0,93 abgerundet. Der Wert in Klammer bedeutet die über die 2 Veranstaltungen gemittelte Urteileranzahl.
Alle anderen Kennziffern in Gleichung 31 und Abbildung 20 bleiben formal gleich.
Bilden wir nun die Urteilerübereinstimmung über die Gleichung 27 mit n’/n = 1/3,5 so folgt $Üb_{total}$ = 0,80 abgerundet.
Die Interraterreliabilität $Irr_{(3,5)}$ ist in diesem Zusammenhang nur eine Berechnungshilfsgröße und ist wenn überhaupt nur schwer zu interpretieren. Die Urteilerübereinstimmung nach Gleichung 33 bildet eine gute Näherung zur Urteilerübereinstimmung nach Gleichung 32.
Aufgrund der unterschiedlichen Urteileranzahl (Veranstaltung 1: 3, Veranstaltung 2:4) würde auch eine Normalisierung der Daten nicht zum Ergebnis nach Gleichung 32 führen.
Der Gewinn an Genauigkeit der Urteilerübereinstimmungen für die verschiedenen Veranstaltungen berechnet über **normalisierte** Daten im Vergleich zu **residualisierten** Daten wird aufgrund unserer Annahme der Varianzhomogenität nichtig im Vergleich zum Aufwand (es müssen zusätzlich zum Mittelwert auch noch die Standardabweichungen jedes Urteilers berechnet werden) und erscheint daher nicht praktikabel.
Unterschiedlichkeiten des Bewertungsmaßstabes innerhalb der Beurteilerstichprobe einer Veranstaltung fließen, wie oben bereits angedeutet, durch die vorher stattfindende Residualisierung nicht in die Interraterreliabilitäts- und damit Urteilerübereinstimmungsberechnung mit ein.
Urteilerübereinstimmungen spiegeln in gewisser Weise die **Zuverlässigkeit** oder Objektivität **individueller Urteile** wider, sie geben an, wie gut man von einem Urteil auf das eines anderen schließen kann. Die unter 3.4.2.1 dargestellten Umrechnungen von Interraterreliabilität zu Urteilerübereinstimmungen haben natürlich nur für Interraterreliabilitäten ≥ 0 Gültigkeit. **Interraterreliabilitäten < 0 werden null gesetzt.**
Eine getrennte Untersuchung der totalen Urteilerübereinstimmung für die diversen Skalen (z. B. die totale Übereinstimmung der Beurteilung nur für Skala 1), die totale Skalenübereinstimmungen, wird nicht erwogen. Zwar kann mit großer

Wahrscheinlichkeit davon ausgegangen werden, dass Unterschiede in der Urteilerübereinstimmung in den verschiedenen Skalen bestehen. So werden zum Beispiel in Skalen, die Verhaltensweisen des Dozenten oder der Veranstaltung beschreiben, sprich Skalen, die sich auf konkret beobachtbares Verhalten beziehen, höhere Urteilerübereinstimmungen zu erwarten sein als in Skalen, die mehrheitlich eine Einschätzung von Einstellungen oder Beurteilungen von Persönlichkeitseigenschaften vornehmen. Eine Analyse der totalen „Skalenübereinstimmungen" getrennt für die verschiedenen Skalen wird aber aus Gründen des damit verbundenen Aufwandes nicht empfohlen. Es müssten vorerst alle Urteilerübereinstimmungen über alle Skalen für Veranstaltung r (Skala 1...Skala s) berechnet werden, was eine vorhergehende Residualisierung der Skaladaten bezüglich der Skalenmittelwerte bedeuten würde Es müsste für Veranstaltung r die s*rt Skalenmittelwerte berechnet werden, wobei s die Anzahl der beschriebenen Skalen darstellt. Die totale „Skalenübereinstimmung" für Skala s würde sich dann wieder analog über alle r Veranstaltungen nach Gleichung 32 berechnen. Außerdem könnten Urteilerübereinstimmungen für Skalen, die nur durch ein Item beschrieben werden, nicht berechnet werden und Übereinstimmungen über die entsprechenden Skalen hinweg werden in der Skalenübereinstimmung durch die methodisch über die Skalen getrennte Berechnungsweise nicht berücksichtigt.
Es soll zum Schluss dieses Kapitels nochmals nachdrücklich darauf hingewiesen werden, dass Veranstaltungsbewertungen durch Studenten nur als Interraterreliabilitätsmaß sprich gemittelte Bewertung über alle beurteilenden Studenten Aussagekraft für sich in Anspruch nehmen können, da Urteilerübereinstimmungen, wie auch andere Untersuchungen schon belegt haben, über eine zu geringe mittlere Korrelation (Üb < 0,5) untereinander aufweisen als, dass man Urteile eines einzelnen Studenten als Beurteilungsgröße heranziehen dürfte.

3.4.2.2. Testgütekriterien für Skalen:

Die Skala s wird durch is Items erhoben.
Es wird angenommen, dass das Charakteristikum der Skalen durch die verschiedenen Items, die Aspekte des Merkmals nach dem Prinzip der inhaltlichen Ähnlichkeit ohne inhaltliche Redundanz und nicht nach dem Prinzip der inhaltlichen Ergänzung abbilden, erklärt wird.
Somit bildet eine hohe mittlere Korrelation zwischen diesen Items eine Grundvoraussetzung für eine adäquate Beschreibung und Erfassung des Skalenmerkmals. In Analogie zur Urteilerübereinstimmung vs. Interraterreliabilitäten, bildet die **Skalenhomogenität** das Pendant zur Interraterreliabilität, während die Urteilerübereinstimmung hier der mittleren Itemkorrelation entspricht. Wie im vorigen Kapitel für die Interraterreliabilität steigt der Grad der Zuverlässigkeit für eine Skala, die Skalenhomogenität, mit steigender Itemzahl. Hieraus lässt sich auch der in 3.4.1 geäußerte Wunsch verstehen, der gemeinsam eingeschätzten und unterschiedlich vermuteten Bedeutung der Skalen zur Erhebung der Vortragsqualität mit einer quantitativen Staffelung zu begegnen. Denn je wichtiger und bedeutender eine Skala für die Erfassung der Veranstaltungsqualität eingeschätzt wurde, umso höher soll auch der Grad der Zuverlässigkeit dieser Skala ausfallen, umso verlässlicher sollen auch die Skalenmittelwerte geschätzt werden. Wir berechnen wieder in Analogie zur Interraterreliabilität und Urteilerübereinstimmung die Skalenhomogenität und mittlere Itemkorrelation der Skalen, indem wieder zuerst die verschiedenen Items der Skala s bezüglich des Itemmittelwertes in den jeweiligen Veranstaltungen residualisiert werden. Damit werden Unterschiede in den Items (Items können verschieden hoch

bewertet werden) und Unterschiede in den Beurteilungsstichproben in den verschiedenen Veranstaltungen aus den Daten herausgefiltert.
Zur Veranschaulichung berechnen wir in unserem Beispiel die Skalenhomogenität und mittlere Itemkorrelation exemplarisch für Skala 1.

			Urteiler 1,1	Urteiler 1,2	Urteiler 1,3		Itemmittelwert:
Veranst. 1	S 1	Item 1	5	4	6		5
		Item 2	5	4	7		5,333333333
		Item 3	4	3	5		4
			Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4	Itemmittelwert:
Veranst. 2	S 1	Item 1	6	3	7	4	5
		Item 2	5	2	4	2	3,25
		Item 3	7	5	7	6	6,25

			Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	
Veranst. 1	S 1	Item 1	0	-1	1	
		Item 2	-0,33333333	-1,33333333	1,666666667	
		Item 3	0	-1	1	
			0,037037037	3,7037037	4,481481481	
			Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4
Veranst. 2	S 1	Item 1	1	-2	2	-1
		Item 2	1,75	-1,25	0,75	-1,25
		Item 3	0,75	-1,25	0,75	-0,25
			4,083333333	6,75	4,083333333	2,08333333

Abbildung 22

Wir berechnen die Skalenhomogenität für die Skala 1 zuerst getrennt für die Veranstaltungen 1 und 2.
Folgende Kennziffern werden zur Berechnung benötigt:

$$(1) = \frac{G^2}{N}$$ [34]

$$(2) = \sum_{i=1}^{p} \sum_{m=1}^{rt} x^2{}_{im}$$

$$(3) = \sum_{i}^{rt} \left(\frac{A_i{}^2}{p} \right)$$

$$N = rt * p$$

In dieser Gleichung bedeuten:
G = die Summe über alle N Werte der Skala s in der Veranstaltung r,
A_i = die Spaltensumme über die in der Spalte i befindlichen p Werte der Skala s in der Veranstaltung r,
p = is die Anzahl der Zeilen (Items), die zur Beschreibung der Skala s dienen (in unserem Beispiel für Skala 1 p = 3),
rt die Anzahl der Urteiler in der Veranstaltung r.

[34] G ist aufgrund der Residualisierung der Itemdaten gleich 0.

Gleichung 34

Die spaltenweise gebildeten Ausdrücke $\frac{A_i^{\,2}}{p}$ sind in Abbildung 22 gelb markiert.

Die Skalenhomogenitäten berechnen sich nach Abbildung 23

	Quadratsumme (QuS):	Freiheitsgrade (df):	Varianz:
Zw._den_Urteilern:	(3)-(1)	$(rt-1)$	$\frac{Qus_{Zw\_den\_Urteilern}}{df_{Zw\_den\_Urteilern}}$
Fehler:	(2)-(3)	$(rt-1)*(is-1)$	$\frac{Qus_{Fehler}}{df_{Fehler}}$
Total:	(2)-(1)	$(rt-1)*is$	$\frac{Qus_{Total}}{df_{Total}}$

Abbildung 23

und Gleichung 23, wobei wieder der Restvarianz hier die Fehlervarianz entspricht. Wir erhalten für Veranstaltung 1 eine Skalenhomogenität von $Skh_{1,(3)} = 0{,}97$ abgerundet und für Veranstaltung 2 $Skh_{2,(3)} = 0{,}93$ aufgerundet. In Klammer steht die Anzahl der Items, die zur Beschreibung der Skala verwendet wurden. Aus diesen Werten folgt die mittlere Itemkorrelation der Skala 1 über die Gleichung 27 (für n'/n = 1/3 entsprechend der 3 Items in Skala 1) für Veranstaltung 1 zu $ItKo_{[1]} = 0{,}92$ abgerundet und $ItKo_{[2]} = 0{,}81$ aufgerundet. Auf Grund der angenommenen Varianzhomogenität bezüglich der Items über die Veranstaltungen hinweg, vereinfacht sich die Berechnung der totalen mittleren Itemkorrelation über die 2 Veranstaltungen aufgerundet zu:

$$\frac{\sum_r rt * ItKo_{[r]}}{\sum_r rt} = \frac{3*0{,}92+4*0{,}81}{7} = 0{,}86 = ItKo_{[tot]}.$$

Gleichung 35

Die totale Skalenhomogenität berechnet sich über die totale mittlere Itemkorrelation durch „Testverlängerung" über Gleichung 27 für n'/n = 3, wobei 3 die Anzahl der Items, die zur Beschreibung der Skala 1 verwendet wurden, bedeutet, zu aufgerundet $Skh_{tot,(3)} = 0{,}95$.

Es ist ebenfalls möglich die totale Skalenhomogenität und totale mittlere Itemkorrelation direkt ,ohne zuvor die Berechnungen für die einzelnen Veranstaltungen durchzuführen, zu erhalten. Dazu müssen nur die residualisierten Daten der verschiedenen Veranstaltungen für die interessierende Skala **hintereinander** angeordnet werden und mit dieser neu entstandenen Matrix der Konsistenzkoeffizient nach Gleichung 23 berechnet werden. Zusätzlich müssen noch die Freiheitsgrade in Abbildung 23 der veränderten Matrix angepasst werden, indem der Ausdruck (rt-1) durch $\left(\sum_r rt - 1\right)$ ersetzt wird. Es ergeben sich somit für unser Beispiel die neuen Freiheitsgrade $df_{Zw\_den\_Urteilern} = 6$ und $df_{fehler} = 12$ und die entsprechenden Quadratsummen berechnen sich nun über alle Werte aller r Veranstaltungen (in unserem Beispiel r = 2).

0	-1	1	1	-2	2	-1
-0,33333333	-1,33333333	1,66666667	1,75	-1,25	0,75	-1,25
0	-1	1	0,75	-1,25	0,75	-0,25
0,03703704	3,7037037	4,48148148	4,08333333	6,75	4,083333333	2,08333333

Veranst.1 Veranst.2

Abbildung 24

Eine Anordnungen der Daten der verschiedenen Veranstaltungen untereinander und anschließende Berechnung des Konsistenzkoeffizienten wäre nicht zulässig, da dann die Kovariation nicht nur von den Itemformulierungen abhängt, sondern auch durch die verschiedenen Urteiler beeinflusst wird, die im Regelfall nicht die gleichen Personen über die Veranstaltungen darstellen (Urteiler 1,1 ≠ Urteiler 2,1) und somit über einen anderen Beurteilungsmaßstab verfügen.
Man berechnet auf diese oben beschriebene Weise eine totale Skalenhomogenität von $Skh_{total,(3)} = 0,94$ und eine mittlere Itemkorrelation von $ItKo_{[total]} = 0,84$ jeweils abgerundet. Die Unterschiede zu den oben genannten Werten $ItKo_{[total]} = 0,86$ und $Skh_{total,(3)} = 0,95$ sind auf Verletzungen der Varianzhomogenität zwischen den Items zurückzuführen. Ansonsten würden beide Berechnungen zu den gleichen Werten $ItKo_{[total]} = 0,93$ abgerundet und $Skh_{total,(3)} = 0,98$ aufgerundet führen.
Ein Nachteil der Berechnung der Skalenhomogenität in dieser beschriebenen Weise über den Konsistenzkoeffizienten, liegt in der einfachen Tatsache, dass für Skalen, die nur durch ein Item beschrieben werden, keine Skalenhomogenität und mittlere Itemkorrelation berechnet werden können. Hier müsste die Zuverlässigkeit bzw. die Messgenauigkeit der „Skala“ über Retestreliabilitäten bestimmt werden, die jedoch zumeist unter den erzielten Konsistenzkoeffizienten liegen, da bei „Testwiederholungen“ Fehleranteile aufgrund von Merkmalsänderungen und verschiedenen Durchführungsbedingungen nicht ausgeschlossen werden können. Eine direkte Vergleichbarkeit mit den Konsistenzkoeffizienten somit nicht gegeben ist.

3.4.2.3. Stabilität der studentischen Beurteilung dozentenbezogener Skalen über verschiedene Veranstaltungen:

In diesem Kapitel sollen Überlegungen zur Generalisierbarkeit studentischer Bewertungen **dozentenbezogener** Daten angestellt werden, wobei die interessierende Fragestellung lautet, inwiefern dozentenbezogene Skalen von unterschiedlichen Studenten in verschiedenen Veranstaltungen übereinstimmend eingeschätzt werden. Im Gegensatz bei Untersuchungen zu Urteilerübereinstimmungen und Interraterreliabilitäten werden hier also Skalen, die Rahmenbedingungen und Veranstaltungsspezifika einfangen, nicht in die Berechnung mit einfließen, da nur die Bewertung der Leistung des Dozenten über unterschiedliche Veranstaltungen den Untersuchungsschwerpunkt darstellt und Rahmenbedingungsdaten aufgrund der Verschiedenartigkeit der Veranstaltungen (andere Vorlesungsthemata, Besucheranzahl der Vorlesungen variiert, Pflicht- versus Wahlveranstaltungen, Anforderung,...) grundsätzlich über die verschiedenen Veranstaltungen als voneinander unabhängig bzw. gegensätzlich angesehen werden und den so berechneten Korrelationskoeffizienten ungerechtfertigter Weise schmälern bzw. beeinflussen würden. Berechnet werden soll hier also nur das Maß der Übereinstimmung der Einschätzungen der Unterrichtsleistung der in der Person des Dozenten als unverändert feststehenden Größe über die verschiedenen Veranstaltungen, dass heißt untersucht wird, wie maßgebend die Veranstaltung mit

den zugehörigen Faktoren Thema, Rahmenbedingung (Größe der Veranstaltung),... die dozentenbezogenen Skalen beeinflussen. Der nun folgend berechnete Korrelationskoeffizient der „Generalisierbarkeit studentischer Urteile des Lehrverhaltens des Dozenten“ über verschiedene Veranstaltungen **Ge** hängt nicht nur von der Zusammensetzung und Ähnlichkeit der Beurteilungsstichproben in den verschiedenen Veranstaltungen ab, sondern auch von der „wirklichen“ Stabilität der tatsächlichen Leistungserbringung seitens des Lehrenden. Ob nun die „Unstimmigkeiten“ in den Bewertungen auf wirkliche, reale Unterschiede in der Lehrleistung des Dozenten oder aber auf ungleiche Bewertungsmaßstäbe der Urteilerstichproben zurückzuführen sind, ist anhand dieser Untersuchung nicht trennbar aber auch nicht zwingend relevant. Vielmehr liegt der Nutzen dieser Untersuchung festzustellen, wie viel Vorlesungsevaluationen eines bestimmten Dozenten in für ihn repräsentativen Fächern notwendig sind, um einigermaßen gesicherte Aussagen bezüglich seiner Lehrkompetenz machen zu können. Grundsätzlich werden Daten aus den verschiedenen Veranstaltungen nicht zusammengefasst oder in irgend einer Weise gemittelt, da mit solchen Operationen stets ein Informationsverlust gegenüber den Originaldaten verbunden ist. Wir benützen zur Berechnung in unserem Beispiel die Rohdaten nach Abbildung 19 oben, wobei die 2 Skalen s1 und s2 nun **dozentenbezogene** Daten aufweisen sollen und dem entsprechend die Urteilermittelwerte die Mittelung über alle dozentenbezogenen Skalen der Urteiler darstellen. Zur korrekten Berechnung des Generalisierungskoeffizienten müssen die dozentenbezogenen Daten nun aber bezüglich der verschiedenen Urteiler normalisiert und in folgender Weise zur Berechnung angeordnet werden:[35]

[35] Siehe Abbildung 25.

			Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	
Veranst. 1	S 1	Item 1	5	4	6	
		Item 2	5	4	7	
		Item 3	4	3	5	
	S 2	Item 1	2	1	3	
		Item 2	2	2	4	
Urteilermittel:			3,6	2,8	5	
Standardabw.			1,356465997	1,166190379	1,414213562	
			Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4
Veranst. 2	S 1	Item 1	6	3	7	4
		Item 2	5	2	4	2
		Item 3	7	5	7	6
	S 2	Item 1	4	3	5	2
		Item 2	5	5	5	4
Urteilermittel:			5,4	3,6	5,6	3,6
Standardabw.			1,019803903	1,2	1,2	1,496662955

			Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4
	S 1	Item 1	1,032093693	1,028991511	0,707106781	0,588348405	-0,5	1,166666667	0,267261242
		Item 2	1,032093693	1,028991511	1,414213562	-0,39223227	-1,333333333	-1,33333333	-1,06904497
		Item 3	0,294883912	0,171498585	0	1,568929081	1,166666667	1,166666667	1,603567451
	S 2	Item 1	-1,179535649	-1,543487266	-1,414213562	-1,372812946	-0,5	-0,5	-1,06904497
		Item 2	-1,179535649	-0,685994341	-0,707106781	-0,39223227	1,166666667	-0,5	0,267261242
Urteilermittel:			0	1,55431E-16	0	-3,21965E-16	-4,44089E-17	2,77556E-16	-8,8818E-17
Standardabw.			1	1	1	1	1	1	1

Veranstaltung 1 Veranstaltung 2

			Veranst. 1	Veranst. 2
	S 1	Item 1	2,768191985	1,522276314
		Item 2	3,475298766	-4,127943905
		Item 3	0,466382497	5,505829866
	S 2	Item 1	-4,137236478	-3,441857914
		Item 2	-2,572636771	0,541695638
Urteilermittel:			0	-8,88178E-17
Standardabw.			2,956122008	3,515998539

Abbildung 25 Datenmanipulation zur Berechnung des Generalisierungskoeffizienten.

Es wäre nun falsch die gemittelte Korrelation über alle Kombinationen der in unserem Beispiel insgesamt 7 Urteiler aus den 2 Veranstaltungen zu berechnen. Es interessiert uns nur die mittlere Korrelation aller Kombinationen der Urteiler aus Veranstaltung 1 mit jenen aus Veranstaltung 2, die gegenseitigen Kombinationen innerhalb der Veranstaltungen wurden schon in der Urteilerübereinstimmung berücksichtigt.
Um nun die gegenseitigen Kombinationen innerhalb der Veranstaltungen aus der Berechnung auszuschließen, müssen die Urteilerwerte innerhalb der Veranstaltung mittels Summation über die Urteiler einer Veranstaltung zusammengefasst werden (siehe Abbildung 25 unten). Die so entstandene Datenmatrix der unterschiedlichen Veranstaltung bildet den Ausgangspunkt zur Berechnung des Generalisierungskoeffizienten **Ge.**
Wir berechnen nun mit diesen Daten den Konsistenzkoeffizienten in Analogie nach Gleichung 31 und Abbildung 20, wobei die Kennziffern unter Gleichung 31 folgendes Aussehen haben:

$$(1) = \frac{G^2}{N}$$

$$(2) = \sum_{i=1}^{p} \sum_{m=1}^{r} x^2{}_{im}$$

$$(3) = \sum_{i}^{p} \left(\frac{A_i{}^2}{r} \right)$$

$$p = \sum_{s} is$$
$$N = r * p$$

In dieser Gleichung bedeuten:
G = die Summe über alle N Werte der über die Urteiler der jeweiligen Veranstaltung summierten dozentenbezogenen Daten,[36]
A_i = die Zeilensumme über die in der Zeile i befindlichen r Werte entsprechend der in der Berechnung einbezogenen Veranstaltungen
p = die Anzahl der Zeilen der Matrix erhoben über die Itemdaten der s dozentenbezogenen Skalen.
Is = die Anzahl der Items, die zur Beschreibung der Dimension s verwendet wird und schließlich bedeutet Ausdruck (2) die Summe über alle quadrierten Werte der Matrix.

Gleichung 36

Und die Freiheitsgrade sich wie folgt ergeben:

	Quadratsumme (QuS):	Freiheitsgrade (df):	Varianz:
Zw._den_Items:	(3)-(1)	$\left(\sum_{s} is - 2\right)$	$\frac{Qus_{Zw\_den\_Items}}{df_{Zw\_den\_Items}}$
Fehler:	(2)-(3)	$\left(\sum_{s} is - 2\right) * (r-1)$	$\frac{Qus_{Fehler}}{df_{Fehler}}$
Total:	(2)-(1)	$\left(\sum_{s} is - 2\right) * r$	$\frac{Qus_{Total}}{df_{Total}}$

Abbildung 26

Diesen Konsistenzkoeffizienten wollen wir mit **KoGe** bezeichnen. Wir erhalten in unserem Beispiel ein KoGe von abgerundet 0,18.[37]
Der Generalisationskoeffizienten berechnet sich über Gleichung 37, indem wir für r = 2 entsprechend der in unserem Beispiel angenommenen Anzahl der Veranstaltungen setzen und diesen so gewonnen Ausdruck noch mit dem Faktor $\frac{\sigma_m^{\,2}}{\sum_{i<j}^{r} it * jt}$ multiplizieren, wobei σ_m^2 für das arithmetische Mittel der Veranstaltungsvarianzen steht und der Nenner des Faktors die Anzahl der Urteilerkombinationen über alle r Veranstaltungen darstellt, wobei jedoch nur Urteilerkombinationen unterschiedlicher Veranstaltungen $i \neq j$ und $i<j$ in die Berechnung eingehen. it bedeutet dabei die Urteileranzahl der Veranstaltung r = i.

$$Ge = \frac{\frac{1}{r} * KoGe}{1 + \left(\frac{1}{r} - 1\right) * KoGe} * \frac{r * (r-1)}{2} * \frac{\sigma_m^{\,2}}{\sum_{i<j}^{r} it * jt}$$

Gleichung 37

Der Ausdruck $\sum_{i<j}^{r} it * jt$ ist für mehrer Veranstaltungen besser zu erhalten, indem wir Gleichung 38 benützen.

[36] G berechnet sich aufgrund der stattgefundenen Normalisierung zu 0.
[37] Vergleiche auch Abbildung 27.

$$\sum_{i<j}^{r} it * jt = \frac{\left(\sum_{r} rt\right)^2 - \sum_{r} rt^2}{2}$$

Gleichung 38

Zur Verdeutlichung sind die entsprechenden Schritte der Berechnung in Abbildung 27 dargestellt, wobei FG für die Freiheitsgrade QuS für die entsprechenden Quadratsummen, die lila Bereiche für die zugehörigen Varianzen stehen, die gelb markierten Bereiche dem Ausdruck (3) nach Gleichung 36 entsprechen, orange das arithmetische Mittel der Veranstaltungsvarianzen darstellt, grün für den Ausdruck $\frac{\frac{1}{r} * KoGe}{1 + \left(\frac{1}{r} - 1\right) * KoGe}$ in Gleichung 37 steht und treat dem entsprechenden Ausdruck Zw._den _Items in Abbildung 26 gleichzusetzen ist.

FG		QuS	V1	V2			KoGe	
6	total	105,504515	2,76819199	1,52227631	9,20405911	0,18203453	0,10013091	
3	treat	58,034389	3,47529877	-4,1279439	0,21297284		0,08803552	Ge
3	fehler	47,4701262	0,4663825	5,50582987	17,8336603			
	treat	19,3447963	-4,13723648	-3,44185791	28,7213359			
	fehler	15,8233754	-2,57263677	0,54169564	2,06236094			
			8,73865733	12,3622457	10,5504515	σ_m^2		

Abbildung 27

Die blau hinterlegte Zahl schlussendlich ist der gesuchte Generalisierungskoeffizient, der sich aufgerundet zu 0,09 ergibt. Dieser so berechnete Generalisierungskoeffizient Ge entspricht der Mittelung aller möglichen Urteilerkombinationen über die verschiedenen Veranstaltungen hinweg, in unserem Beispiel der Mittelung der 3*4 = 12 Urteilerkombinationen. Wird nun dieser Koeffizient mittels der Gleichung 27 „verlängert“, so kann bei Vorliegen einer grundsätzlich frei wählbaren Zuverlässigkeitsuntergrenze von z.B. 0,8 die benötigte Anzahl an bewerteten Vorlesungen abgeschätzt werden, um zu gesicherten Aussagen bezüglich der Lehrkompetenz des betreffenden Dozenten zu gelangen.

$$\frac{r * Ge}{1 + (r-1) * Ge} \geq 0{,}8 \quad .$$

Gleichung 39

In unserem Beispiel müssten 42 repräsentative Vorlesungen evaluiert werden, um den Wert von 0,8 zu übertreffen.[38]
Es soll noch einmal deutlich gemacht werden, dass obige Daten nur ein fiktives Beispiel zur Erläuterung der notwendigen Berechnungsschritte darstellt. Daher resultiert auch dieser kleine Generalisierungskoeffizient und in weiterer Folge die enorm große Anzahl an benötigten Vorlesungen, die zur gesicherten Aussage bezüglich der Lehrkompetenz benötigt werden. In der Literatur findet man Aussagen zur Zuverlässigkeit in der Bewertung von Dozentenskalen, die schon nach Vorliegen von mindesten 5 Vorlesungen gesicherte Rückschlüsse auf dieselbigen erlauben.[39]

[38] Vgl. auch Bemerkungen zur repräsentativen Zusammensetzung der Beurteilungsstichprobe
[39] Vgl. [9], Seite 199.

3.4.2.4. Faktorenanalyse:

Da die Faktorenanalyse ein zentrales statistisches Element in der sozialwissenschaftlichen Forschung darstellt, soll auf diese etwas ausführlicher in diesem Kapitel eingegangen werden, wobei sich die Darstellungen auf die Hauptkomponentenanalyse beschränken.
Ziel einer Faktorenanalyse ist es für eine größere Menge von Variablen eine ordnende Struktur zu unterlegen, die angibt welche und wie viele unabhängige Variablengruppen, die sich innerhalb der Variablengruppe durch wechselseitig hohe Korrelationen und zwischen den Gruppen durch eher kleine Korrelationen um 0 auszeichnen, existieren und am besten durch welche Faktoren diese Variablenstrukturen abgebildet werden können, die voneinander unabhängig sein sollen. Die Faktorenanalyse liefert Ladungszahlen, die angeben wie gut oder schlecht eine Variable zu einem Faktor sprich Variablengruppe passt. Aus diesen Ladungszahlen können interpretative Schlussfolgerungen über das Gemeinsame einer Variablengruppe angestellt werden. Hinter dem Wort Faktor verbirgt sich also die für Zusammenhänge zwischen Variablen gemeinsam wirksame Größe, die die festgestellten korrelativen Beziehungen erklärt. Je höher die Korrelation zwischen Variablen ausfällt desto ähnlicher sind die Informationen die durch sie erfasst werden, dass heißt die Messung einer Variablen erübrigt bei hohen Korrelationen weitgehend die Messung der anderen Variablen. Nach diesen Ausführungen ist die Aufgabe der Faktorenanalyse leicht zu verstehen. Ausgehend von stark korrelierenden Variablen wird eine „synthetische“ Größe sprich ein Faktor konstruiert, die mit den zu erklärenden Variablen so hoch wie möglich korreliert. Ein Faktor stellt somit die für die wechselseitig hoch gemessen Korrelationen zwischen den Variablen die bestimmende Größe dar. Wird nun dieser gemeinsame Faktor aus den Variablen herauspartialisiert, ergeben sich restliche Partialkorrelationen, die diejenige Zusammenhänge zwischen den Variablen widerspiegeln, die nicht durch den extrahierten Faktor erklärt werden. Zur Klärung dieser Restkorrelation wird deshalb ein weiterer Faktor bestimmt, der vom ersten Faktor unabhängig ist und die verbleibenden Zusammenhänge möglichst gut erklärt. Dieser zweite gefundene Faktor wird nun erneut aus den noch verbliebenen Korrelationen herauspartialisiert, was zu einer erneuten Reduktion der Zusammenhänge zwischen den Variablen führt. Auf diese Art und Weise werden die Restkorrelationen sukzessive zum Verschwinden gebracht. Das Ergebnis einer Faktorenanalyse stellen somit wechselseitig unabhängige Faktoren dar, die die Zusammenhänge zwischen den ursprünglichen Variablen in datenreduzierender Weise erklären. Es soll noch darauf hingewiesen werden, dass mit Hilfe der Faktorenanalyse das Vorliegen von Ein- oder Multidimensionalität komplexer Befragungsmerkmale untersucht werden kann, was in weiterer Folge wichtig für den Testentwurf und Auswertung desselben ist. Bei Vorliegen von Multidimensionalität des Untersuchungsmerkmals ist es nicht statthaft Teilergebnisse der Untersuchung summativ in ein Gesamtergebnis überzuführen. Hier liegt vielmehr ein Testprofil vor, deren Bestandteile die unterschiedlichen Dimensionalitäten der Merkmalsausprägung ausdrücken. Unter 3.4.2.5 wird eine Faktorenanalyse exemplarisch durchgerechnet, um Methode und Grenzen der Aussagekraft der Faktorenanalyse zu verdeutlichen. Es ist aber generell empfehlenswert die Faktorenanalyse mittels EDV-Anlagen durchzuführen, da sie ein relativ rechenaufwendiges Analyseverfahren darstellt. Es existiert eine Vielzahl von

Softwarepaketen, die die gängigsten Statistik-Prozeduren zur Verfügung stellen (SPSS, BMDP, SAS, STATISTICA, SPLUS).[40]
Im folgenden soll nun kurz das Prinzip der Hauptkomponentenanalyse erklärt werden. Es sollen von n Personen p Merkmale (Variablen) erhoben werden. Diese Daten sollen in der (n x p) Matrix X abgelegt werden.

$$X = \begin{bmatrix} x_{11} & x_{12} & . & . & . & x_{1p} \\ . & . & & & & \\ . & & . & & & \\ . & & & . & & \\ x_{n1} & . & . & . & . & x_{np} \end{bmatrix}$$

Diese über die n Personen gemessenen p Merkmale sollen mehr oder weniger über die Personen miteinander korrelieren, sodass wir vermuten können, dass für das Zustandekommen der p Merkmale, gemessen für eine beliebige Person j, gemeinsame Faktoren der Person j maßgeblich beteiligt bzw. verantwortlich zu machen sind, die sich hinter den Korrelationen der Merkmale verbergen. Die Grundgleichung der Faktorenanalyse bildet also folgendes in Matrixschreibweise dargestelltes Gleichungssystem.

$$\begin{bmatrix} x_{11} & x_{12} & . & . & . & x_{1p} \\ . & . & & & & \\ . & & . & & & \\ . & & & . & & \\ x_{n1} & . & . & . & . & x_{np} \end{bmatrix} = \begin{bmatrix} f_{11} & f_{12} & . & . & . & f_{1q} \\ . & . & & & & \\ . & & . & & & \\ . & & & . & & \\ f_{n1} & . & . & . & . & f_{nq} \end{bmatrix} * \begin{bmatrix} a_{11} & a_{21} & . & . & . & a_{p1} \\ . & . & & & & \\ . & & . & & & \\ . & & & . & & \\ a_{1q} & . & . & . & . & a_{pq} \end{bmatrix}$$

oder kurz

$$X = F * A' \quad ,$$

Gleichung 40

wobei die j-te Zeile in der Matrix F die Ausstattung der Person j mit den gemeinsam vermuteten Faktoren f_{j1}, f_{j2},... f_{jq} angeben, die für das Zustande kommen der Merkmalsausprägungen x_{j1}, x_{j2},... x_{jp} bestimmend sind. Dabei sind die gemeinsam gefundenen Faktoren der Person j in unterschiedlicher Weise für die ebenfalls unterschiedlichen Merkmalsausprägungen „wichtig“ und werden dem entsprechend mit den dazugehörenden Gewichtungen je nach Merkmal multipliziert. Das bedeutet, dass z.B. für das Merkmal x_{j1} andere Gewichtungen a_{j1}, $a_{j2,}$... a_{jq} für die Faktoren zum Tragen kommen als für das Merkmal x_{j2}.
Die $f_{i,j}$ und die $a_{i,j}$-Werte werden nun so bestimmt, dass nach Gleichung 40 Messwerte vorhergesagt werden können, die möglichst wenig von den tatsächlichen $x_{i,j}$-Werten abweichen. Die Hauptkomponentenanalyse geht somit ähnlich vor wie die multiple Regressionsrechnung. Den unbekannten Gewichten in der Regressionsrechnung, die üblicherweise mit b bezeichnet werden, entsprechen hier in der Faktorenanalyse die unbekannten $a_{i,j}$-Werte und den bekannten Prädiktorvariablen in der Regression entsprechen die unbekannten $f_{i,j}$-Werte.
Zur Vereinfachung der nun folgenden Berechnungsschritte werden die Daten in der Matrix X zuvor noch normalisiert, dass heißt der Mittelwert einer Spalte über alle n Werte wird vom ursprünglichen Wert abgezogen und dieses Ergebnis noch durch die Standardabweichung, berechnet über die Spaltenwerte der Matrix X, dividiert, sodass die sich ergebende neue Datenmatrix Z eine Standardabweichung und einen

[40] Vgl. auch [8].

Mittelwert, berechnet jeweils über die neuen Spaltenwerte, von 1 bzw. 0 hat. Nicht nur die Berechnungsschritte werden durch die Normalisierung erleichtert, sondern auch eine Vergleichbarkeit der p Merkmale (Variablen), die in unterschiedlichen Maßeinheiten vorliegen können, ermöglicht.
Somit ergibt sich die veränderte Grundgleichung der Faktorenanalyse zu:

$$Z = F * A'$$

Gleichung 41

Gesucht in der Hauptkomponentenanalyse ist nun jene orthogonale Rotationstransformation der ursprünglichen Koordinaten (z-Werte) auf neue Faktoren (Achsen), sodass die transformierten Werte (neue Koordinaten) sukzessiv maximale Varianz aufklären. Eine Rotationstransformation zu den neuen Koordinaten wird durch folgende allgemeine Matrixgleichung realisiert:

$$\begin{bmatrix} y_{11} & y_{12} & \cdot & \cdot & \cdot & y_{1p} \\ \cdot & \cdot & & & & \\ \cdot & & \cdot & & & \\ \cdot & & & \cdot & & \\ y_{n1} & \cdot & \cdot & \cdot & \cdot & y_{np} \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdot & \cdot & \cdot & z_{1p} \\ \cdot & \cdot & & & & \\ \cdot & & \cdot & & & \\ \cdot & & & \cdot & & \\ z_{n1} & \cdot & \cdot & \cdot & \cdot & z_{np} \end{bmatrix} * \begin{bmatrix} v_{11} & v_{12} & \cdot & \cdot & v_{1p} \\ \cdot & \cdot & & & \\ \cdot & & \cdot & & \\ \cdot & & & \cdot & \\ v_{p1} & \cdot & \cdot & \cdot & v_{pp} \end{bmatrix}$$

oder kurz

$$Y = Z * V$$

Gleichung 42

Wobei die Zeilenvektoren in Z und Y jeweils die in p Variablen gemessenen Koordinaten darstellen. Für die Transformationsmatrix V als Rotationsmatrix müssen folgende Bedingungen gelten:

$$V' * V = I$$
$$|V| = 1$$

Gleichung 43

V' bedeute die transformierte Matrix von V und I beschreibt die Einheitsmatrix. Die untere Formel in Gleichung 43 bedeutet die Determinante von V. Zur Generierung einer Transformationsmatrix V, die sukzessiv maximale Varianz aufklärt, muss folgende Gleichung, die aus der Maximierung der Diagonalelemente der Matrix (Y'*Y)/n unter der Nebenbedingung v'*v = 1 nach einigen Umformungen folgt, erfüllt sein:

$$(R - \lambda * I) * v = 0$$

wobei R = (Z'*Z)/n, die (pxp) Korrelationsmatrix der p Variablen, v, ein (px1) Spaltenvektor der gesuchten Matrix V bildet, und 0, rechts vom Gleichheitszeichen, ein mit 0-en belegter (px1) Vektor darstellt.

Gleichung 44

Obige Gleichung ist für einen Vektor v ≠ 0 dann erfüllt, wenn die Determinante des Ausdrucks (R-λ*I) = 0 ist. Dies führt zum Lösen des Eigenwertproblems obigen Ausdrucks und die gesuchten Spalten von V bilden dabei die zu den Eigenwerten gehörenden Eigenvektoren, die auf die Länge 1 normiert sind (dann gilt auch V'*V = I). Dabei werden die Eigenvektoren der Reihe nach, beginnend mit dem Eigenvektor mit dem größten, zugehörigen Eigenwert, in die Matrix V aufgenommen. Dass heißt die (px1) Eigenvektoren (Spalten der Matrix V) v_1, v_2,... v_p werden entsprechend ihrer zugehörigen Eigenwerte $\lambda_1 \geq \lambda_2,...\geq \lambda_p$ gereiht angeordnet.
Die Matrix

$$\frac{Y'*Y}{n} = \frac{(V'*Z')*(Z*V)}{n} = V'*\frac{(Z'*Z)}{n}*V = V'*R*V$$

ergibt sodann die Diagonalmatrix Λ der Eigenwerte von R

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & . & . & 0 \\ 0 & \lambda_2 & & & \\ . & & . & & \\ . & & & . & \\ 0 & . & . & . & \lambda_p \end{bmatrix}$$

Gleichung 45

Wird die Matrix Y mittels der Matrix $\Lambda^{(-1/2)}$ normalisiert folgt die gesuchte Matrix der neuen Faktorwerte F:

$$F = Y * \Lambda^{-\left(\frac{1}{2}\right)}$$

Gleichung 46

$\Lambda^{(-1/2)}$ ist eine Diagonalmatrix in deren Diagonale sich die Reziprokwerte der Wurzeln der Eigenwerte ($\frac{1}{\sqrt{\lambda_i}}$) befinden.

Für F gilt die Beziehung:

$$\frac{F'*F}{n} = \frac{\left(\Lambda^{-\left(\frac{1}{2}\right)} * V'*Z'\right)*\left(Z*V*\Lambda^{-\left(\frac{1}{2}\right)}\right)}{n} = \Lambda^{-\left(\frac{1}{2}\right)} * \Lambda * \Lambda^{-\left(\frac{1}{2}\right)} = I$$

Gleichung 47

Und mit Gleichung 41 folgt schließlich:

$$R = \frac{Z'*Z}{n} = \frac{(F*A')'*(F*A')}{n} = A*\frac{F'*F}{n}*A' = A*A'$$

Gleichung 48

Die Korrelationsmatrix der ursprünglichen p Variablen kann also nach Gleichung 48 durch die Faktorladungsmatrix abgebildet werden. Diese Beziehung gilt exakt falls alle p Faktoren extrahiert werden.
A bezeichnet die Matrix der Faktorladungen der p Variablen auf den neuen Faktoren, diese entsprechen den Korrelationen der Variablen mit den Faktorwerten.
Aus der Beziehung in Gleichung 40 sieht man, dass sich z.B. die Faktorladung a_{p1} als Korrelation der Faktorwerte des ersten Faktors mit den Werten der p-ten Variable ergibt.
Die Matrix der Faktorladungen erhalten wir aus den Beziehungen nach Gleichung 41 und Gleichung 46 schließlich zu:

$$A = V * \Lambda^{\left(\frac{1}{2}\right)}$$

Gleichung 49

Zur Ermittlung der Faktorwerte nach Gleichung 46 muss noch folgendes gesagt werden. Durch rechtsseitige Multiplikation der Gleichung 41mit A folgt:

$$Z*A = F*(A'*A)$$

Gleichung 50

(A'*A) entspricht hierbei aber genau der Diagonalmatrix Λ, wie durch einsetzten der Beziehung nach Gleichung 49 leicht verifiziert werden kann. Enthält nun aber diese Diagonalmatrix Λ Eigenwerte $\lambda_i = 0$ so ist (A'*A) nicht mehr invertierbar und die Berechnung der Faktorwerte nach Gleichung 46 nicht möglich.
Streichen wir die 0er Spalten aus der (pxp) Matrix nach Gleichung 49, erhält man die verkürzte (pxq) Matrix $\widetilde{A}_v$ mit q < p.

Aufgrund der Tatsache, dass meistens weniger Faktoren als zu Grunde liegende Merkmale (p Variablen) extrahiert werden, wird die Faktorladungsmatrix in den meisten Fällen nicht quadratisch sein. Es soll noch einmal ausdrücklich darauf hingewiesen werden, dass die Beziehung, nämlich R, nach Gleichung 48 nur exakt reproduziert wird, falls p Faktoren extrahiert werden, Unabhängigkeit zwischen den Faktoren und lineare Abhängigkeiten zwischen den Variablen bestehen. Nur für den Fall, dass einige Eigenwerte $\lambda_i = 0$ sind, wird die (pxp) Korrelationsmatrix mit einer verkürzten (pxq) mit q < p Faktorladungsmatrix (die q extrahierten Faktoren entsprechen allen Eigenwerten ≠ 0) $\widetilde{A}_v$ exakt widergegeben. Werden, wie allgemein üblich, auch Faktoren, die Eigenwerte ≠ 0 besitzen, nicht extrahiert, folgt eine reproduzierte Korrelationsmatrix durch die (pxq) Faktorladungsmatrix $\widetilde{A}$, die sich von der ursprünglichen Korrelationsmatrix unterscheidet. Dieser nicht reproduzierte Varianzanteil aufgrund fehlender Extraktion von Faktoren wird bewusst in Kauf genommen und zu Gunsten einer Datenreduktion als Informationsverlust deklariert. Denn das Ziel der Hauptkomponentenanalyse stellt ja gerade eine möglichst umfassende Reproduktion der Datenstruktur durch möglichst wenig Faktoren dar. Gleichung 50 stellt die Bedingungsgleichung für die gesuchte Faktorwertmatrix dar. Dieses Gleichungssystem ist für den Fall, dass die extrahierten Faktoren q < p (p ist die ursprüngliche Anzahl der Merkmale, Variablen) sind und der Rang der Matrix $(A|\underline{z_i})$ = q+1 ein nicht exakt lösbares für F.

Dies stellt den Regelfall in der Faktorenanalyse dar, da meistens nur q < p Faktoren extrahiert werden und die nicht extrahierten Faktoren auch meistens Eigenwerte ≠ 0 besitzen.
Die Faktorwerte müssen dann geschätzt werden. Je nach Wahl des Schätzverfahrens variiert die Lösung für die Faktorwertmatrix $\widetilde{F}$.
Wird nun Gleichung 50 mit dem jetzt invertierbarem Ausdruck $\left(\widetilde{A}'*\widetilde{A}\right)$, die (pxq) Matrix $\widetilde{A}$ enthält keine 0er-Spalten (die Berechnung für $\widetilde{A}_v$ ist formal gleich) von rechts multipliziert, erhalten wir die Faktorwerte zu:

$$\widetilde{F} = Z * \widetilde{A} * \left(\widetilde{A}'*\widetilde{A}\right)^{-1}$$

Gleichung 51

Diese Berechnung der Faktorwerte nach obiger Vorschrift entspricht der Kleinstquadratschätzung. $\widetilde{F}$ wird dabei so bestimmt, dass die Summe der quadrierten Abweichungen des ursprünglichen Wertes Z' vom Schätzwertes $\hat{Z}'$ über die p Merkmale für einen Probanden j minimal wird, dass heißt für den (px1) Residualvektor $\underline{u_i}$ gilt für alle i = 1..n:

$$\underline{u_i}'*\underline{u_i} = \left(\underline{z_i} - \underline{\hat{z}_i}\right)' * \left(\underline{z_i} - \underline{\hat{z}_i}\right) = \left(\underline{z_i} - \widetilde{A} * \underline{f_i}\right)' * \left(\underline{z_i} - \widetilde{A} * \underline{f_i}\right) = \min$$

mit

$$\hat{Z}' = \widetilde{A} * \left\lfloor \underline{f_1} \quad . \quad . \quad . \quad \underline{f_n} \right\rfloor = \widetilde{A} * \widetilde{F}' \approx \left\lfloor \underline{z_1} \quad . \quad . \quad . \quad \underline{z_n} \right\rfloor = Z'$$

$\underline{f_i}$ bzw. $\underline{z_i}$ bilden dabei die (qx1) bzw. (px1) Spaltenvektoren der entsprechenden Matrizen.

Gleichung 52

Folgende Beziehungen gelten im speziellen für eine verkürzte (pxq) Matrix $\widetilde{A}_v$ bzw. für die (qxq) Eigenwertmatrix $\widetilde{\Lambda}_v$ (Gleichung 53-Gleichung 56):

$$\widetilde{A}_v = V * \widetilde{\Lambda}_v^{\left(\frac{1}{2}\right)}$$

Gleichung 53

$$\left(\widetilde{\Lambda}_v^{\left(\frac{1}{2}\right)}\right)' * \widetilde{\Lambda}_v^{\left(\frac{1}{2}\right)} = \widetilde{\Lambda}_v$$

und

$$\widetilde{\Lambda}_v^{\left(\frac{1}{2}\right)} * \left(\widetilde{\Lambda}_v^{\left(\frac{1}{2}\right)}\right)' = \Lambda$$

Λ entspricht der originären (pxp) Eigenwertmatrix von R, die nun auch Eigenwerte = 0 und damit 0er-Spalten enthalten kann.

Gleichung 54

Mit Gleichung 43, Gleichung 53 und Gleichung 54 folgt:

$$\widetilde{A}_v' * \widetilde{A}_v = \widetilde{\Lambda}_v$$

Gleichung 55

Mit Gleichung 45, Gleichung 43, Gleichung 54 und Gleichung 53 folgt:

$$\widetilde{A}_v * \widetilde{A}_v' = R$$

,

im Gegensatz zu

$$\widetilde{A} * \widetilde{A}' \neq R$$

Gleichung 56

Betrachten wir für den allgemein eintretenden Fall, dass nicht alle Faktoren (auch solche die Eigenwerte ≠ 0 aufweisen) extrahiert werden, die Eigenschaften der Matrizen $\widetilde{A}$ und $\widetilde{F}$ etwas genauer:
In Analogie zur Gleichung 49 erhalten wir

$$\widetilde{A} = V * \widetilde{\Lambda}^{\frac{1}{2}}$$

wobei $\widetilde{\Lambda}$ die (qxp) Eigenwertmatrix beschreibt.

Gleichung 57

Für die nach Gleichung 51 gefundenen Faktorwerte $\widetilde{F}$ gilt:

$$\frac{\widetilde{F}' * \widetilde{F}}{n} = \left(\widetilde{A}' * \widetilde{A}\right)^{-1} * \widetilde{A}' * R * \widetilde{A} * \left(\widetilde{A}' * \widetilde{A}\right)^{-1} = I$$

Gleichung 58

Obiges wird evident, wenn wir für R = V*Λ*V' setzen und für $\widetilde{A}$ die Beziehung nach Gleichung 57 benützen (Λ ist die vollständige (pxp) Eigenwertmatrix von R).
Bezeichnen wir die Residualmatrix mit U so folgt:

$$\widetilde{U} = Z - \hat{Z}$$

und mit Gleichung 52 und Gleichung 51

$$\widetilde{U} = Z - \widetilde{F} * \widetilde{A}' = Z * \left[I - \widetilde{A} * \left(\widetilde{A}' * \widetilde{A} \right)^{-1} * \widetilde{A}' \right]$$

Gleichung 59

Multiplizieren wir Gleichung 59 von rechts mit der Transponierten und Division durch n erhält man die residuale Korrelationsmatrix zu:

$$\frac{\widetilde{U}'\widetilde{U}}{n} = \left[I - \widetilde{A} * \left(\widetilde{A}' * \widetilde{A} \right)^{-1} * \widetilde{A}' \right] * R * \left[I - \widetilde{A} * \left(\widetilde{A}' * \widetilde{A} \right)^{-1} * \widetilde{A}' \right] = M * R * M$$

Gleichung 60

Bezeichnen wir den Ausdruck $\widetilde{A} * \left(\widetilde{A}' * \widetilde{A} \right)^{-1} * \widetilde{A}'$ mit K und berücksichtigen die Gültigkeit der Beziehung nach Gleichung 58 so lässt sich zeigen, dass

$$K * R = K * R * K = \widetilde{A} * \widetilde{A}' = \frac{\hat{Z}' * \hat{Z}}{n} = \frac{Z' * \hat{Z}}{n}$$

gilt.

Gleichung 61

Ferner lässt sich aus Gleichung 61

$$\frac{Z' * \widetilde{F}}{n} = \widetilde{A}$$

ableiten.

Gleichung 62

Benützen wir Gleichung 60 und Gleichung 61, erhält man

$$M * R * M = R - \widetilde{A} * \widetilde{A}'$$

oder anders ausgedrückt

$$R = \frac{\widetilde{U}' * \widetilde{U}}{n} + \frac{\hat{Z}' * \hat{Z}}{n}$$

Gleichung 63

Die Korrelationsmatrix R setzt sich also additiv durch die erklärte Korrelationsmatrix $\frac{\hat{Z}' * \hat{Z}}{n}$ und die residuale Korrelationsmatrix $\frac{\widetilde{U}' * \widetilde{U}}{n}$ zusammen.

Für eine verkürzte Faktorladungsmatrix $\widetilde{A}_v$ bzw. auch für A wird der Ausdruck in Gleichung 63 M*R*M = 0, da die Beziehungen nach Gleichung 56 bzw. Gleichung 48 gilt. Dass heißt die Matrix Z wird durch die extrahierten Faktoren, dass heißt durch die Faktorenmatrix $\widetilde{F}_v$ und die Faktorladungsmatrix $\widetilde{A}_v$, durch $Z = \widetilde{F}_v * \widetilde{A}_V'$ vollständig erklärt. Selbiges gilt für A.

Ferner gilt $\widetilde{A} * \left(\widetilde{A}' * \widetilde{A} \right)^{-1} * \widetilde{A}' \neq I$ und $\widetilde{A}_v * \left(\widetilde{A}_v' * \widetilde{A}_v \right)^{-1} * \widetilde{A}_v' \neq I$, wie man sich leicht überzeugen kann, während für die (pxp) Faktorladungsmatrix A (R besitzt nur Eigenwerte ≠ 0) $A * \left(A' * A \right)^{-1} * A' = I$ gilt.

Bei einer etwaigen Berechnung der (nxq) Faktorwertmatrix $\widetilde{F}$ ist es dabei unerheblich, ob $\widetilde{F}$ mittels der Beziehung $\widetilde{F} = Z * \widetilde{A} * \left(\widetilde{A}' * \widetilde{A} \right)^{-1}$ mit der (pxq) Ladungsmatrix $\widetilde{A}$ oder mit der (pxp) Matrix A berechnet wird und anschließend die (nxp) Matrix F durch streichen der nicht „benötigten“ (p-q) Spalten auf $\widetilde{F}$ verkürzt wird. Beide Ansätze führen zum gleichen Ergebnis, da ja der Ausdruck

$\widetilde{A} * \left(\widetilde{A}' * \widetilde{A}\right)^{-1} = \widetilde{A} * \widetilde{\Lambda}^{-1}$ die (pxq) Matrix $\widetilde{A}$ dividiert durch die jeweiligen Eigenwerte der entsprechenden Spalte darstellt.

Dies gilt natürlich nicht für den Schätzwert $\hat{Z} = \widetilde{F} * \widetilde{A}' = Z * \widetilde{A} * \left(\widetilde{A}' * \widetilde{A}\right)^{-1} * \widetilde{A}' \approx Z$, der verschiedene Werte bei Verwendung unterschiedlich mächtiger Ladungsmatrizen $\widetilde{A}$ = (pxq) oder A = (pxp) berechnet:
Dabei fallen die Schätzwerte umso genauer im Sinne einer Reproduktion von Z aus, je mehr Faktoren extrahiert werden, dass heißt je größer q ≤ p gewählt wird.
Nach diesen eher theoretischen Ausführungen soll im nächsten Kapitel 3.4.2.5 ein kurzes Beispiel folgen und die Vor- und Nachteile der Faktorenanalyse besprochen werden.

3.4.2.5. Schwierigkeiten beim Einsatz der Faktorenanalyse und Verwendung dieser als Gütekriterien für Items:

Wie schon unter 3.4.2.4 angedeutet wurde, lässt sich mit einer über alle Befragungsitems und über alle Skalen hinweg durchgeführten Faktorenanalyse, deren Aufgabe es ist, Zusammenhangsmuster der verschiedenen Items bzw. Skalen zu strukturieren und auf grundlegende Faktoren zurückzuführen, feststellen, ob es sich um ein ein- oder multidimensionales Erhebungsinstrument handelt.
Dabei spricht nun die Anzahl der gefundenen Faktoren, die sich aus „ähnlichen" Variablenbündeln konstituieren, mit relevantem Erklärungsbeitrag hinsichtlich der Varianzaufklärung aller Befragungsitems entweder für das Vorliegen von Ein- oder Multidimensionalität des Befragungsinstrumentes. Multidimensionalität liegt dann vor, wenn mehr als ein relevanter Faktor gefunden werden kann. Dieser Fragestellung, ob studentische Lehrevaluationsfragebögen bei ausreichender Berücksichtigung der Mehrdimensionalität in den Fragestellungen auch eine empirisch mehrdimensionale Beantwortung desselben durch die Studenten nach sich zieht, wurde schon in anderen wissenschaftlichen Untersuchungen nachgegangen und der interessierte Leser sei auf die entsprechende Literatur verwiesen.[41]
In diesem Kapitel wollen wir die Faktorenanalyse als Mittel zur Gewinnung eines Kriteriumswertes für die verschiedenen Items einsetzen, die angeben wie geeignet die Items das zu messende Skalenmerkmal darstellen bzw. wie eng sie mit diesem verbunden sind. Wie schon früher erwähnt, soll davon ausgegangen werden können, dass sich das zu messende Skalenmerkmal durch verschiedene Items konstituiert, die gleiche bzw. ähnliche Sachverhalte bzw. Eigenschaften ansprechen oder offen legen. Die Zweckmäßigkeit eines Items zur Beschreibung des Skalenmerkmals könnte z.B. mit der Feststellung der Ähnlichkeit eines Items i mit den übrigen Items bewerkstelligt werden, die durch die mittlere Korrelation des Items i mit den übrigen Items der Skala gefunden werden könnte. Diese Art der Berechnung ist jedoch für größere Skalen recht aufwendig.[42]
Die Eignung eines Items zur Abbildung einer Skala drückt sich also durch den korrelativen Zusammenhang des Items mit dem die Skala begründenden, allen Items gemeinsamen, quasi der Skala zu Grunde liegenden Faktor aus.
Daher wird auch für die folgende Berechnung des Gütekriteriums eines Items die Faktorenanalyse jeweils getrennt für die entsprechende Skala durchgeführt, da korrelative Zusammenhangsmuster zwischen verschiedenen Skalen, die nicht nur empirisch belegbar sondern auch theoretisch durchaus denkbar und erklärbar sind, andere Erklärungen der Faktorenstruktur bewirken und somit nicht mehr die

[41] Vgl. z.B.[9]
[42] Vgl. Berechnungsmethode zur Generalisierbarkeit studentischer Aussagen unter 3.4.2.3

gesuchte Ladung bzw. die korrelative Beziehung des Befragungsitems auf bzw. mit dem gemeinsamen Faktor der zu messenden Skala resultiert.
Ausgangspunkt einer Faktorenanalyse bildet die Korrelationsmatrix, die wie im letzten Kapitel berichtet zur Vereinfachung der Rechenschritte aus normalisierten Daten gewonnen wird.
Dabei sind mehrere Möglichkeiten der Normalisierung und Aggregierung der Itemdaten einer Skala s über mehrere Veranstaltungen hinweg denkbar, die alle in der Folge zu unterschiedlichen Ergebnissen der faktorenanalytischen Untersuchung führen und hier kurz besprochen werden sollen.
Wenden wir uns wieder unserem fiktiven Beispiel zu, deren Items der Skala 1 zur Beschreibung dieser nach obigen Kriterien geprüft werden sollen:
Die unter 1 in Abbildung 28 aufgeführte, obere Tabelle enthält die Rohdaten über die Veranstaltungen für die betreffende Skala, die untere Tabelle die bezüglich des Mittelwertes und der Standardabweichung normalisierten Daten, jeweils über die 2 Veranstaltungen berechnet. Eine solchermaßen durchgeführte Normalisierung der Ausgangsdaten zur Berechnung der Korrelationsmatrix und anschließender Faktorenanalyse birgt statistische und interpretatorische Probleme. Die unter 1 normalisierten Daten enthalten 2 verschiedene Varianzquellen. Zum einen beinhalten sie die Varianz der Itemdaten innerhalb einer Veranstaltung zwischen den einzelnen Urteilern und zum anderen die Varianz, die auf die Unterschiedlichkeit der über die Veranstaltungen gemittelten Itembeurteilung zwischen den Veranstaltungen zurückzuführen ist. Die Faktorenstruktur ist somit nicht mehr eindeutig auf eine Varianzquelle rückführbar. Faktorenanalysen auf Basis einer Normalisierung der Rohdaten bezüglich Mittelwert und Standardabweichung, berechnet über alle Veranstaltungen, stellen somit nicht die Methode der Wahl dar.
Die unter 2a oben abgebildete Tabelle enthält die Mittelwerte der Itembeurteilungen je Veranstaltung, dass heißt die Urteilerwerte der betreffenden Items und Veranstaltungen werden durch die entsprechenden Veranstaltungsmittel ersetzt. Die untere Tabelle enthält nun wiederum die bezüglich des Mittelwertes und der Standardabweichung normalisierten Daten, die beide über alle Urteiler und Veranstaltungen berechnet wurden.
Die unter 2b dargestellte Tabelle oben enthält ebenfalls die Itemmittel pro Veranstaltung, die hier allerdings nur einmal pro Veranstaltung aufgeführt wird. In der unteren Tabelle wurden die Daten wiederum bezüglich Mittelwert und Standardabweichung berechnet über alle Veranstaltungsmittelwerte normalisiert.
Faktorenanalysen ermittelt über Daten von Veranstaltungsmitteln stellen die in der Literatur die meist verbreitete und angefundene Methode dar. Grund hierfür liefert die Argumentation, dass zur Analyse Daten herangezogen werden sollen, die frei von individuellen Verzerrungen und Urteilsfehlern, die Lehrsituation geeignet abbilden. Durch Mittelwertbildung werden nun solche individuellen Verzerrung ausgeblendet bzw. verringert. Kritisch zu dieser Vorgehensweise ist anzumerken, dass oftmals die Anzahl der untersuchten Veranstaltungen für eine Faktorenanalyse nicht ausreicht, da die Anzahl der Beobachtungen sprich evaluierten Vorlesungen größer sein muss als die Zahl der in die Untersuchung eingehenden Merkmale. In unserem Fall muss also die Zahl der untersuchten Veranstaltungen größer sein als die Anzahl der erhobenen Itemmerkmale in der größten, dass heißt durch die meisten Items beschriebenen Skala ($r > max(is)$ über alle s), da ja die Faktorenanalyse getrennt für die einzelnen Skalen durchgeführt werden soll. Genauer wird empfohlen, dass die Fallzahl der geprüften Veranstaltungen mindestens das 3-fache der Anzahl an Items betragen ($r \geq max(is)$) sollte. Die verschiedenen methodischen Ansätze nach **2a** oder **2b** spiegeln unterschiedliche Betrachtungsweisen der Anwender wider. Das

Vorgehen nach **2b** unterstreicht die Gleichgewichtung der Veranstaltungsbewertungen durch die jeweils einmalige Nennung des Itemmittelwertes. Schließlich müssen die Lehrleistungen, die unabhängig von der Größe der Hörerschaft erbracht werden und über mehrere Veranstaltungen hinweg beobachtet werden, bei einer Ermittlung einer mittleren Leistung über alle Veranstaltungen hinweg gleich behandelt und daher gleich gewichtet werden. **2b** betont also in seinem methodischem Zugang eher den Aspekt der Leistungserbringung des Dozenten ohne auf die „messtechnischen" Unzulänglichkeiten bei der Ermittlung des „wahren" Wertes der Leistung einzugehen, die sich durch das Verwenden unterschiedlicher Bewertungsmaßstäbe der Verschiedenen Urteiler und durch das Problem der Zusammensetzung der Beurteilungsstichprobe ergeben. Dem gegenüber lenkt die Vorgehensweise nach **2a** das Hauptaugenmerk aber genau auf diese Tatsache der „unsicheren" Beurteilung der „wahren" Leistung des Dozenten durch einen Studierenden. Gemittelte Itemwerte über verschiedene Beurteilungen einer Veranstaltung durch Studierende entsprechen umso mehr der „wahren" Leistung des Dozenten in dieser Veranstaltung je größer die Beurteilungsstichprobe ausfällt, die zur Mittelwertbildung herangezogen werden kann. Werden Einzelbewertungen noch im größerem Ausmaß durch subjektive Verzerrungen und unterschiedliche Bewertungsmaßstäbe beeinflusst, so werden summative Beurteilungen oder Mittelwertbildungen über mehrere Urteiler in dem Sinne stabiler als sich unterschiedliche Bewertungsmaßstäbe von ungleichen Beurteilern durch die Mittelung zunehmend ausgleichen. Dass heißt je größer die Stichprobe der Urteiler in einer Veranstaltung ist umso größer ist die Wahrscheinlichkeit einer repräsentativen Zusammensetzung der Hörerschaft, die eine Verzerrung der „wahren" Leistung des Lehrenden durch eine zu einseitige Zusammenstellung der Stichprobe ausschließt. Daher werden die Itemmittel der Veranstaltungen entsprechend ihrer Veranstaltungsstichprobengröße, wie in **2a** dargestellt, aufgeführt, was bei einer durchschnittlichen Leistungsermittlung über alle Veranstaltungen hinweg einer unterschiedlichen Gewichtung der Itemmittelwerte proportional zur Stichprobengröße zur Folge hat.

Da die Wahl der methodischen Auflistung der Itemmittelwerte nach **2a** oder **2b** vom Anwender abhängt und darüber hinaus für eine einzige Veranstaltung (z. B. für Veranstaltung 1, egal ob nach **2a** oder **2b**) überhaupt keine Berechnung von Korrelationen zwischen verschiedenen Items aufgrund fehlender Variation des Itemmittelwertes in einer Veranstaltung möglich ist (Informationsverlust der Mittelung gegenüber den Rohdaten, wo das möglich ist), wird auch diese Art der Datenmanipulation für die Normalisierung der Daten verworfen.

Schließlich enthält die unter **3** dargestellte obere Tabelle die Rohdaten der Items in den verschiedenen Veranstaltungen, die Itemmittelwerte sowie die Standardabweichungen je Veranstaltung. Die untere Tabelle enthält wiederum die bezüglich jeder Vorlesung normalisierten Daten. Die Vorgehensweise der Normalisierung der entsprechenden Veranstaltungsdaten nach der unter Punkt **3** dargestellten Methode lässt sich wie folgt begründen:

Die Berechnung von Korrelationen über mehrere Veranstaltungen setzt wie bei einer jeden Korrelationsberechnung zwischen zwei Variablen x, y voraus, dass die Daten über die die Korrelation berechnet wird, dass heißt die Realisationen sowohl von x als auch von y, aus einer Grundgesamtheit mit einem entsprechenden Erwartungswert und Varianz stammen. Da die Itemmittelwerte der verschiedenen Veranstaltungen in der Regel jedoch unterschiedliche Werte annehmen, sei es, dass das entsprechende Merkmal realiter sich geändert hat, dass heißt der gemessene Umstand sich tatsächlich im Mittel von einer Vorlesung zu der anderen verbessert

oder verschlechtert hat, oder nur durch eine andere Zusammensetzung der Beurteilungsstichprobe anders bewertet wurde, müssen die Itemdaten, wenn eine Korrelation über mehrere Veranstaltungen berechnet werden soll, bezüglich Mittelwert und Streuung der Veranstaltungsdaten normalisiert werden, quasi auf gleiches Niveau gebracht werden. Der korrelative Zusammenhang zwischen 2 Itemvariablen berechnet über eine Veranstaltung ändert sich durch die vorhergehende Normalisierung der Daten nicht. Wird eine Normalisierung der Itemdaten pro Veranstaltung durchgeführt und eine Korrelation in der Folge über mehrere Veranstaltungen berechnet (alle Daten verfügen nunmehr über denselben Mittelwert und Streuung pro Veranstaltung), entspricht dieses Korrelationsmaß der mit der Stichprobengröße gewichteten Mittelung der Korrelationen, berechnet über die einzelnen Veranstaltungen.[43] Wir werden daher aus obig genannten Gründen eine Normalisierung der Itemdaten nach der in Tabelle **3** beschriebenen Art und Weise für die Berechnung der Korrelationsmatrix vornehmen.

[43] Vgl. auch Ausführungen weiter oben.

1		Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4	Itemmittel:	Standardabw.:
S 1	Item 1	5	4	6	6	3	7	4	5	1,309307341
	Item 2	5	4	7	5	2	4	2	4,142857143	1,641303613
	Item 3	4	3	5	7	5	7	6	5,285714286	1,385051388
S 2	Item 1									
	Item 2									

Veranstaltung 1

Veranstaltung 2

		Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4	Itemmittel:	Standardabw.:
S 1	Item 1	0	-0,763762616	0,763762616	0,763762616	-1,52752523	1,527525232	-0,76376262	0	1
	Item 2	0,522232968	-0,087038828	1,74077656	0,522232968	-1,30558242	-0,08703883	-1,30558242	-2,53765E-16	1
	Item 3	-0,928279122	-1,650273994	-0,20628425	1,237705496	-0,20628425	1,237705496	0,515710623	1,11022E-16	1
S 2	Item 1									
	Item 2									

Veranstaltung 1

Veranstaltung 2

2 a		Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4	Itemmittel:	Standardabw.:
S 1	Item 1	5	5	5	5	5	5	5	5	0
	Item 2	5,333333333	5,333333333	5,333333333	3,25	3,25	3,25	3,25	4,142857143	1,030982624
	Item 3	4	4	4	6,25	6,25	6,25	6,25	5,285714286	1,113461233
S 2	Item 1									
	Item 2									

Veranstaltung 1

Veranstaltung 2

		Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4	Itemmittel:	Standardabw.:
S 1	Item 1	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
	Item 2	1,154700538	1,154700538	1,154700538	-0,8660254	-0,8660254	-0,8660254	-0,8660254	-4,44089E-16	1
	Item 3	-1,154700538	-1,154700538	-1,154700538	0,866025404	0,866025404	0,866025404	0,866025404	9,5162E-17	1
S 2	Item 1									
	Item 2									

Veranstaltung 1

Veranstaltung 2

Abbildung 28

2 b		Veranstaltung 1:	Veranstaltung 2:	Itemmittel:	Standardabw.:
S 1	Item 1	5	5	5	0
	Item 2	5,333333333	3,25	4,291666667	1,041666667
	Item 3	4	6,25	5,125	1,125
S 2	Item 1				
	Item 2				

		Veranstaltung 1:	Veranstaltung 2:	Itemmittel:	Standardabw.:
S 1	Item 1	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
	Item 2	1	-1	4,44089E-16	1
	Item 3	-1	1	0	1
S 2	Item 1				
	Item 2				

3		Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	Itemmittel:	Standardabw.:	Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4	Itemmittel:	Standardabw.:
S 1	Item 1	5	4	6	5	0,816496581	6	3	7	4	5	1,58113883
	Item 2	5	4	7	5,333333333	1,247219129	5	2	4	2	3,25	1,299038106
	Item 3	4	3	5	4	0,816496581	7	5	7	6	6,25	0,829156198
S 2	Item 1											
	Item 2											

Veranstaltung 1

Veranstaltung 2

		Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4	Itemmittel:	Standardabw.:
S 1	Item 1	0	-1,224744871	1,224744871	0,632455532	-1,26491106	1,264911064	-0,63245553	0	1
	Item 2	-0,267261242	-1,069044968	1,336306211	1,347150628	-0,96225045	0,577350269	-0,96225045	1,26883E-16	1
	Item 3	0	-1,224744871	1,224744871	0,904534034	-1,50755672	0,904534034	-0,30151134	1,58603E-17	1
S 2	Item 1									
	Item 2									

Veranstaltung 1

Veranstaltung 2

Abbildung 28

(Fortsetzung)

Mit den aus Abbildung 28 nach Tabelle **3** unten normalisierten Daten Z folgt nun die Korrelationsmatrix nach der Beziehung $R = \frac{Z'*Z}{n}$:

1	0,9077	0,97340719	R
0,907713	1	0,91821352	
0,973407	0,9182	1	

Abbildung 29

Diese Korrelationsmatrix R besitzt die 3 Eigenwerte $\lambda_1 = 2.8665 \geq \lambda_2 = 0.1073 \geq \lambda_3 = 0.0261$ und die zugehörigen Eigenvektoren $[\underline{v}_1, \underline{v}_2, \underline{v}_3] = V$, die der Größe nach entsprechend der Eigenwerte in der Matrix V angeordnet werden:

0,580481	-0,4545	0,67562943	V
0,568891	0,82	0,06284504	
0,582585	-0,3479	-0,73455794	

Abbildung 30

V erfüllt die Beziehung nach Gleichung 43.
Werden die Wurzeln der 3 Eigenwerte in einer Diagonalmatrix angeordnet, so folgt $\Lambda^{1/2}$:

1,693081	0	0	$\Lambda^{1/2}$
0	0,3276	0	
0	0	0,16163188	

Abbildung 31

Aus Gleichung 49 folgt A zu:

0,982802	-0,1489	0,10920325	A
0,963178	0,2687	0,01015776	
0,986363	-0,114	-0,11872798	

Abbildung 32

Wollen wir noch die Faktorwertmatrix $\widetilde{F}$ nach Gleichung 51 bestimmen, so folgt mit der unter Tabelle **3** ,in Abbildung 28 beschriebenen Normalisierung der Daten die (pxn) Matrix Z'[44]

		Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4	
s 1	Item 1	0	-1,22474487	1,224744871	0,632455532	-1,264911064	1,264911064	-0,63245553	Z'
	Item 2	-0,26726124	-1,06904497	1,33630621	1,347150628	-0,962250449	0,577350269	-0,96225045	
	Item 3	0	-1,22474487	1,224744871	0,904534034	-1,507556723	0,904534034	-0,30151134	
		Veranstaltung 1			Veranstaltung 2				

und mit A schließlich nach Abbildung 32 $\widetilde{F}$ zu:

-0,08980222	-0,66888173	-0,10391541	F
-1,20055023	0,323758873	0,030861581	
1,290352453	0,345122853	0,073053832	
0,980743223	1,533844798	-0,94328764	
-1,27575228	0,947029115	1,189765009	
0,938923798	-1,27006828	1,40109794	
-0,64391474	-1,21080564	-1,6475753	

-0,08980222	$\widetilde{F}$
-1,20055023	
1,290352453	
0,980743223	
-1,27575228	
0,938923798	
-0,64391474	

Abbildung 33

[44] p entspricht der Anzahl der Items, die die Skala s beschreiben, und stimmt daher mit is überein.

Wobei F mit der (pxp) sprich (3x3) Matrix A und $\widetilde{F}$ mit der verkürzten (pxq) Matrix $\widetilde{A}$ (in Abbildung 32 die (3x1), orange gefärbte Matrix) berechnet wurde.[45]
Der orange, gefärbte erste Spaltenvektor von A oder gleichbedeutend, die verkürzte Ladungsmatrix $\widetilde{A}$ stellt nun die gesuchten Bestimmungsgrößen, nämlich die Gütekriterien der 3 Befragungsitems, dar. Diese geben an, wie gut sich die entsprechenden Items der Skala 1 (Item 1, Item 2, Item 3) hinsichtlich der „Fähigkeit zur Beschreibung des gemeinsamen, quasi zu Grunde liegenden Faktors der Skala“ eignen. Denn es gilt ja wie früher festgestellt wurde, die Beziehung:

$$\frac{Z'*\widetilde{F}}{n} = \widetilde{A} \quad .$$

Gleichung 64

Dieser (3x1) Spaltenvektor gibt also die korrelative Beziehungen (Ladungen) der p Items mit dem extrahierten, gemeinsamen Faktor an.
Diese Faktorladungen sollten einer allgemeinen Übereinkunft oder Konvention entsprechend mindestens ≥ 0.5 sein,[46] was in unserem angenommenen Beispiel für alle Items zutrifft. Sollte ein Item eine geringere Ladung als 0.5 haben, muss das Befragungsitem entweder ausgeschlossen bzw. anders formuliert werden. Kritisch bei dieser Ermittlung der Gütekriterien bleibt wiederum anzumerken, dass für Skalen, die nur aus einem Befragungsitem bestehen, keine sinnvolle Ermittlung nach der vorgestellten Art möglich ist. Hier muss entweder darauf vertraut werden bzw. „überprüft“ werden, dass das Befragungsitem das zu messende Merkmal inhaltlich geeignet und treffend wiedergibt.
Ein weiteres Problem bei der Anwendung der Faktorenanalyse besteht in der unvollständigen Beantwortung der Fragebögen, das sogenannte „**Missing Value-Problem**“. Grundsätzlich bieten sich dem Anwender 3 Möglichkeiten, der Schwierigkeit fehlender Daten zu begegnen:

1. Listenweiser Fallausschluss,
2. Paarweiser Fallausschluss,
3. Verwendung von Mittelwerten.

Um die 3 verschiedenen Ansätze, wie mit fehlenden Daten zu verfahren ist, deutlich zu machen, kehren wir zu unserem kleinen Datenbeispiel zurück, das nun zufällig gewählte, nicht ausgefüllte Itemdaten enthält (lila Bereiche in Abbildung 34).

			Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	
Veranst. 1	S 1	Item 1		4	6	
		Item 2	5	4	7	
		Item 3	4		5	
	S 2	Item 1				
		Item 2				

			Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4
Veranst. 2	S 1	Item 1	6	3	7	
		Item 2	5	2	4	2
		Item 3	7		7	6
	S 2	Item 1				
		Item 2				

Abbildung 34

[45] Vgl. hierzu die Schlussbemerkung unter 3.4.2.4, Bei einer etwaigen Berechnung...
[46] Vgl. [6].

Ad 1): Beim listenweisen Fallausschluss wird der gesamte Fragebogen des Urteilers, sobald ein fehlender Wert bei den für die Untersuchung relevanten Variablen auftritt, aus der weiteren Analyse ausgeschlossen. Da sich die Analysen in diesem Kapitel jeweils nur auf eine einzelne Skala beschränken (Faktoranalyse wird über die Skalen einzeln gerechnet), sind auch nur diese Daten in dieser Skala über die Veranstaltungen maßgebend. Dass heißt, wurde in irgendeiner Skala s ein Itemwert von einem Urteiler nicht ausgefüllt, streichen wir den Urteiler für die weitere Faktorenanalyse dieser Skala. Bei einem listenweisen Fallausschluss resultiert also folgende Datenmatrix für Skala 1.

			Urteiler 1,3	Itemmittel:	Standardabw.:
Veranst. 1	S 1	Item 1	6	6	0
		Item 2	7	7	0
		Item 3	5	5	0
	S 2	Item 1			
		Item 2			

			Urteiler 2,1	Urteiler 2,3	Itemmittel:	Standardabw.:
Veranst. 2	S 1	Item 1	6	7	6,5	0,5
		Item 2	5	4	4,5	0,5
		Item 3	7	7	7	0
	S 2	Item 1				
		Item 2				

Abbildung 35

Wir bilden wiederum die Mittelwerte und Standardabweichungen über die verschiedenen Veranstaltungen und normalisieren die Daten zu der Ausgangsdatenmatrix.

		Urteiler 1,3	Urteiler 2,1	Urteiler 2,3	
S 1	Item 1	#DIV/0!	-1	1	Z'
	Item 2	#DIV/0!	1	-1	
	Item 3	#DIV/0!	#DIV/0!	#DIV/0!	

Veranstaltung 1 Veranstaltung 2

Abbildung 36

Wie wir sofort aus Abbildung 35 und Abbildung 36 sehen, kann das Verfahren des listenweisen Fallausschlusses nur bei einer hinreichend großen Datenmenge in den jeweiligen Veranstaltungen und einer ebenfalls ausreichenden Datengrundlage über alle Veranstaltungen hinweg angewendet werden. Es handelt sich um ein Verfahren, dass erheblich die ursprüngliche Datenmenge reduzieren kann. Zumindest müssen nach Ausschluss jener Urteiler, die nicht alle Befragungsitems der entsprechenden Skala beantwortet haben, aber Berechnungen von Itemstreuungen ≠ 0 aller Merkmale (Items) in jeder Veranstaltung möglich sein, um eine Normalisierung der Itemdaten in den jeweiligen Veranstaltungen durchführen zu können und die Fallzahl über alle Veranstaltungen sollte noch mindestens das 3-fache der Itemanzahl der untersuchten Skala sein. Methodisch stellt der listenweise Fallausschluss die korrekte Vorgehensweise dar, da es aufgrund der gleichen Fallzahlen in den verschiedenen Items zu keiner Verzerrung der Korrelationsmatrix kommt.[47] Kritisch anzumerken bleibt der mögliche, riesige Datenverlust.

[47] Siehe Absatz Diese so berechnete Matrix... unter Ad) 3.

Ad 2): Beim paarweisen Fallausschluss berechnen wir wiederum zuerst die Itemmittelwerte und –streuungen in den jeweiligen Veranstaltungen. Dabei kann sich die Itemmittelung und Streuung für die verschiedenen Items in der gleichen Veranstaltung nun aber über unterschiedliche Anzahlen von Einzelbeurteilungen berechnen (z.B. für Veranstaltung 2 und Item 2 resultieren 4 Einzelurteile, die zur Mittelungs- und Streuungsberechnung herangezogen werden, und für Veranstaltung 2 und Item 1 kommen 3 Einzelurteile zum Einsatz).[48]

			Urteiler 1,1	Urteiler 1,2	Urteiler 1,3		Itemmittel:	Standardabw.:
Veranst. 1	S 1	Item 1		4	6		5	1
		Item 2	5	4	7		5,333333333	1,247219129
		Item 3	4		5		4,5	0,5
	S 2	Item 1						
		Item 2						

			Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4	Itemmittel:	Standardabw.:
Veranst. 2	S 1	Item 1	6	3	7		5,333333333	1,699673171
		Item 2	5	2	4	2	3,25	1,299038106
		Item 3	7		7	6	6,666666667	0,471404521
	S 2	Item 1						
		Item 2						

Abbildung 37

Werden die Veranstaltungsdaten nunmehr noch normalisiert so folgt die Ausgangsmatrix in Abbildung 38.

		Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4	
S 1	Item 1		-1	1	0,39223227	-1,372812946	0,980580676		Z'
	Item 2	-0,26726124	-1,06904497	1,33630621	1,347150628	-0,962250449	0,577350269	-0,962250449	
	Item 3	-1		1	0,707106781		0,707106781	-1,414213562	

Veranstaltung 1 Veranstaltung 2

Abbildung 38

Zur Ermittlung einer korrelativen Beziehung zwischen 2 Größen benötigen wir zur rechnerischen Durchführung Schätzungen von unbekannten Populationsparametern der 2 Variablen. Diese unbekannten Populationsparameter, nämlich die Streuungen σ_1, σ_2 und die Erwartungswerte μ_1, μ_2 der 2 Variablen werden über die Werte der jeweiligen Variablen aus der Stichprobe geschätzt.
Betrachten wir nun zur Veranschaulichung die Daten der 2 Items 1 und 3 aus Abbildung 38. Beide Datenreihen besitzen aufgrund der Normalisierung Schätzungen für die Erwartungswerte von $\mu_1 = \mu_3 = 0$ (Mittelwerte über die n1 = 5 bzw. n3 = 5 Werte von Item 1 bzw. Item 3) und Schätzungen für die Streuungen $\sigma_1 = \sigma_3 = 1$ (ebenfalls über n1 = n2 = 5 berechnet). Während die Populationsparameter μ, σ der 2 Variablen aus allen zur Verfügung stehenden Werten geschätzt werden, wird die Korrelation zwischen Item 1 und Item 3 über die Anzahl ng = 3 an gemeinsamen Datenpaaren bestimmt.[49]
Die Korrelation ergibt sich über die bekannte Beziehung:

$$r_{1,3} = \frac{\frac{1}{ng} * \sum_{i=1}^{ng} \left(z_{1,i} - \mu_1\right) * \left(z_{3,i} - \mu_3\right)}{\sigma_1 * \sigma_3}$$

Gleichung 65

[48] Vgl. Abbildung 37
[49] Obwohl in ng, der Bezeichnung für die Anzahl an gemeinsamen Datenpaaren zwischen 2 Items, kein Laufindex verwendet wurde, variiert diese aber natürlich von Fall zu Fall der verschiedenen Itemkombinationen (Item i mit Item j, i,j = 1..n).

Setzen wir für die Parameterwerte μ, σ die über n1 bzw. n3 errechneten Schätzwerte in Gleichung 65 ein, so berechnet sich die Korrelation zwischen Item 1 und Item 3 aufgrund der Normalisierung als einfache Produktsumme der entsprechenden Wertepaare aus Abbildung 38 über die Anzahl der gemeinsam, bestehenden Datenpaare ng = 3, die noch durch die Anzahl ng dividiert wird, dass heißt die Korrelationsberechnung der 2 Items vereinfacht sich zur Kovarianzbestimmung über die ng Datenpaare.
Achtung!!: Eine **Schätzung der Populationsparameter** μ und σ der 2 Items für die Korrelationsbestimmung **aus** der verkleinerten Datenmenge der gemeinsamen Wertepaare **ng** = 3 ist zwar methodisch für sich alleine genommen nicht falsch (kritisch anzumerken bliebe aber immer noch der Informationsverlust der nicht berücksichtigten Daten), eine auf diese Weise konstruierte Korrelationsmatrix ist jedoch durch die Verwendung unterschiedlicher Schätzungen für ein und denselben Parameter einer Itemvariablen in einer nicht zulässigen Weise verzerrt und damit **nicht statthaft**. Dies sieht man leicht ein, wenn man z.B. die unterschiedlichen Parameterschätzungen μ für Item 2, die sich über die ungleichen Stichprobengrößen ng je nach Korrelationsbestimmung (Item 1-Item 2; Item 2-Item 2; Item 2-Item 3) ergeben, mit ng gewichtet und anschließend mittelt. Dieses so entstandene gewichtete Mittel entspricht nun aber nicht mehr dem arithmetischen Mittel über alle Daten von Item 2, was einer Gleichgewichtung der Zahlen zur Mittelwertbildung entspräche. Es handelt sich hierbei um eine Mittelung mit einer unterschiedlichen Gewichtung der Zahlen von Item 2, wobei die Gewichtung mit der Häufigkeit der Verwendung dieser Zahl bei den Parameterschätzungen einhergeht. Dabei werden Zahlen, die in Abbildung 38 rot bzw. orange hinterlegt sind 3- bzw. 2-fach in der Mittelung gewertet, was aber ursprünglich nicht zu begründen ist.
Es folgt also die Korrelationsmatrix nach oben beschriebener Berechnungsvorschrift zu:

1	0,9641751	0,65690845
0,9641751	1	0,86504454
0,65690845	0,86504454	1

Abbildung 39 Korrelationsmatrix R = Varianz-Kovarianzmatrix Σ^2 aus den Ausgangsdaten nach Abbildung 38 berechnet.

Wie wir sehen, berechnen sich die Korrelationen bei der Methode des paarweisen Fallausschlusses über eine im Regelfall größere Urteileranzahl als im listenweisen Fallausschluss. Jedoch kommt es durch den paarweisen Fallausschluss zu unterschiedlichen Fallzahlen bei den Korrelations- oder Varianz-Kovarianzbestimmungen. Dadurch kann es zu Ungleichgewichtungen der Variablen und in weiterer Folge zu Verzerrungen der Korrelationsmatrix kommen.[50]
Multiplizieren wir die entsprechenden (pxp) Matrixeinträge der Varianz-Kovarianzmatrix Σ^2 aus Abbildung 39 mit den zugehörigen Fallzahlen ng so folgt die Matrix Ξ^2. In der Diagonale befinden sich die Quadratsummen und an die Stelle der Kovarianz tritt die Produktsumme, berechnet jeweils über die gemeinsam vorhandenen Datenpaare der normalisierten Ausgangsdaten nach Abbildung 38. Ξ^2 nimmt also folgende Form an:

[50] Siehe Absatz Diese so berechnete Matrix... unter Ad) 3.

5	4,82087552	1,97072534
4,82087552	7	4,32522272
1,97072534	4,32522272	5

Abbildung 40: Quadratsummen-Produktsummenmatrix Ξ^2

Ad 3): Bei diesem Verfahren werden keine Daten ausgeschlossen. Die Normalisierung der Veranstaltungsdaten geschieht auf die gleiche Weise wie in Abbildung 37 und Abbildung 38 beschrieben. Dann werden die fehlenden Daten durch eine 0 ersetzt. Dies entspricht dem Mittelwert der Veranstaltungsdaten vor einer erfolgten Normalisierung.
Wir erhalten also folgende Ausgangsdaten:

		Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	Urteiler 2,1	Urteiler 2,2	Urteiler 2,3	Urteiler 2,4	
S1	Item 1	0	-1	1	0,39223227	-1,372812946	0,980580676	0	Z'
	Item 2	-0,26726124	-1,06904497	1,33630621	1,347150628	-0,962250449	0,577350269	-0,962250449	
	Item 3	-1	0	1	0,707106781	0	0,707106781	-1,414213562	
		Veranstaltung 1			Veranstaltung 2				

Abbildung 41

Das unter Punkt 3 genannte Verfahren lässt sich folgendermaßen begründen. Durch das Ersetzen der fehlenden j Werte durch eine 0 wird der (p-j) dimensionale Vektor (Spaltenvektor in Abbildung 41) nicht verändert, sondern lediglich auf die größere Dimension „gehoben“. Weder Länge noch Lage des ursprünglichen Vektors wird durch diese Art der Manipulation verändert.
Multiplizieren wir nun die so entstandene Ausgangsdatenmatrix nach Abbildung 41 mit ihrer Transponierten, so folgt $Z'*Z = \Xi^2$:

5	4,82087552	1,97072534
4,82087552	7	4,32522272
1,97072534	4,32522272	5

Abbildung 42

Die Quadratsummen-Produktsummen Matrix Ξ^2, berechnet aus den ursprünglichen, unverfälschten Ausgangsdaten nach Abbildung 38 entspricht also der Matrix Z'*Z der veränderten Daten nach Abbildung 41. Quadratsummen und Produktsummen originärer Daten werden durch obig beschriebene Art der Datenmanipulation nicht verändert. Die so veränderten Daten in der (pxn) Matrix Z' können nun wieder auf neue Koordinatenachsen rotationstransformiert werden, sodass die sich neu ergebenden p Koordinaten eine sukzessiv maximale Varianz oder gleichbedeutend sukzessiv maximale Quadratsummen über die n Urteiler aufweisen. Dazu berechnen wir die Eigenwerte von Ξ^2, die sich zu $\lambda_1 = 13.4586$, $\lambda_2 = 3.0476$, $\lambda_3 = 0.4937$ ergeben. Die Summe der Eigenwerte entspricht der gesamten quadratischen Abweichung aller Itemvariablen und ergibt sich durch Summation der Diagonalelemente der Matrix Ξ^2 nach Abbildung 40 ($5+7+5 = 17 = \lambda_1+\lambda_2+\lambda_3$) zu 17. Hatte Item 1, Item 2 bzw. Item 3 ursprünglich quadratische Abweichungen von 5, 7 bzw. 5, so besitzen nun die rotationstransformierten, neuen Koordinaten sukzessiv maximale, quadratische Abeichungen von $\lambda_1 = 13.4586$, $\lambda_2 = 3.0476$, $\lambda_3 = 0.4937$. Der Eigenwert λ_i, der gleichsam die Quadratsumme der neuen, rotierten Koordinatenwerte über die n Urteiler bedeutet, ist stets ≥ 0. Dividiert man Ξ^2 durch n = 7 (durch die Ergänzungen der fehlenden Daten durch 0 besitzt jedes Item wiederum „7 Beurteilungen“ und auch die transformierten Daten weisen Fallzahlen von n = 7 auf), so erhalten wir die (pxp) Matrix $\frac{Z'*Z}{n}$.

0,71428571	0,6886965	0,28153219
0,6886965	1	0,61788896
0,28153219	0,61788896	0,71428571

Abbildung 43 Substitutionsmatrix für R

Diese so berechnete Matrix stellt nun die geeignete Hilfsgröße für den Fall fehlender Daten zur Berechnung der Faktorladungen dar. Die Eigenwerte dieser Matrix ergeben sich zu $\lambda_1 = 1.9227$, $\lambda_2 = 0.4354$, $\lambda_3 = 0.0706$. Sie entsprechen den Eigenwerten von Ξ^2 dividiert durch 7 und gleichsam den sukzessiv maximalen Varianzen der neuen, rotationstransformierten Koordinaten. Alle Eigenwerte obiger Matrix λ_i sind erneut ≥ 0. Während die Varianz-Kovarianzmatrix R normalisierter Ausgangsdaten, die gleiche Fallzahlen in den verschiedenen Variablen besitzen, stets Eigenwerte ≥ 0 besitzen, können unterschiedliche Fallzahlen in den Ausgangsdaten Verzerrungen von R, durch die unterschiedliche Gewichtung der Variablen bewirken und Eigenwerte kleiner 0 verursachen. So besitzt z.B. die Korrelationsmatrix nach Abbildung 39 eine Determinante < 0 und damit mindestens einen Eigenwert < 0.
Um Eigenwerte < 0 zu verhindern und mögliche Verzerrungen der Varianz-Kovarianzmatrix durch unterschiedliche Fallzahlen in den diversen Variablen zu umgehen, ersetzen wir zur Berechnung der Faktorladungen die Korrelationsmatrix nach Abbildung 39 durch das Substitut $\frac{Z'*Z}{n}$ bei unvollständigen Daten. Gleichung 57-Gleichung 63 gelten unverändert. Während der erste Spaltenvektor von $\widetilde{A}$ in Gleichung 62 für vollständige Daten in Z die korrelative Beziehung zwischen den (nx1) Spaltenvektoren $\underline{z_1}$, $\underline{z_2}$,... $\underline{z_p}$ von Z und dem ersten (nx1) Spaltenvektor der aus Z berechneten Faktorwertmatrix $\widetilde{F}$ angibt, unterscheidet sich nun die Kovarianzbeziehung nach Gleichung 62 für die mit 0-en ergänzte Z-Ausgangsmatrix von der Korrelationsberechnung entsprechender Variablen. Aufgrund der veränderten Varianzen der originalen Z-Daten durch die Ergänzung der fehlenden Daten mit einer 0.
Zum Schluss dieses Kapitels soll darauf hingewiesen werden, dass die beschriebenen 3 Arten des methodischen Zugangs, mit fehlenden Daten umzugehen, nicht auf den Fall der faktorenanalytischen Untersuchung beschränkt sind, sondern sinngemäß auf alle mit Fehldaten durchzuführenden Analysen Anwendung finden können.
So es die Menge der Daten sowohl innerhalb der verschiedenen Veranstaltungen als auch des Gesamtdatensatzes über die verschiedenen Veranstaltungen hinweg erlaubt, bildet die Vorgangsweise des listenweisen Fallausschlusses (1. Methode), die bevorzugte Vorgangsweise, da es aufgrund der gleichen Fallzahlen in den untersuchten Merkmalen zu keinen Verzerrungen der Varianz-Kovarianzmatrix kommt. In den verschiedenen Veranstaltungen muss der Datenumfang nicht nur die Berechnung von Mittelwert und Streuung zur Normalisierung der Werte ermöglichen, sondern die Fallzahlen müssen nach erfolgter Streichung unzureichend ausgefüllter Fragebögen aus der Veranstaltungsdatenliste vielmehr auch noch die Bedingung einer noch genügend großen Mindestrücklaufquote von Veranstaltungsbewertungen, die eine repräsentative und aussagekräftige Auswertung der Fragebögen sicherstellt, erfüllen. Dass heißt, die restlichen Fallzahlen je Veranstaltung nach einem listenweisen Fallausschluss müssen noch den geforderten Mindestrücklaufquoten

entsprechen.[51] Des weiteren sollte, wie bereits schon erwähnt wurde, die Fallzahl über alle Veranstaltungen hinweg mindestens das 3-fache der Itemanzahl der untersuchten Skala sein. Diese beiden Bedingungen sollten nach erfolgtem listenweisen Fallausschluss erfüllt sein, wobei die Repräsentativitätsforderung der diversen Veranstaltungsbewertungen durch eine entsprechende Mindestrücklaufquote sicherlich die schärfere zu erfüllende Voraussetzung darstellt. Die unter Punkt 2) beschriebene Methode des paarweisen Fallausschlusses wird aufgrund der unterschiedlichen Gewichtungen der Variablen und den damit einhergehenden möglichen Verzerrungen der Varianz-Kovarianz Matrix nicht empfohlen.
Sollten zuwenig Daten zur Verfügung sein, um die 1.) Methode durchführen zu können, sollte das Prinzip der Ergänzung der fehlenden Daten durch die Veranstaltungsmittel nach Punkt 3.) Verwendung finden.

3.5. Das Problem der Zusammensetzung und Repräsentativität von Stichproben:

Ziel von statistischen Analysen ist es, Kenntnis über bestimmte Merkmale einer interessierenden Grundgesamtheit zu erlangen. Unter Grundgesamtheit versteht man die Menge von allen potentiell untersuchbaren Einheiten, die über ein gemeinsames Merkmal oder eine gemeinsame Merkmalkombination verfügen. Dabei stellt sich nun die Frage, wie man denn zur Kenntnis dieser Merkmale kommt, da eine Prüfung aller Untersuchungsgegenstände oftmals einen Umfang einnehmen, der den Prüfer vor eine nicht zu bewältigende, wenn nicht gar unmögliche Aufgabe stellt. Daher werden in aller Regel nur die Merkmale eines Teilbereiches der Grundgesamtheit erhoben, um aus diesen Rückschlüsse auf die Verteilung der Merkmale in der Grundgesamtheit zu ermöglichen. Von entscheidender Bedeutung für diese Rückschlüsse ist es nun aber, wie sehr dieser untersuchte Teilbereich die Eigenschaften der Grundgesamtheit widerspiegelt, dass heißt wie sehr der Telbereich der Grundgesamtheit ähnelt bzw. diesen repräsentiert und wie groß der untersuchte Teilbereich sprich der Stichprobenumfang ist. Aussagen, wie sich das untersuchte Merkmal in der Grundgesamtheit darstellt, gewinnen dabei mit zunehmendem Stichprobenumfang an Präzision.
Stichproben sollten also „Miniaturabbildungen“ der zu untersuchenden Grundgesamtheit darstellen. Ob die Stichprobe die Grundgesamtheit treffend beschreibt, kann durch Vergleiche von Stichprobenmerkmalen mit für die Grundgesamtheit und Untersuchung charakteristischen Merkmalen verifiziert werden (z.B. Geschlechterverteilung, Altersverteilung, soziodemographische Merkmale,...). Wenn eine Erhebung von Merkmalen der Grundgesamtheit aufgrund einer unendlich großen Population nicht möglich ist oder doch zumindest nicht zumutbar erscheint, können die charakteristischen Merkmale der Untersuchungsstichprobe auch behelfsmäßig mit korrespondierenden Merkmalen einer bzw. mehrerer Vergleichsstichproben geprüft werden. Bei der Auswahl oder Zusammenstellung der Stichprobe muss darauf geachtet werden, dass die Stichprobe nicht durch systematische Fehler im Auswahlverfahren verzerrt wird.[52] Grundsätzlich kann davon ausgegangen werden, dass mit zunehmender Größe der Stichprobe grobe Unzulänglichkeiten in der Stichprobenzusammensetzung weitgehend vermieden werden und damit Verzerrungen der Stichprobenergebnisse hinsichtlich der

[51] Angaben zur vorgeschlagenen und geforderten Mindestrücklaufquote in Abhängigkeit von der Teilnehmerzahl der Veranstaltung können Kapitel 3.5 entnommen werden.
[52] Vgl. dazu auch Bemerkungen zur Auswahl der Repräsentationsmerkmale weiter unten. ...Weiters müssen die Repräsentationsmerkmale...

Aussagefähigkeit für die interessierende Grundgesamtheit unwahrscheinlicher werden.
Mit dem Terminus „charakteristische“ Merkmale für die Grundgesamtheit ist natürlich automatisch auch eine vermutete Relevanz dieser Merkmale auf die zu untersuchenden Kriterien verbunden. So scheint es einleuchtend und nicht Zielführend für eine Studie z.B. über „das Konsumverhalten der Österreicher“ die Stichprobe nach den Merkmalen „Anzahl der plombierten Zähne, Länge der Fingernägel,...“ repräsentativ hinsichtlich der österreichischen Grundgesamtheit zu gestalten, da mit großer Wahrscheinlichkeit diese Merkmale das Konsumverhalten der Österreicher nicht entscheidend beeinflussen und somit auch nicht die Aussagekraft der Untersuchung verbessern. Nicht die bloße Anzahl der Merkmale nach denen die Stichprobe „zusammengestellt“ wird, sondern die vermutete Relevanz der Merkmale für das Untersuchungsergebnis fördert die höhere Aussagekraft der Analyse.[53]
Es existieren nun mehrere Möglichkeiten bzw. Tests, die Zugehörigkeit einer Stichprobe zu einer Grundgesamtheit bzw. 2 Stichproben auf ihre Ähnlichkeit sprich Zugehörigkeit zu derselben Grundgesamtheit zu überprüfen. Betrachten wir zur Darstellung dieser verschiedenen Möglichkeiten zuerst den univariaten Fall, bei dem nur eine interessierende Variable (p = 1) über die Stichprobe erhoben wird. Besitzt diese Variable nun **Nominalskalenniveau** werden die Daten nach bestimmten Ausprägungen des Merkmals klassifiziert. Zur Bestimmung der Ähnlichkeit der Datenstruktur können nun die Häufigkeitsverteilungen der Stichprobe/Grundgesamtheit über die Merkmalsausprägungen (Klassenzugehörigkeiten) auf Unterschiedlichkeiten untersucht werden (χ^2-Test oder auch „goodness of fit test“). Natürlich können auch Merkmale, die über ein höheres Skalenniveau verfügen z.B. Variablen mit Intervallskalenniveau, mit dieser Methode auf Repräsentativität der Stichprobe getestet werden. Zu beachten ist dabei, dass sowohl die Information der „exakten“ Merkmalsausprägung der einzelnen, intervallskalierten Variable mit der bloßen Zuordnung in das jeweilige Klassenintervall aufgegeben wird, als auch die letztendlich willkürlich mögliche Festlegung der Klassenintervallgrenzen und damit einhergehend die Anzahl der Klassen den Test in erheblichem Maße beeinflusst.
Aus diesem Grund werden für **intervallskalierte** oder höher skalierte Daten gerne Methoden verwendet, die Stichprobengrößen mit Populationsparameter „vergleichen“ bzw. prüfen, ob die ermittelten Stichprobenkennwerte nur zufällig oder in signifikanter Weise vom vermuteten Populationsparameter abweicht (z.B. Mittelwertsvergleich). Bei diesen Parametertests werden alle Daten der Stichprobe zu einem Stichprobenkennwert aggregiert bzw. verdichtet. Zwar werden hier die einzelnen Daten mit ihrer „exakten“ Merkmalsausprägung bei der Berechnung des interessierenden Stichprobenkennwertes berücksichtigt, die Information der einzelnen Merkmalsausprägungen geht aber ebenfalls nach erfolgter Aggregierung verloren (z.B. Information die Verteilungsform betreffend geht verloren).
Im multivariaten Fall werden nun allgemein p für die Untersuchung relevant vermutete Repräsentativ-Merkmale (Variablen) über die Stichprobe erhoben und werden in K_j (j = 1..p) Klassen je Merkmal eingeteilt (Nominalstruktur der Daten als niedrigstes, allen Variablen gemeinsames Skalenniveau). Die Studenten verteilen sich nun mit einer gewissen Häufigkeit über die $\prod_{j=1}^{p} K_j$ möglichen

[53] Vgl. auch die Bemerkungen zur geschichteten Stichprobe in [8].

Klassenkombinationen der Merkmale. Diese empirisch festgestellte Merkmalsverteilung der Stichprobe auf die möglichen Klassenkombinationen ist dabei nicht nur von den theoretischen Verteilungen der einzelnen, verschiedenen Merkmale abhängig (z.B. normalverteilt -> $N(\mu_i, \sigma_i)$), sondern wird auch durch die gegenseitige, korrelative Beziehung zwischen den einzelnen Variablen (r_{ij}) bestimmt. Zur Veranschaulichung einer solchen p-variaten Merkmalsverteilung dient die in Abbildung 44 dargestellte theoretische Binormalverteilung (p = 2).

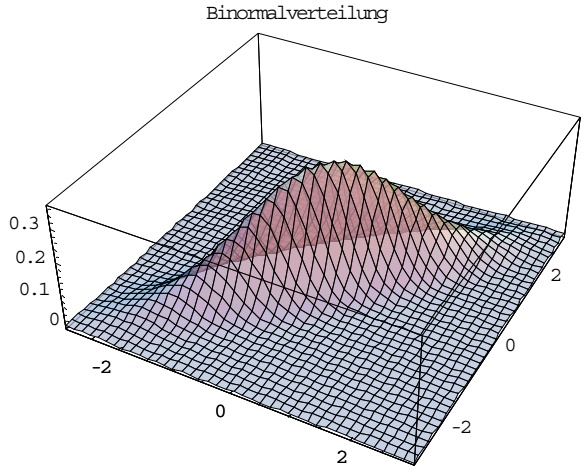


Abbildung 44 Binormalverteilung mit $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ und $r_{12} = 0.9$

Abbildung 45 zeigt einen Vergleich, wie sich eine binormalverteilte Zufallsstichprobe ($\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ und $r_{12} = 0.9$) der Größe n = 100, links theoretisch und rechts empirisch, auf die Klassen $i \le X < i+1$ und $j \le Y < j+1$ mit i,j = -3..2 verteilt.

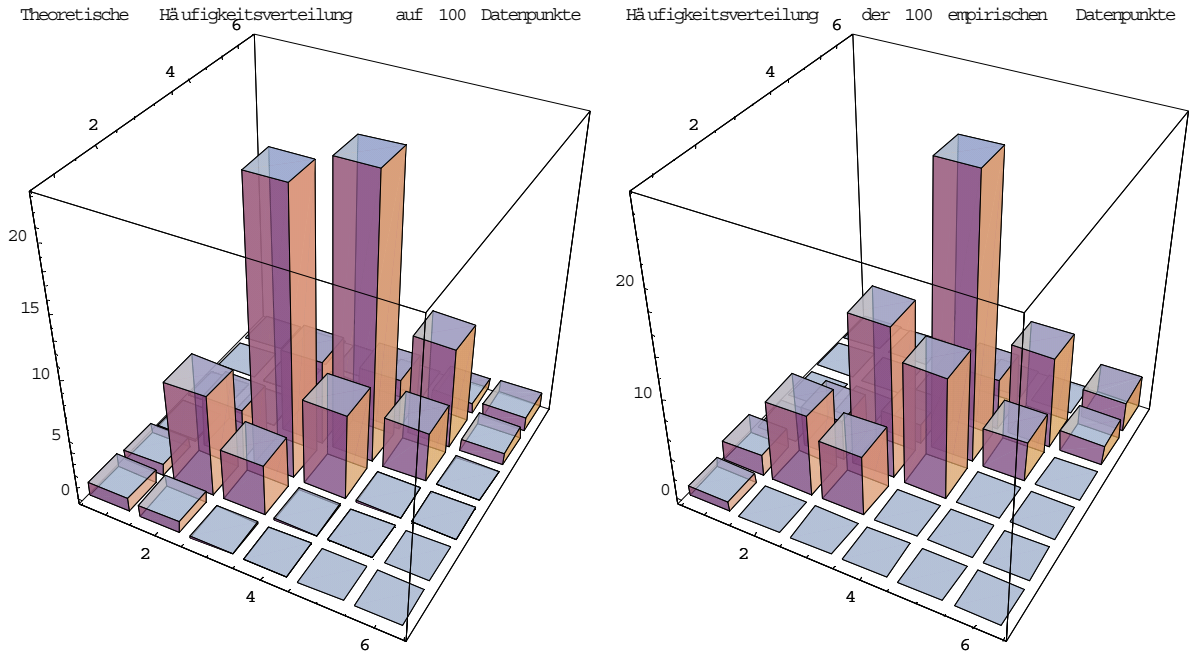


Abbildung 45: Links die theoretisch zu erwartende Verteilung und rechts die Verteilung der tatsächlich aufgetretenen Realisation der Zufallsstichprobe.

Auch im multivariaten Fall, könnten wiederum in Analogie zum weiter oben Gesagten[54] χ^2-Tests (nominale Datenstruktur) oder multivariate

[54] Vgl. Ausführungen weiter oben zum univariaten Fall, Es existieren nun...

Mittelwertsvergleiche (Hotellings T_1^2-Test oder Hotellings T_3^2-Test)[55], für intervallskalierte Daten Aufschluss über die Stichprobenrepräsentativität geben. Jedoch besteht die Restriktion, dass alle p Variablen für den multivariaten Mittelwertvergleich zumindest Intervallskalenniveau besitzen müssen und für den multivariaten χ^2-Test müssen die Häufigkeitsverteilungen über alle $\prod_{j=1}^{p} K_j$ Klassenkombinationen festgehalten werden, was mit zunehmender Merkmalsanzahl p und K_j-facher Stufung derselben einen nicht zu unterschätzenden Aufwand darstellt.
Um diese Probleme zu umgehen, begnügen wir uns damit die Stichprobe durch p univariate Teste (für jedes erhobene Merkmal separat) auf Repräsentativität zu kontrollieren. Diese Vorgehensweise ist nicht ganz unproblematisch, da beim univariaten χ^2-Test über nur eine interessierende Variable, die Randverteilung der multivariaten Verteilung auf Übereinstimmung mit der Grundgesamtheit geprüft wird. Informationen, wie sich dieses interessierende Merkmal auf die anderen Klassenkombinationen der übrigen (p-1) Merkmale verteilt, gehen verloren.
Abbildung 46 gibt beispielhaft Auskunft über die Randverteilungen X und Y. Wir erhalten die Randverteilung X, indem wir die Häufigkeiten aus Abbildung 45 rechts (Häufigkeiten der Zufallsstichprobe) bei einer festgehaltenen Klasse i ≤ X < i+1 (fixes i, i = -3..2) über alle möglichen Klassen j ≤ Y < j+1, j = -3..2 der Variable Y aufsummieren.

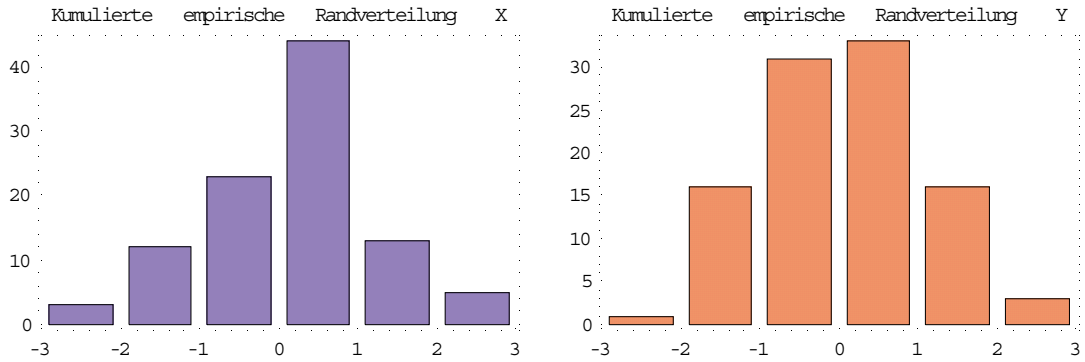


Abbildung 46 Randverteilungen X, Y einer binormalverteilten Zufallsstichprobe der Größe 100.

Selbiges gilt natürlich auch für die p-fache Durchführung univariater Mittelwertvergleiche, die im Gegensatz zum multivariaten Mittelwertvergleich die Information der gegenseitigen, korrelativen Zusammenhänge der p Variablen und damit die gegenseitige Abhängigkeit, wie sich die einzelnen Variablen über die anderen Variablen verteilen, nicht berücksichtigen.
Einer p-fachen, univariaten Prüfung der p erhobenen Merkmale anstelle einer multivariaten Prüfung kann nur unter den folgenden Randbedingungen zugestimmt werden:

- Die p Variablen sind zumindest theoretisch als wechselseitig unabhängig vorstellbar.
- Man möchte anhand der Untersuchung die Äquivalenz der Stichprobe mit einer zweiten Stichprobe oder Grundgesamtheit bezüglich möglichst vieler Variablen nachweisen.[56]

[55] Die p Variablen für die Hotellings T-Teste müssen in der Population multivariat normalverteilt sein., vgl. dazu [8].
[56] Vgl. [8]

Wir wollen davon ausgehen, dass die oben genannte Voraussetzung, der gegenseitigen Unabhängigkeit für die p erhobenen Variablen erfüllt sei. Als Repräsentativ-Merkmale der Beurteilungsstichprobe je Veranstaltung könnten z.B.:

- die durchschnittliche Note im Fach Mathematik im Maturazeugnis (intervallskaliert)
- die Zusammensetzung der Hörerschaft hinsichtlich der Art des Hochschulzuganges/ der besuchten Schule (Gymnasium, HTL, Studienberechtigungsprüfung, HAK, Abendmatura,...-nominalskaliert),
- die Geschlechterverteilung (nominalskaliert)

dienen.

Es ist nun vorstellbar, dass die Verteilung der Hörerschaft z. B. auf die verschiedenen Schulen nicht unabhängig und damit in verschiedener Weise für Studenten und Studentinnen ausfällt, somit eine gegenseitige Abhängigkeit der Variablen Geschlecht und Art des Hochschulzuganges vorliegt. Wir wollen dennoch diese Effekte getrennt überprüfen und annehmen, dass die gegenseitigen Abhängigkeiten (so vorhanden) sich in gewissen Grenzen halten (ebenso die mögliche Abhängigkeit von durchschnittliche Mathematiknote und Art der besuchten Schule). Würde eine perfekte lineare Abhängigkeit zwischen diesen 2 Größen existieren, so würde aus der Repräsentativität des einen Merkmals auch automatisch die Repräsentativität des anderen Merkmals folgen, dass heißt der 2. Test auf Repräsentativität würde seine Aussagekraft verlieren. Die Unabhängigkeit der Variablen Geschlecht-durchschnittliche Mathematiknote ist zumindest theoretisch vorstellbar.

Des weiteren soll angenommen werden dürfen, dass die Merkmale Mathematiknote, Hochschulzugang, sowie Geschlechterverteilung, als Variablen zum einen der punktuellen Erfassung der Fähigkeit des Studenten sich mit wissenschaftlich-technischer Materie zu befassen, der Hochschulzugang aufgrund unterschiedlicher schulischer Vorbildung und die Geschlechterverteilung mit möglichen Sympathieeffekten einen gewissen Einfluss auf die Evaluation des Vortragenden haben können und somit eine Überprüfung der repräsentativen Zusammensetzung der Beurteilungsstichprobe in diesen Merkmalen nachvollziehbar und legitim erscheint.

Weiters müssen die Repräsentationsmerkmale für die Stichprobe unabhängig von curricularen Faktoren ausgesucht werden. So wäre es wenig sinnvoll, Stichproben z.B. hinsichtlich der Altersverteilung auf charakteristische Zusammensetzung zu prüfen, da die Zusammensetzung der Stichprobe in den verschiedenen Vorlesungen im Gegensatz zur Grundgesamtheit durch Faktoren des Studienplans bedingt werden.

Um nun die Beurteilungsstichproben der verschiedenen Studenten in den unterschiedlichen Veranstaltungen auf typische, charakteristische und für die Studienrichtung repräsentative Zusammensetzung zu prüfen, können gleichsam Daten von Lehrberichten der Studienrichtung oder Immatrikulationsdaten als Referenzquelle Verwendung finden.

Die Überprüfung der Repräsentativität **kleiner bis mittlerer Veranstaltung**[57] ($5 \leq n \leq 50$) setzt sich nun aus 2 Teilen zusammen:

1. Hinsichtlich der verwertbaren Stichprobengröße der Veranstaltung,
2. Bezüglich der Stichprobenzusammensetzung der Urteiler anhand charakteristischer Merkmale.

[57] Zur Festlegung der Veranstaltungsgröße für „kleine bis mittlere Veranstaltungen“ vgl. Bemerkungen weiter unten zu $n_{opt.}$. ...Bei Bortz werden optimale Stichprobenumfänge...

Ad 1): Für die auswertbare Anzahl der zur Verfügung stehenden Daten der Studenten nach Ausschluss ungenügend beantworteter Fragebögen, wenn sie als repräsentativ und aussagekräftig in Abhängigkeit von der Teilnehmerzahl gelten soll, wird folgende Mindestrücklaufquoten in Anlehnung an Theall & Franklin nach Abbildung 47 je Veranstaltung gefordert:[58]

Teilnehmerzahl:	Rücklaufquote:
5-10	100%
11-20	80%
21-30	75%
31-50	66%
51-100	60%
>101	55%

Abbildung 47

Ad 2): Die Stichprobenzusammensetzung der einzelnen Vorlesungen wird im Falle Intervallskalierter Merkmale wie die durchschnittliche Mathematiknote mit dem univariaten Mittelwertvergleich der Stichprobe mit dem Populationsparameter vorgenommen. Wobei sich der Populationsparameter über die durchschnittliche Mathematiknote aller Studenten N für die jeweilige Studienrichtung berechnet.[59]

$$\mu_0 = \frac{\sum_{i=1}^{N} X_i}{N}$$

Gleichung 66 Erwartungswert der Population

Dieser Populationsparameter wird nun mit dem Mittelwert der Stichprobe der Größe n verglichen und auf signifikante Abweichung von diesem getestet.

$$(1) = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$$

mit

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ und } \sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu_0)^2}{N}}$$

Gleichung 67[60]

μ_0, σ kennzeichnen 2 Parameter (Erwartungswert und Standardabweichung) der Grundgesamtheit, die sich über alle N Studenten der interessierenden Studienrichtung berechnen (Daten aus Lehrbericht oder Immatrikulationsamt). $\bar{x}$ bezeichnet den Mittelwert berechnet über die Beurteilungsstichprobe der n Studenten.
Ausdruck (1) wird nun auf Signifikanz getestet. Für Fallzahlen n > 30 wird Ausdruck (1) mit $z_{crit.}$ der Standardnormalverteilung bei einem bestimmten Signifikanzniveau z.B. $\alpha = 25\%$ (ungerichtete H_1 -> $z_{crit.} = \pm 1{,}15$) verglichen. Liegt der Ausdruck (1) außerhalb $z_{crit.}$ liegt keine Repräsentativität der Stichprobe bezüglich des Merkmales

[58] Vgl. Rindermann S135.
[59] Die Verwendung inferenzstatistischer Methoden für unendliche Grundgesamtheiten, die eigentlich endlicher Natur sind, sind für praktische Zwecke immer dann anwendbar wenn für das Verhältnis $n/N \geq 100$ gilt. Diese Vorraussetzung kann jedoch bei Lehrevaluationen von Vorlesungen mittlerer Größe als gegeben angesehen werden. Vgl. dazu [8].
[60] Großbuchstaben (X, N) in den Berechnungsformeln stehen für Werte der Grundgesamtheit, Kleinbuchstaben (x, n) für solche der Stichprobe.

x vor. Für Fallzahlen n ≤ 30 wird Ausdruck (1) analog mit $t_{n,crit.}$ mit n Freiheitsgraden verglichen.
Für nominalskalierte Größen wie z.B. die Variable „Art des Hochschulzuganges“ werden die beobachteten Häufigkeiten des K_1-gestuften Merkmals (Gymnasium, HTL,...; p = 1 für univariate Prüfung) nach den Klassenzugehörigkeiten des Merkmals mit den erwarteten Häufigkeiten verglichen.
Die erwartete Häufigkeit in der Klasse j, j = 1..K_1 errechnet sich zu:

$$f_{e(j)} = \frac{N_j}{N} * n$$

mit

$$\sum_{j=1}^{K_1} N_j = N$$

Gleichung 68

n bzw. N stehen für die Größe der Stichprobe bzw. Grundgesamtheit (alle Studenten in der interessierenden Studienrichtung). N_j gibt die Häufigkeit der Studenten in der Klasse j der Grundgesamtheit wieder.
Die beobachtete Häufigkeit in der Klasse j, j = 1..K_1 ergibt sich zu:

$$f_{b(j)} = n_j$$

mit

$$\sum_{j=1}^{K_1} f_{b(j)} = \sum_{j=1}^{K_1} f_{e(j)} = n$$

Gleichung 69

Den χ^2 Wert erhalten wir nach Gleichung 70:

$$\chi^2 = \sum_{j=1}^{K_1} \frac{\left(f_{b(j)} - f_{e(j)}\right)^2}{f_{e(j)}}$$

Gleichung 70

Dieser χ^2 Wert wird nun mit einem kritischem Wert mit K_1-1 Freiheitsgraden bei einem bestimmten α-Niveau (z.B. α = 25% -> $\chi^2_{(K_1-1;75\%)}$) verglichen. Fällt der nach Gleichung 70 berechnete Wert kleiner als der kritische χ^2-Wert aus, so kann die Zusammensetzung der Stichprobe als repräsentativ hinsichtlich dieses Merkmals für die Grundgesamtheit angesehen werden.
Voraussetzungen für diesen eindimensionalen χ^2-Test sind, dass:

1. jede untersuchte Einheit eindeutig einer Kategorie zugeordnet werden kann,
2. die erwarteten Häufigkeiten in jeder Kategorie größer als 5 sind.

Für die Überprüfung der Stichprobe auf repräsentative Zusammensetzung muss noch folgendes angemerkt werden. Egal, ob es sich nun um eine intervallskalierte oder nominalskalierte Variable handelt, besteht die „Wunschhypothese in der Beibehaltung der H_0, dass heißt nicht der α-Fehler stellt die entscheidende Größe dar, sondern der β-Fehler. Dieser gibt nämlich an, wie groß die Wahrscheinlichkeit ist die H_0 zu akzeptieren, obwohl sie eigentlich falsch ist. Wir müssen uns also bei der Prüfung auf eine repräsentative Zusammensetzung der Stichprobe gegen den β-Fehler absichern, der den fälschlichen Schluss einer repräsentativen Zusammensetzung der Stichprobe im entsprechenden Merkmal gering hält. Da jedoch keine spezifische Alternativhypothese vorliegt, sind wir darauf angewiesen

den β-Fehler indirekt klein zuhalten, indem wir aufgrund der gegenläufigen Beziehung den α-Fehler vergrößern. Daher wurde in den obigen Tests entgegen der allgemein üblichen Konvention α = 25% anstatt α = 5% gesetzt.[61]
Die Forderung nach repräsentativen Veranstaltungen bezüglich der Beurteilungsstichprobe gewinnt dabei an Bedeutung je weniger Veranstaltungen beurteilt wurden. Ist die Anzahl der untersuchten und beurteilten Veranstaltungen genügend groß, kann zumeist auf eine Überprüfung aller Veranstaltungen auf Repräsentativität der Beurteilungsstichprobe aus auswertungsökonomischen Gründen verzichtet werden.[62] Dabei wird empfohlen, Vorlesungen beginnend mit den geringsten Besucherzahlen auf Repräsentativität der Stichprobe zu prüfen, bis die geforderte Mindestanzahl repräsentativer Vorlesungen erreicht wird.
Diese Forderung ist leicht zu verstehen, da mit zunehmender Größe der Stichprobe der Ausgang des Testes auf repräsentative Zusammensetzung derselben unwahrscheinlicher wird, dass heißt zunehmend zu ungunsten einer charakteristischen Zusammensetzung entscheidet.
Zur Veranschaulichung obigen Sachverhaltes betrachten wir den Test, der das Stichprobenmittel gegen einen bestimmten Populationsparameter (Mittelwertsparameter μ_0) testet etwas genauer.
Die Hypothese 0 kurz H_0 unterstellt nun, dass die Stichprobe aus einer Grundgesamtheit mit dem Mittelwertsparameter μ_0 gezogen wurde. Diese Hypothese wird nun gegen die konkurrierende Hypothese H_1 getestet, die besagt, dass die untersuchte Stichprobe einer Population angehört, deren Parameter μ vom Parameter μ_0 der Referenzpopulation abweicht (ungerichtet) oder kurz:

H_0: $\mu_0 = \mu$
H_1: $\mu_0 \neq \mu$

Die Entscheidung, ob die H_0 verworfen wird, hängt nun vom Ausdruck (1) in Gleichung 67 ab. Mit zunehmendem Stichprobenumfang wird nun aber der Ausdruck $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, bei endlicher Streuung der Grundgesamtheit (σ) immer kleiner ($\sigma_{\bar{x}} \to 0$ für $n \to \infty$), dass heißt Ausdruck (1) wird bei einer noch so kleinen und unbedeutenden Differenz des Mittelwertes $\bar{x}$ von μ_0 signifikant mit größer werdendem n.
Dazu Bortz aus „Statistik für Sozialwissenschaftler“: „Nimmt man ferner an, dass die H_0: $\mu_0 = \mu$ eine theoretische Fiktion ist (es ist unrealistisch anzunehmen, dass zwei verschiedene, real existierende Populationen exakt identische Mittelwertsparameter aufweisen), dürfte jede H_0 bei genügend großen Stichproben zu verwerfen sein. Die H_0 ist bei sehr großen Stichproben gewissermaßen chancenlos. Oder: Jede Alternativhypothese lässt sich als statistisch signifikant absichern, wenn man nur genügend große Stichproben untersucht.“
Um nun Abweichungen des Mittelwertes von einem bestimmten Populationsparameter nicht nur auf Zufälligkeit (Signifikanz), sondern auch auf eine bestimmte Bedeutung zu prüfen, wurde in der Statistik der Begriff der Effektgröße eingeführt.[63] Wie der Signifikanztest ist auch die Effektgröße eng mit dem Stichprobenumfang verbunden. Optimale Stichprobenumfänge zeichnen sich nun dadurch aus, dass unbedeutende Unterschiedseffekte zwischen dem Stichprobenmittelwert und dem Mittelwertsparameter der Grundgesamtheit nicht signifikant werden, bedeutende Unterschiede aber sich signifikant von μ_0

[61] Vgl. dazu [8]
[62] Als Richtzahl der Mindestanzahl an repräsentativen zu überprüfenden Veranstaltungen könnte die unter 3.4.2.3 mit Hilfe des Generalisierungskoeffizient berechnete Anzahl der beurteilten Veranstaltungen dienen, vgl. Wird nun dieser Koeffizient...
[63] Vgl. [8].

unterscheiden. Oder mit anderen Worten: Der optimale Stichprobenumfang ist gerade nicht so groß, dass praktisch unbedeutende Effekte signifikant werden und nicht so klein, dass praktisch bedeutende Effekte nicht signifikant werden.
Bei Bortz werden optimale Stichprobenumfänge für den Mittelwertsvergleich von $n_{opt.}$ = 20 für starke Effekte bis $n_{opt.}$ = 50 für mittlere Effekte angegeben. Ab einem Umfang $n_{opt.}$ = 310 werden auch schwache Effekte als signifikant unterschiedlich vom Populationsmittelwert angezeigt.[64]
Wir prüfen also wie oben angedeutet die kleineren bis mittleren Vorlesungen hinsichtlich beider Kriterien (Mindestrücklaufquote und Zusammensetzung) der Repräsentativität. So wird sichergestellt, dass signifikante Unterschiede in der Zusammensetzung auch mit bedeutenden Unterschieden mit entweder starken zumindest aber mit mittleren Effekten einhergehen.
Größere Veranstaltungen (n > 50) werden hingegen nur noch auf die Erfüllung der Mindestrücklaufquote nach Abbildung 47 überprüft.[65]
Der Vorteil in obig beschriebener Prüfordnung besteht nun darin, dass Veranstaltungen kleinerer Größe und damit verbunden ungenauerer Schätzung der Veranstaltungskennwerte bei nicht Erfüllung der Repräsentativitätsanforderungen sukzessive beginnend mit den kleinsten Vorlesungen ausgeschieden werden.

3.6. Die Multiple Regression als Instrument zur Aufdeckung möglicher Wirkzusammenhänge von Prozessdaten und Verzerrungsvariablen mit einer definierten Kriteriumsvariablen:

Die folgend vorgestellte Analysemethode fragt nun, ob und falls in welchem Maße denn Prozessdaten, Rahmenbedingungsvariablen und Biasvariablen mit einer zuvor definierten Kriteriumsvariablen zusammenhängen. Dabei stellt sich nicht die Frage des bivariaten Zusammenhanges der verschiedenen Variablen mit der Kriteriumsvariable als Korrelation, sondern wie groß sich dieser mögliche Zusammenhang mit der Kriteriumsvariablen im gegensätzlichen Zusammenspiel mehrerer beteiligter Prozess-/Rahmenbedingungsvariablen und Biasvariablen darstellt. Es wird also der Effekt der verschiedenen Variablen auf die zuvor festgesetzte Kriteriumsvariable gesucht, wobei nicht individuelle Rohdaten im Lichte des Interesses stehen, sondern Wirkzusammenhänge in aggregierter Form gesucht werden.[66]
Als Kriteriumsvariable soll hier z.B. in Übereinstimmung mit schon bestehenden Untersuchungen die **Arbeitsleistungen** sprich **Prüfungsleistungen** der Studenten als „reduziertes“ Kriterium für die zu messende Ausbildungsqualität gelten.[67]
Denn für die Prüfungsleistung gilt:
„Das hierfür maßgebliche Instrument ist der Leistungstest, der zudem weniger der Gefahr unwillkürlicher Fehler und Verzerrungen bzw. absichtlicher Fälschungen ausgesetzt ist, als ein Fragebogen, der fast nur mit „selbstbezogenen“ Auskünften

[64] Vgl. [8] Seite 137.
[65] Die ungenaueren Schätzungen der Veranstaltungskennwerte schlecht besuchter Veranstaltungen aufgrund ungenügender Kompensation interindividueller Unterschiede in der Beurteilung im Vergleich zu größeren werden durch die zusätzliche Forderung einer repräsentativen Zusammensetzung quasi egalisiert.
[66] Vgl. dazu auch die Bemerkungen unter 3.2 Bildungspolitisch relevante Indikatoren...
[67] Falls die durchschnittlichen Prüfleistungen über die verschiedenen Veranstaltungen sich als zu wenig variabel darstellt, kann auch zusätzlich nach dem selbst eingeschätzten Lernfortschritt gefragt werden. Die Kriteriumsvariable lässt sich dann z.B. gemeinsam als Mittelwert der Prüfleistung/Lernfortschritt berechnen. Allerdings muss dann zusätzlich die Frage nach dem Lernfortschritt in den Fragebogen (vgl. Abbildung 15) aufgenommen werden.

arbeitet und somit besonders stark von Erinnerungsvermögen und Aufmerksamkeit des Probanden abhängig ist, mithin kein „objektives“ Befragungsinstrument darstellt.“
Und:
„Der knappe Überblick über die Anfänge der Erhebung von Bildungsindikatoren auf einer empirischen Basis zeigt, dass schulische Arbeitsresultate die zentralen Indikatoren bzw. zentraler Bestandteil von Indikatorensystemen sowohl international als auch innerhalb bestimmter nationaler Grenzen waren und bis heute sind, dass in diesem Kontext standardisierte Tests einschließlich der Erhebung von Hintergrundmerkmalen wesentliche Quelle von Indikatorisierungen sind,...“[68]
Dieser multiple Regressionsansatz, der Veranstaltungsmittel der verschiedenen, interessierenden Prozessdaten den Veranstaltungsmitteln der Kriteriumsvariable in datenreduzierender Weise gegenüberstellt, kann folgendermaßen begründet werden.[69]

$$Y = K + X1 * b1 + ... + Xp * bp + U$$

$$\begin{pmatrix} Veranst.1 \\ Veranst.2 \\ . \\ . \\ . \\ Veranst.r \end{pmatrix} \begin{pmatrix} \bar{y}_{1,1} \\ \bar{y}_{1,2} \\ . \\ . \\ . \\ \bar{y}_{1,1t} \end{pmatrix} = \begin{pmatrix} k \\ k \\ . \\ . \\ . \\ k \end{pmatrix} + \begin{pmatrix} \bar{x}1_{1,1} \\ \bar{x}1_{1,2} \\ . \\ . \\ . \\ \bar{x}1_{1,1t} \end{pmatrix} * b1 + ... + \begin{pmatrix} \bar{x}p_{1,1} \\ \bar{x}p_{1,2} \\ . \\ . \\ . \\ \bar{x}p_{1,1t} \end{pmatrix} * bp + \begin{pmatrix} u_{1,1} \\ u_{1,2} \\ . \\ . \\ . \\ u_{1,1t} \end{pmatrix}$$

$$\begin{pmatrix} \bar{y}_{2,1} \\ \bar{y}_{2,2} \\ . \\ . \\ . \\ \bar{y}_{2,2t} \end{pmatrix} = \begin{pmatrix} k \\ k \\ . \\ . \\ . \\ k \end{pmatrix} + \begin{pmatrix} \bar{x}1_{2,1} \\ \bar{x}1_{2,2} \\ . \\ . \\ . \\ \bar{x}1_{2,2t} \end{pmatrix} * b1 + ... + \begin{pmatrix} \bar{x}p_{2,1} \\ \bar{x}p_{2,2} \\ . \\ . \\ . \\ \bar{x}p_{2,2t} \end{pmatrix} * bp + \begin{pmatrix} u_{2,1} \\ u_{2,2} \\ . \\ . \\ . \\ u_{2,2t} \end{pmatrix}$$

$$\begin{pmatrix} \bar{y}_{r,1} \\ \bar{y}_{r,2} \\ . \\ . \\ . \\ \bar{y}_{r,rt} \end{pmatrix} = \begin{pmatrix} k \\ k \\ . \\ . \\ . \\ k \end{pmatrix} + \begin{pmatrix} \bar{x}1_{r,1} \\ \bar{x}1_{r,2} \\ . \\ . \\ . \\ \bar{x}1_{r,rt} \end{pmatrix} * b1 + ... + \begin{pmatrix} \bar{x}p_{r,1} \\ \bar{x}p_{r,2} \\ . \\ . \\ . \\ \bar{x}p_{r,rt} \end{pmatrix} * bp + \begin{pmatrix} u_{r,1} \\ u_{r,2} \\ . \\ . \\ . \\ u_{r,rt} \end{pmatrix}$$

Y bezeichnet die Kriteriumsvariable X1,...,Xp die Prozessdaten, Rahmenbedingungsvariablen bzw. Biasvariablen, K eine Konstante und U stellt die stochastische Störgröße dar. b1,...,bp stellen die gesuchten, unbekannten b-Gewichte der multiplen Regression dar.

Gleichung 71

[68] Vgl. [4], [5].
[69] Siehe Gleichung 71.

Wie aus Gleichung 71 ersichtlich wird, werden die Veranstaltungsmittel der diversen in der Regression berücksichtigten Skalen bzw. Variablen entsprechend der Urteilerstichprobengröße der jeweiligen Veranstaltung aufgelistet. Wir entscheiden uns somit für den Ansatz, der mehr die messtechnische Schwierigkeit der Erfassung des „wahren" Wertes für das untersuchte Merkmal unterstreicht.[70] Für die Berechnung der Regression werden r evaluierte Veranstaltungen benötigt, wobei r > p+1 ist. p steht für die Anzahl der in der Regression berücksichtigten Prädiktorvariablen.
Die Prädiktorvariablen X1,..,Xp können dichotome Nominalskalen oder Intervallskalen sein. Ist Xj eine dichotome nominalskalierte Variable, die Charakteristika der Urteiler beschreiben (z.B. männlich/weiblich-0/1), so geht in die Regressionsrechnung das Veranstaltungsmittel als eine Größe der anteilsmäßigen (geschlechtlichen) Zusammensetzung der Beurteilungsstichprobe der jeweiligen Veranstaltung und betreffenden Variablen ein. Dichotome, nominalskalierte Eigenschaften der Veranstaltung werden durch zunächst willkürliche aber konstante Zuordnungen codiert (z.B. Pflicht- versus Wahlveranstaltung-0/1). Y ist eine intervallskalierte Kriteriumsvariable.
Da mittels der Regression hier nur empirische Zusammenhänge aufgedeckt bzw. analysiert werden und keine Prognosen von Arbeitsleistungen/Prüfungsleistungen erstellt werden sollen, werden die in die Regression eingehenden Variablen zuvor noch normalisiert.[71] Eine solche Normalisierung der Variablen birgt den Vorteil der besseren Vergleichbarkeit der verschiedenen Variablen, da Unterschiede der Skalierung in den verschiedenen Variablen kompensiert werden und damit auch eine bessere Vergleichbarkeit der b-Gewichte folgt, da diese ebenfalls in der zuvor normalisierten Regression frei von Varianzunterschiedlichkeiten der diversen Variablen und betragsmäßig alle samt ≤1 sind.
Es folgt daher:

$$Z = Z1 * b1 + ... + Zp * bp + U$$

Gleichung 72

Z steht für den ($\sum_{j=1}^{r} jt$) dimensionalen, normalisierten Spaltenvektor der Kriteriumsvariable Y. Z1,...,Zp sind die normalisierten Prädiktorvariablen. Die Konstante K wird aufgrund der Normalisierung zu 0 und fällt aus der Regressionsgleichung heraus. Jeder Spaltenvektor in Gleichung 72 besitzt nach erfolgter Normalisierung einen Mittelwert von 0 und eine Standardabweichung von 1.[72] Gleichung 72 kann für U = 0, r > p und r(Z1,...,Zp|Z) = p+1 nicht exakt gelöst werden.[73] Die multiple Regression bestimmt nun wiederum die b-Gewichte auf solche Art und Weise, dass für die vorhergesagten Werte für Z nämlich $\hat{Z}$ minimal quadratisch von diesen abweichen.

$$(Z - \hat{Z})' * (Z - \hat{Z}) = \min$$

Gleichung 73

[70] Vgl. auch die Ausführungen weiter oben unter 3.4.2.5-Die verschiedenen methodischen Ansätze...
[71] Die Mittelwerte für die Normalisierung der Variablen berechnen sich als mit den Besucherzahlen rt gewichtetes Mittel über die durchschnittlichen Veranstaltungsbewertungen der betreffenden Variablen.
[72] Bei der Normalisierung wird der nicht erwartungstreue Maximum Likelihood Schätzer der Standardabweichung verwendet, wodurch Übereinstimmung von Korrelation und Kovarianz der normalisierten Daten resultiert.
[73] r(Z1,...,Zp|Z) = p+1 bezeichnet den Rang der um die normalisierte Kriteriumsvariable erweiterte Prädiktorenmatrix.

Die Mittelung der diversen Skalen über alle Werte der verschiedenen Urteiler und Items einer Skala je Veranstaltung zur mittleren Prädiktorvariable der Veranstaltung kann folgendermaßen begründet werden.[74] Die Mittelung über die verschiedenen Skalenmittelwerte der verschiedenen Urteiler zum Veranstaltungsmittel der betreffenden Dimension stellt sicher, dass der gesuchte „wahre“ Wert der Beurteilung durch die Mittelung über die Urteiler in zunehmendem Maße unverzerrt im Sinne der Angleichung der verschienen Wertemaßstäbe dargestellt wird. Die Zusammenfassung der verschiedenen Itemwerte einer Skala zum Skalenmittelwert eines Urteilers kann durch die korrelativ belegte Ähnlichkeit der diversen Befragungsitems für das gesuchte Skalenmerkmal begründet werden. Außerdem wird durch die Zusammenfassung der ähnlichen Befragungsitems zu einem Skalenwert die Multikollinearität des Regressionsmodells der verschiedenen Prädiktorvariablen gering gehalten. Diese würde nicht nur eine zunehmend ungenauere Schätzung der b-Gewichte bewirken, sondern würde auch zu Verzerrungen der Teststatistiken und interpretatorischen Schwierigkeiten der b-Gewichte führen, da geringfügige Veränderungen der Multikollinearität zu drastischen Veränderungen der b-Gewichte führen können.[75]
Bezeichnen wir die normalisierte Prädiktormatrix mit ZP bestehend aus den Spaltenvektoren [Z1,...,Zp] = ZP, so folgt der KQ-Schätzer[76] für die b-Gewichte zu:

$$b = \left(ZP'*ZP\right)^{-1} * ZP'*Z = \begin{bmatrix} b1 \\ . \\ . \\ . \\ bp \end{bmatrix}$$

b bezeichnet den (px1) Spaltenvektor.

Gleichung 74

Für inferenzstatistische Absicherung der multiplen Korrelation wird gefordert, dass alle beteiligten Variablen multivariat normalverteilt sind. Die Überprüfung dieser Voraussetzung stößt auf das Problem, dass derzeit kein ausgereifter Test existiert, der alle möglichen Abweichungen von einer multivariaten Normalverteilung der Variablen gleich gut aufzeigt. Mehrere Behelfslösungen werden von unterschiedlichsten Autoren vorgeschlagen, die sich z.B. auf die Überprüfung von Schiefe und Exzess einer multivariaten Normalverteilung konzentrieren oder eine sequentielle Teststrategie unter Verwendung mehrerer Normalverteilungstests vorschlagen.
Die Überprüfung der multivariaten Normalverteilung der Variablen kann jedoch für einen Stichprobenumfang r > 40 bei p < 10 entfallen.[77]
Eine Überprüfung der multiplen Regression auf Signifikanz bezüglich der H_0: ρ = 0 oder anders ausgedrückt die Überprüfung, ob mindestens ein b-Gewicht signifikant von 0 verschieden ist, kann mit folgendem F-Test bewerkstelligt werden:

$$F = \frac{R^2 * \left(r - p - 1\right)}{\left(1 - R^2\right) * p}$$

Gleichung 75

[74] Siehe Abbildung 17.
[75] Vgl. [8]
[76] KQ-Schätzer bedeutet die Kleinstquadratschätzung der b-Gewichte nach Gleichung 73.
[77] Vgl. [8]

Ist obiger F-Wert größer als ein bei einem bestimmten α-Niveau (z.B. α = 1%) kritischer $F_{crit.}$-Wert mit p Zählerfreiheitsgraden und r-p-1 Nennerfreiheitsgraden so ist mindestens ein b-Gewicht signifikant von null verschieden.
Wobei R^2 das Bestimmtheitsmaß der multiplen Regression kennzeichnet oder anders ausgedrückt R den bivariaten Korrelationskoeffizienten zwischen der Kriteriumsvariablen Z und mittels der multiplen Regression vorhergesagten Werte $\hat{Z}$ darstellt.

$$R^2 = \frac{\hat{Z}'*\hat{Z}}{Z'*Z}$$

$$\hat{Z} = ZP * b$$

$$R = \frac{\hat{Z}'*Z}{\left(\left(\hat{Z}'*\hat{Z}\right)*\left(Z'*Z\right)\right)^{\frac{1}{2}}} = \sqrt{b1*r_{1,c} + ... + bp*r_{p,c}}$$

$$r_{i,c} = \frac{Z'*Zi}{\sum_r rt} \quad \text{für} \quad i = 1,..., p$$

Gleichung 76

Die Frage, welche Prädiktorvariable im Kontext der übrigen einen signifikanten Beitrag zur Vorhersage der Kriteriumsvariablen leistet (Signifikanztest der einzelnen b-Gewichte), kann mit folgendem Test überprüft werden:

$$t = \frac{bi}{\sqrt{\frac{r_{i,i}*(1-R^2)}{r-p-1}}}$$

wobei

$$\frac{\left(ZP'*ZP\right)^{-1}}{\sum_r rt} = \begin{bmatrix} r_{1,1} & r_{1,2} & . & . & . & r_{1,p} \\ . & . & & & & \\ . & & . & & & \\ . & & & . & & \\ . & & & & . & \\ r_{p,1} & . & . & . & . & r_{p,p} \end{bmatrix}^{-1}$$

$r_{i,i}$ das Diagonalelement obiger Matrix darstellt.

Gleichung 77

Der Betrag obigen t-Wertes sollte bei Vorliegen einer Signifikanz größer als ein kritischer $t_{crit.}$-Wert mit r-p-1 Freiheitsgraden sein.
Da oftmals das optimale Set an Prädiktorvariablen mit bester Varianzaufklärung weder mit der sogenannten Vorwärtstechnik oder Rückwärtstechnik noch mit einer Kombination beider Techniken gefunden werden kann, beschränken wir uns darauf nach erfolgter vollständiger Regression, bei der alle Variablen Berücksichtigung finden, jene Variablen, die keinen signifikanten Erklärungsbeitrag nach Gleichung 77 besitzen, aus der Regression auszuschließen.
Die so gefundene reduzierte Prädiktorenmatrix $ZP_{red.-mit\_Bias}$ kann dann dazu verwendet werden ein „subjektives“ Validitätsmaß zu kreieren. Berechnet man das Bestimmtheitsmaß $R^2_{mit\_Bias}$ nach Gleichung 76 mit der reduzierten Prädiktorenmatrix $ZP_{red.-mit\_Bias}$, die noch signifikante Biasvariablen enthalten kann, und zusätzlich das Bestimmtheitsmaß $R^2_{ohne\_Bias}$, das mit den restlichen Variablen ohne die Biasvariablen ermittelt wird, so folgt der Validitätskoeffizient zu:

$$r_{Val.} = \sqrt{1 - \frac{R^2_{mit\_Bias} - R^2_{ohne\_Bias}}{R^2_{mit\_Bias}}}$$

Gleichung 78

Dieser so berechnete Validitätskoeffizient besitzt leider noch die Unzulänglichkeit der Abhängigkeit der in die multiple Regression eingehenden Variablen. Er unterliegt sozusagen der subjektiven Auffassung des Analysten, welche Variablen erhoben werden und damit für wichtig genug erachtet werden.
Diese mit Hilfe der multiplen Regression gefundenen b-Gewichte geben nun erste Anhaltspunkte, wie die untersuchten Prädiktorvariablen mit der Kriteriumvariablen zusammenhängen. Hieraus können strategische, finanzpolitische Entscheidungen abgeleitet werden. Drückt beispielsweise ein entsprechend großes, positives b-Gewicht für die Rahmenbedingungsvariable der Veranstaltungsgröße mit der Kriteriumsvariablen eine Beziehung aus, dass größere Veranstaltungen schlechtere Prüfungsleistungen prognostizieren, können etwaige Korrekturmaßnahmen dieser Situation durch Vergrößerung des Lehrkörpers (Professoren-, Assistenten-, Dozentenschaft) gesetzt werden.[78]

3.7. Die kanonische Regression als Instrument der Analyse zur Aufdeckung von Zusammenhängen von Prozessdaten mit einem Set von Kriteriumsvariablen:

Während die multiple Regression den Zusammenhang zwischen mehreren Prädiktorvariablen und einer Kriteriumsvariablen überprüft, wird durch die kanonische Korrelationsanalyse die Beziehung zwischen mehreren Prädiktorvariablen und mehreren Kriteriumsvariablen ermittelt. Wird eine Reduktion der Qualität von Studium und Lehre auf das bloße Kriterium der Prüfungsleistung der Studenten, obzwar sich daraus eine einfachere statistische Handhabung ergibt, nicht erwogen und möchte man die Komplexität die Qualität des Ausbildungsprozesses zu messen und darzustellen durch weitere Kriterienvariablen einfangen[79], so müssen wir auf die von Hotelling entwickelte kanonische Korrelationsanalyse zurückgreifen, die den Zusammenhang zwischen 2 Variablenkomplexen aufdeckt. Eine Durchführung mehrerer multipler Korrelationen für jede Kriteriumsvariable führen nach Bortz zu Ergebnissen die den Gesamtzusammenhang der Variablenkomplexe unterschätzen.[80]
Es soll hier nur kurz das Grundprinzip dieser Analysemethode angerissen werden. Für genauere Informationen und die Herleitung der Berechnung wird auf das Lehrbuch von Bortz verwiesen.[81]
Bezeichnet hier Y die (nxq) Matrix der q normalisierten Kriteriumsvariablen

[78] Siehe auch Bemerkungen unter 3.2 Bildungspolitisch relevante Indikatoren und Evaluation bedeutet also... unter 4.
[79] Weitere Kriterienvariablen könnten z.B. die Studiendauer, die Höhe des ersten Bruttogehaltes bei Absolventen, die eigenen Einschätzungen oder Fremdeinschätzungen der Arbeitgeber der beruflich geforderten Fähigkeiten und Qualifikationen.
[80] Vgl. [8], so wie eine multiple Korrelation immer größer oder zumindest genau so groß ist wie die Einzelkorrelation, ist die kanonische Korrelation immer größer oder zumindest genau so groß wie die größte der einzelnen multiplen Korrelationen.
[81] Vgl. [8].

$$Y = \begin{bmatrix} y_{11} & y_{12} & . & . & . & y_{1q} \\ . & . & & & & \\ . & & . & & & \\ . & & & . & & \\ y_{n1} & . & . & . & . & y_{nq} \end{bmatrix} = [y_1, ..., y_q]$$

und X die (nxp) Matrix der p normalisierten Prädiktorvariablen,

$$X = \begin{bmatrix} x_{11} & x_{12} & . & . & . & x_{1p} \\ . & . & & & & \\ . & & . & & & \\ . & & & . & & \\ x_{n1} & . & . & . & . & x_{np} \end{bmatrix} = [x_1, ..., x_p]$$

Gleichung 79

so folgt, da die (nx1) Spaltenvektoren der entsprechenden Matrizen alle einen Mittelwert von 0 und über eine Standardabweichungen von 1 über die n Werte berechnet besitzen:

$$R_x = \frac{1}{n} * X'*X$$

$$R_y = \frac{1}{n} * Y'*Y$$

$$R_{xy} = \frac{1}{n} * X'*Y = R_{yx}'$$

Gleichung 80

R_x bezeichnet die (pxp) Korrelationsmatrix der Prädiktorvariablen, R_y die (qxq) Korrelationsmatrix der Kriterienvariablen und schließlich R_{xy} die (pxq) Korrelationsmatrix zwischen den Prädiktor- und Kriterienvariablen
Soll der kanonische Zusammenhang zwischen p Prädiktorvariablen und q Kriteriumsvariablen berechnet werden, ermitteln wir zunächst folgende Supermatrix von bivariaten Korrelationen:

$$R = \left[\begin{array}{c|c} R_X & R_{XY} \\ \hline R_{YX} & R_Y \end{array} \right]$$

Gleichung 81

Alle Kriterien- und Prädiktorvariablen sind intervallskaliert und werden multivariat normalverteilt vorausgesetzt.
Die weitere Vorgehensweise hat Gemeinsamkeiten mit der Hauptkomponentenanalyse unter 3.4.2.4. In der Hauptkomponentenanalyse wurden aus p Variablen diejenigen Linearkombinationen oder Faktoren bestimmt, die sukzessiv maximale Varianz aufklären, wobei die extrahierten Faktoren orthogonal sein sollen. Das kanonische Modell impliziert im Prinzip zwei getrennt durchzuführende Hauptkomponentenanalysen[82], wobei eine PCA über die Prädiktorvariablen und die andere über die Kriteriumsvariablen gerechnet wird. Während jedoch die erste Hauptachse in der PCA nach dem Kriterium der

[82] Auch als PCA bezeichnet. PCA steht für principal component analysis.

maximalen Varianzaufklärung festgelegt wird , werden in der kanonischen Korrelationsanalyse die ersten Achsen in den beiden Variablensätzen so bestimmt, dass zwischen ihnen eine maximale Korrelation, die als kanonische Korrelation bezeichnet wird, besteht.
Formal lässt sich das Problem folgendermaßen veranschaulichen: Aus dem Satz der Kriteriumsvariablen werden Linearkombinationen (Rotationstransformation)[83] $\hat{y}_m$ (m = 1,...,n) bestimmt, die maximal mit den aus den Prädiktorvariablen linear kombinierten $\hat{x}_m$-Werten korrelieren.

$$\begin{aligned}
\hat{y}_1 &= w_1 * y_{11} + w_2 * y_{12} + ... + w_q * y_{1q} \\
\hat{y}_2 &= w_1 * y_{21} + w_2 * y_{22} + ... + w_q * y_{2q} \\
&\cdots \\
\hat{y}_n &= w_1 * y_{n1} + w_2 * y_{n2} + ... + w_q * y_{nq}
\end{aligned}$$

und

$$\begin{aligned}
\hat{x}_1 &= v_1 * x_{11} + v_2 * x_{12} + ... + v_p * x_{1p} \\
\hat{x}_2 &= v_1 * x_{21} + v_2 * x_{22} + ... + v_p * x_{2p} \\
&\cdots \\
\hat{x}_n &= v_1 * x_{n1} + v_2 * x_{n2} + ... + v_p * x_{np}
\end{aligned}$$

Gleichung 82

Das obere Gleichungssystem bezieht sich auf die q Kriteriumsvariablen, unteres auf die p Prädiktorvariablen. Die kanonische Korrelation ist nichts anderes als die Produkt-Moment-Korrelation zwischen den $\hat{x}$-Werten und den $\hat{y}$-Werten.
Die Lösung des Problems läuft auf die Ermittlung der Eigenwerte der folgenden nicht symmetrischen quadratischen Matrix hinaus:

$$\left(R_x^{-1} * R_{xy} * R_y^{-1} * R_{yx} - \lambda^2 * I\right) * v = 0$$

Gleichung 83

Die Anzahl der kanonischen Korrelationen entspricht der Anzahl der Variablen im kleineren Variablensatz min(p,q) = g. Wird die Determinante obiger Matrix null gesetzt, folgt ein Polynom max(p,q)-ter Ordnung, das min(p,q) nicht-negative Lösungen hat. Diese Eigenwerte sind die Quadrate der kanonischen Korrelationen λ_s^2 (s = 1,..,g).
Die Frage, ob der durch alle g kanonischen Korrelationen erfasste Gesamtzusammenhang der beiden Variablensätze statistisch bedeutsam ist, überprüfen wir mit folgendem Test:

$$V = -[n - \frac{3}{2} - \frac{(p+q)}{2}] * \sum_{s=1}^{g} \ln(1 - \lambda_s^2)$$

Gleichung 84

[83] Siehe auch Gleichung 42 (rotationstransformierter erster Spaltenvektor von Y sprich erster extrahierte Faktor).

Der V-Wert ist mit p*q Freiheitsgraden approximativ χ^2-verteilt. Wurden bereits t kanonische Korrelationen bestimmt, überprüfen wir mit, ob die verbleibende Kovarianz noch signifikant ist:

$$V_t = -[n - \frac{3}{2} - \frac{(p+q)}{2}] * \sum_{s=t+1}^{g} \ln(1-\lambda_s^2)$$

Gleichung 85

Dieser V_t-Wert hat (p-t)*(q-t) Freiheitsgrade. Ist V_t nicht signifikant, sind nur die ersten t kanonischen Korrelationen statistisch bedeutsam, und die übrigen (g-t) kanonischen Korrelationen müssen auf stichprobenbedingte Zufälligkeit zurückgeführt werden.
Es existieren mehrere Kennwerte, die die Bedeutsamkeit des empirisch gefundenen Zusammenhanges der kanonischen Korrelation beschreiben. Wir empfehlen die Verwendung der von Cohen vorgeschlagenen „set-correlation“ R_{xy}^2, die den Gesamtzusammenhang zweier Variablensätze als verallgemeinerte, gemeinsame Varianz widerspiegelt.

$$R_{xy}^2 = 1 - (1-\lambda_1^2) * (1-\lambda_2^2) * ... * (1-\lambda_g^2)$$

Gleichung 86

Finden wir noch zu den Eigenwerten λ_s^2 nach Gleichung 83 die zugehörigen Eigenvektoren v_s (s = 1,...,min(p,q) = g) und normieren diese, dass gilt:

$$v_s' * R_x * v_s = 1$$

Gleichung 87

so berechnen sich die entsprechenden Eigenvektoren w_s der Kriteriumsvariablen für s = 1,...,g nach:

$$w_s = \frac{1}{\lambda_s} * R_y^{-1} * R_{xy}' * v_s$$

Gleichung 88

Ordnet man noch die so gefundenen Vektoren in einer Matrix an, so folgt:

$$V = [v_1,...,v_g]$$

und

$$W = [w_1,...,w_g]$$

Gleichung 89

Die Faktorwerte der Prädiktor- und Kriteriumsvariablen folgen zu:

$$F_x = \hat{X} = X * V = [\hat{x}_1,...,\hat{x}_g]$$

$$F_y = \hat{Y} = Y * W = [\hat{y}_1,...,\hat{y}_g]$$

Gleichung 90

Wir erhalten die kanonischen Korrelationen λ_s für s = 1,...,g wie weiter oben schon angedeutet zu:

$$\frac{\hat{X}'*\hat{Y}}{n} = \begin{bmatrix} \lambda_1 & 0 & . & . & . & 0 \\ 0 & \lambda_2 & & & & \\ 0 & & . & & & \\ . & & & . & & \\ . & & & & . & \\ 0 & & & & & \lambda_g \end{bmatrix}$$

Gleichung 91

Möchte man noch die Ladungen der einzelnen Variablen auf den entsprechenden Faktoren wissen, so erhält man diese über:

$$A_x = R_x * V$$

und

$$A_y = R_Y * W$$

Gleichung 92

Die kanonische Korrelation stellt den allgemeinen Lösungsansatz dar. Wie im Lehrbuch von Bortz gezeigt wird, folgt aus dem Signifikanztest nach Gleichung 84 für q = 1 der Signifikanztest der multiplen Korrelation nach Gleichung 75. Für p = 1 und q = 1 kann gezeigt werden, dass der Signifikanztest für eine einfache Produkt-Moment-Korrelation folgt, somit beides als Spezialfälle in der kanonischen Korrelation enthalten ist.

4. Plädoyer gegen die Vergleichbarkeit oder das Problem der Abnahme der Produkte durch die Professorenschaft und Überprüfung der Prozessdaten durch die studentische Hörerschaft:

Wie Eingangs schon erwähnt wurde besteht zumindest die zwei geteilte Aufgabe des Qualitätssicherungsprozesses Studium und Lehre bei bloßer Konzentration auf den Wissenserwerb bzw. Wissensvermittlung in eben diesen Bereichen zu prüfen. Der Professorenschaft obliegt es Qualitätsstandards hinsichtlich der Abnahmekriterien der Produkte festzusetzen, welche sich in einer mindestens zu erreichenden Prüfungsleistung für die Studenten manifestiert. Der Anfangs gestellte Vergleich der Qualitätssicherung von Studium und Lehre mit der Qualitätssicherung eines beliebigen Fertigungsprozesses, mit der die Notwendigkeit der getrennten Qualitätsprüfung hinsichtlich dieser Bereiche nämlich der Produkte und der Prozessdaten veranschaulicht wird, hinkt allerdings, wenn es um die Objektivität der Messergebnisse geht. So werden z.B. in der Werkstoffkunde Probestäbe auf Zugfestigkeit geprüft, um aus einer Reihe von Versuchen Rückschlüsse auf die Festigkeitswerte der untersuchten Legierung schließen zu können. Während eine Probe bei einem Festigkeitsversuch aufgrund der Zerstörung derselben bezeichnender Weise nur einmal geprüft werden kann, so besteht bei Prüfungsleistungen das Problem, dass diese nicht nur innerhalb der Produkte sprich von Student zu Student variieren, sondern dass diese auch eine unterschiedliche Beurteilung eines Probanden über verschiedene Prüfer erfahren können und dies in der Regel auch geschieht. Diese also mehr einem subjektiven Charakter unterliegen.

Gleiches trifft selbstverständlich auf die studentische Beurteilung über die Art und Weise der stattgefundenen Wissensvermittlung zu. Diese unterliegt zwar in geringerem Ausmaße der Subjektivität der Urteilsfindung durch die interindividuelle Mittelung der verschiedenen Urteilermaßstäbe, dennoch ist auch die studentische Veranstaltungsbewertung abhängig von der Zusammensetzung der Hörerschaft. Eine Vergleichbarkeit der dozentenbezogenen Veranstaltungsbewertung über unterschiedliche Fächer oder über korrespondierende Fächer ist aufgrund der unterschiedlichen Zusammensetzung der Hörerschaft und damit einhergehend die Möglichkeit eines von Hörerschaft zu Hörerschaft verschiedenen Wertemaßstabes für gleiche Leistung von vornherein nicht statthaft.
Vergleichbarkeit der Dozentenbewertung wäre ausschließlich bei gleichzeitiger Beurteilung korrespondierender Hörerschaften über mehrere Veranstaltungen möglich. Damit würde sichergestellt, dass gleiche Leistungen des Lehrkörpers von fachbezogenen Studenten beurteilt werden und über die Analyse mehrerer Veranstaltungen Unterschiedlichkeiten der mittleren Bewertungshöhe der verschiedenen Hörerschaften aufgedeckt werden könnten. Des weiteren besteht das anders geartete Problem, dass für eine korrekte Vergleichbarkeit der Lehrleistung ähnlichen Stoffinhaltes folgende Vorraussetzungen erfüllt sein müssten:

- Verwendung einer einheitlichen Skala in der Evaluation zur Messung der Untersuchungsmerkmale,[84]
- Verwendung einer annähernd einheitlichen Qualität und Dimensionalität der erhobenen Daten,[85]
- Verwendung einheitlicher statistischer Modelle und Berechnungsvorschriften,
- Verwendung einheitlicher Standards hinsichtlich der Repräsentativität der Stichproben,
- Verwendung einer Mindestanzahl an repräsentativen Lehrevaluationen um eine zuverlässiges Beurteilung des Dozenten zu sichern,[86]

und

- Verwendung und Berücksichtigung urteilverzerrender Effekte und Filterung dieser Biasgrößen aus den mittleren Dozentenbeurteilungen über eine multiple Regression,[87]

Zum letztgesagtem Punkt ist noch folgendes zu sagen. Die betreffende Lehrvariable als Kriteriumsvariable müsste also noch um die verzerrenden Effekte bereinigt werden. Dazu schätzen wir mittels der multiplen Regression die b-Gewichte möglicher Biasvariablen über die repräsentativen Veranstaltungen.[88]
Formal entspricht der Ansatz Gleichung 71, wobei Y die Kriteriumsvariable sprich eine der entsprechenden Lehrvariablen darstellt und als Prädiktoren die Biasvariablen X1,...,Xp in die Gleichung eingehen. Wir erhalten die b-Gewichte zu:

[84] In dieser Arbeit wurde eine siebenstufige Bewertungsskala mit den Endpolen völlig unzutreffend (1) bis völlig zutreffend (7) vorgeschlagen.
[85] Vgl. 3.4.1 und 3.4.2.4 zur Dimensionalität und 3.4 und 3.5 zur Überprüfung der Qualität der erhobenen Daten.
[86] Vgl. Ausführungen zum Generalisationskoeffizienten unter 3.4.2.3
[87] Kriteriumsvariable stellt die Dozentenbewertung der interessierenden Skala dar. Dabei werden wiederum die Veranstaltungsmittel entsprechend der Größe der Hörerschaft aufgelistet (vgl. dazu die Ausführungen zu Gleichung 71). Die Lehrleistungen werden bezüglich der Biasvariablen residualisiert.
[88] Die Mindestanzahl an repräsentativen Veranstaltungen ergibt sich wiederum aus dem Generalisierungskoeffizienten.

$$b = \left(XP'*XP\right)^{-1} * XP'*Y = \begin{bmatrix} K \\ b1 \\ b2 \\ . \\ . \\ . \\ bp \end{bmatrix}$$

und

$$XP = \left[1, X1, X2, ..., Xp\right]$$

Gleichung 93

XP bedeutet dabei die ($\sum_{r=1}^{r} rt$ x(p+1)) dimensionale Prädiktormatrix der Biasvariablen.

Wir erhalten mit den gefundenen b-Gewichten nach Gleichung 93 die von den Verzerrungseffekten befreiten Lehrvariablen Y_{ohne_Bias} zu:

$$Y_{ohne\_Bias} = Y - X1*b1 - ... - Xp*bp = K + U$$

Gleichung 94

Für eine Vergleichbarkeit der Prüfungsleistungen müssten folgende Vorraussetzungen erfüllt sein:

- Verwendung einer einheitlichen Testauswertung,
- Überprüfen des Leistungstests auf objektive Aufgaben und auf Objektivität des Tests im Ganzen,
- Verwendung desselben oder aber mindestens eines gleichwertigen Leistungstests für die Probanden,

und

- Schulung der Testauswerter auf einheitliche, gleiche Beurteilung der Testaufgaben.[89]

Wären oben genannte Bedingungen erfüllt, wäre zumindest theoretisch eine Vergleichbarkeit, sowohl der Prüfungsleistungen als auch der Lehrleistungen möglich. Eine solche Vorgehensweise würde aber eine übergeordnete, unabhängige Stelle zur Überprüfung und Einhaltung obig genannter Vorraussetzungen quasi eine Institutionalisierung erfordern, was aber innerhalb der verschiedenen Institute zu Widerständen führen könnte.

Eine solche Vorgehensweise scheint aufgrund des damit verbundenen enormen Aufwandes oder besser gesagt praktischen Unmöglichkeit, dass koresondierende Hörerschaften in ein und derselben Vorlesung sitzen, und der geringen Aussagekraft über dozentische Vergleiche für die Qualität der Vorlesung nicht ratsam. Schlechter bewertete Dozenten könnten die Evaluation durch die Studenten in Frage stellen und die Sinnhaftigkeit des methodischen Ansatzes ablehnen.

Eine direkte Vergleichbarkeit sowohl der Prüfungsleistungen als auch der Lehrleistung über unterschiedliche Fächer als auch über korrespondierende Fächer hinweg ist demnach nicht „gegeben“, aber auch nicht notwendig. So soll es ja gerade dem jeweiligen Professor, der gewissermaßen als Experte für seinen Fachbereich gelten soll, in Absprache mit den entsprechenden wissenschaftlichen und wirtschaftlichen Gremien „frei“ stehen, die notwendigen Qualitätsansprüche

[89] Vgl. dazu Kapitel 3.3 Grundsätzliche Begriffsbestimmungen und Gütekriterien in der Testanalyse:

festzusetzen. Ebenso verhält es sich mit der Lehrevaluation durch die Hörerschaft. Eben diese, welche auch die Adressaten des Vortrages sind, sollen die Vorlesung hinsichtlich der Verständlichkeit und logischer Einbettung der Vorlesung in den Lehrplan bewerten.
In den Kapiteln 3.6 und 3.7 wurden Analysen zur Aufdeckung von Zusammenhängen der Produktdaten (Kriteriumsvariablen) mit Prozessdaten, Rahmenbedingungsvariablen und Biasvariablen vorgestellt. Da die funktionale Beziehung der verschiedenen Variablen unbekannt ist, begnügen wir uns darauf den Einfluss der Prädiktorvariablen mittels der b-Gewichte auf die Kriteriumsvariablen linear zu schätzen.[90] Wie schon angedeutet wurde, sollten vor allem für jene Prozessvariablen und Rahmenbedingungsvariablen, die über die größten b-Gewichte verfügen, Aussagen über zu erreichende Werte getroffen werden. Diese b-Gewichte sollen jedoch nur ein Anhaltspunkt sein mindestens für jene Variablen mit hohen b-Gewichten Richtwerte festzusetzen.[91] Um eine Benachteiligung von Variablen mit niedrigen b-Gewichten zu verhindern, können selbstverständlich Prozessvariablen per se als Zielgrößen definiert werden, da kein empirisch signifikanter Zusammenhang der Kriteriumsvariable mit der Prädiktorvariable sprich Indikators (b-Gewicht) noch lange kein Garant für einen qualitativ unwirksamen Indikator ist. Dies bedeutet lediglich, dass die Prozess- oder Rahmenbedingungsvariable keinen empirischen Zusammenhang mit der bestimmten Kriteriumsvariablen besitzt, die jedoch nur unzulänglich alle möglichen Faktoren, die auf die Qualität von Studium und Lehre Einfluss haben, abdeckt.[92]
Wir fragen nun also nach der Art des zu verwendenden bzw. anwendbaren Maßstabes und nach der Festlegung, welche Werte für die entsprechenden Skalen erreicht werden sollten, um Qualität der Lehre ähnlich den Qualifikationsetappen für Studenten, die zumindest die Prüfung positiv absolvieren müssen, auch in quantitativer Hinsicht festschreiben zu können. Der unter Kapitel 3.1 angeführte soziale Maßstab wird aufgrund des oben gesagten und der Schwierigkeiten, die sich über eine korrekte Vergleichbarkeit ergeben, nicht zur Anwendung kommen. Betrachten wir, um uns den folgenden Gedankengang klarer zu machen, eine beliebige Lehrverhaltensskala (z.B. s1), die über eine beliebige Veranstaltung r (r=1) ermittelt wird:

			Urteiler 1,1	Urteiler 1,2	Urteiler 1,3
Veranst. 1	S 1	Item 1	5	4	6
		Item 2	5	4	7
		Item 3	4	3	5

			Urteiler 1,1	Urteiler 1,2	Urteiler 1,3	$\bar{y}_1$
Veranst. 1	S 1		4,66666667	3,66666667	6	4,77777778

Abbildung 48

[90] Diese Schätzungen der b-Gewichte werden bei Verletzung der linearen Abhängigkeit zwischen Kriteriumsvariablen und Prädiktoren und zunehmender Streuung der Prädiktorvariablen unbrauchbarer.
[91] Als hohe b-Gewichte können z.B. als signifikant von 0 verschiedene b-Gewichte gelten.
[92] Gleiches gilt natürlich für die kanonische Regression, wobei hier davon auszugehen ist, dass durch die unterschiedlichen Kriteriumsvariablen im zu untersuchenden Set die verschiedenen Faktoren, die auf die Qualität von Studium und Lehre wirken, zumindest besser abgebildet werden. Vgl. dazu auch die Bemerkungen unter 3.2, Ein Versteifen auf einige, wenige Qualitätsindikatoren...

Wir gehen davon aus, dass die Urteiler in der betreffenden Veranstaltung alle Itemwerte der interessierenden Lehrverhaltensskala beantwortet haben.[93] Wir mitteln die Itemwerte eines Urteilers und erhalten die untere Tabelle in Abbildung 48. Mitteln wir diese durchschnittlichen Skalenwerte noch über die entsprechende Urteilerzahl 1t, so folgt die durchschnittliche Veranstaltungsbewertung (r=1) der Skala 1, die mit der Kriteriumsvariablen $\bar{y}_1$ der Veranstaltung 1 nach Gleichung 71 übereinstimmt.
Werden die durchschnittlichen Veranstaltungswerte noch bezüglich der Verzerrungseffekte nach Gleichung 94 befreit, erhalten wir, wie weiter oben schon beschrieben, Y_{ohne_Bias} und damit für die erste Veranstaltung $\bar{y}_{1,ohne\_Bias}$
Diese so erhaltenen Werte über die verschiedenen Veranstaltungen der betreffenden dozentischen Skala stellen nun den Ausgangspunkt für den unter Kapitel 3.1 dargestellten **ipsativen** Maßstab dar. Das Vorliegen mehrerer auf diese Weise ermittelter Befunde können Anhaltspunkt qualitativer Entwicklung dozentischer Lehrverhaltensweise im Zeitreihenvergleich liefern. Auch können aus der Höhe mehrerer Bewertungen erstmals Rückschlüsse auf die durchschnittliche Leistung im betreffenden Bereich gezogen werden und diese mit zukünftigen Leistungen verglichen werden.[94]
Des weiteren könnte in Übereinstimmung mit studentischen Prüfungsleistungen ein **kriteriumsbezogener** Maßstab zur Anwendung kommen, der unterdurchschnittliche Leistungen als ungenügend qualifiziert. Da eine positive Bewertung studentischer Leistung eine zufriedenstellende oder richtige Beantwortung von mehr als der Hälfte aller an ihn gestellten Aufgaben erfordert, könnte somit eine unterdurchschnittliche Bewertung, die sich bei Verwendung einer Bewertungsskala von 1...n=7 durch

$$\bar{y}_{1,ohne\_Bias} \leq \frac{n+1}{2}$$

Gleichung 95

abzeichnet, als Aufforderung verstanden werden, Schritte zur Steigerung der Vortragsqualität in eben diesem Bereich zu veranlassen.
Eine interessante Ergänzung zum oben Angeregten könnte durch das Miteinfließen der Varianzen in die Bewertung geschehen. Diese kommt dadurch zustande, dass z.B. eine mittlere Bewertung von 4, die sich aus den zwei Wertungen 7 und 1 ergibt sicherlich negativer zu bewerten ist als eine gleiche mittlere Bewertung, die sich aus den zwei Wertungen 4 und 4 ergibt. Denn es herrscht im einen Fall (7 und 1) eine größere Unsicherheit bezüglich der „wahren“ Leistung und muss im Sinne einer Qualitätssicherung für das System Studium und Lehre als kritischer sprich eher als vermeintliche Minderleistung angesehen werden.
So könnte Gleichung 95 folgend ergänzt werden:

$$\frac{\bar{y}_{1,ohne\_Bias}}{1+\frac{s_{1,s1}^2}{s_e^2}} \leq \frac{\frac{n+1}{2}}{1+\frac{s_g^2}{s_e^2}}$$

In dieser Gleichung bedeutet:
$s_{1,s1}^2$ = Die Varianz der Skalenmittelwerte berechnet über die 1t Urteiler der interessierenden Veranstaltung (r=1, erster Index) und betreffenden dozentischen Skala (s1, zweiter Index).[95]

[93] Wie mit fehlenden Daten umgegangen wird, vergleiche Kapitel 3.4.2.5 unter **Missing Value-Problem**...
[94] Siehe Bemerkungen weiter unten zu Obige Überlegungen der Einbeziehung der Varianz...
[95] Siehe Abbildung 48 unten.

s_e^2 = Die größtmögliche und damit extremste Varianz bei Verwendung einer Skala von 1...n.
s_g^2 = Die Varianz einer gleichverteilten, diskreten Zufallsvariablen von 1...n.

Gleichung 96

Gleichung 96 fordert also, dass zumindest für eine mittlere Lehrveranstaltungsbewertung $\bar{y}_{1,ohne\_Bias}$ von 4, die Varianz der einzelnen Urteilerschätzungen der betreffenden Skala und Veranstaltung zumindest kleiner sein muss als eine gleichverteilte Varianz. Oder anders ausgedrückt: Es wird zumindest gefordert, dass sich die mittlere Beurteilung nicht aus Werten zusammensetzt, die alle Werte der Skala in gleicher Weise berücksichtigt und damit eher ein unklares Bild der Beurteilung suggeriert, die Bewertungen also ein Mindestmaß an Übereinstimmung oder Homogenität erfüllen sollten.
Da sich die gemittelte, dozentische Veranstaltungsbewertung (für z.B. r=1) $\bar{y}_1$ der interessierenden Skala nur bezüglich eines pro Veranstaltung unveränderten Verzerrungseffekt von $\bar{y}_{1,ohne\_Bias}$ unterscheidet, besitzen $\bar{y}_1$ und $\bar{y}_{1,ohne\_Bias}$ dieselbe Varianz.
s_e^2, s_g^2 ermitteln sich allgemein für eine verwendete diskrete Skala von 1..n zu:

$$s_e^{\,2} = \frac{(n-1)^2}{4}$$

und

$$s_g^{\,2} = \frac{(n+1)*(n-1)}{12}$$

Gleichung 97

Für eine diskrete Skala ist aber der Term s_g^2/s_e^2 noch abhängig von der verwendeten Skalenbreite n. Verwenden wir hingegen näherungsweise die Berechnungsvorschriften für kontinuierlich Zufallsvariablen so folgt:

$$s_e^{\,2} = \frac{(n-1)^2}{4}$$

und

$$s_g^{\,2} = \frac{(n-1)^2}{12}$$

Gleichung 98

Nun ist obige Division unabhängig von der Skalenbreite und aus Gleichung 96 folgt:

$$\frac{\bar{y}_{1,ohne\_Bias}}{1+\frac{s_{1,s1}^{\,2}}{\frac{(n-1)^2}{4}}} \le \frac{\frac{n+1}{2}}{1+\frac{1}{3}}$$

Gleichung 99

Setzen wir noch unsere verwendete Skalenbreite n=7 ein, so muss die unverzerrte Lehrbewertung mindestens folgende Ungleichung erfüllen, um einen kriteriumsbezogenen qualitativen Mindestanspruch für sich behaupten zu können.

$$\frac{\bar{y}_{1,ohne\_Bias}}{1+\frac{s_{1,s1}^{\ 2}}{9}} > 3$$

Gleichung 100

Obige Überlegungen der Einbeziehung der Varianz kann natürlich auch bei der Ermittlung der durchschnittlichen Leistungsbewertung über r Veranstaltungen Anwendung finden.
Zuerst berechnen wir die durchschnittliche Leistungsfähigkeit auf herkömmliche Art und Weise.

$$\frac{[1 \quad . \quad . \quad . \quad 1] * Y_{ohne\_Bias}}{\sum_{r} rt} = \frac{\sum_{j=1}^{r} y_{j,ohne\_Bias} * jt}{\sum_{j=1}^{r} jt} = \bar{Y}_{ohne\_Bias}$$

Gleichung 101

$[1 \quad . \quad . \quad . \quad 1]$ steht dabei für den (1 x $\sum_{r} rt$)dimensionalen Zeilenvektor und Y_{ohne_Bias} der nach Gleichung 94 gefundene, verzerrungsfreie ($\sum_{r} rt$ x 1) Spaltenvektor.

Die varianzberücksichtigende durchschnittliche Lehrleistung erhalten wir durch:

$$\frac{\bar{Y}_{ohne\_Bias}}{1+\frac{s_{Y_{ohne\_Bias}}^{\ 2}}{\frac{(n-1)^2}{4}}}$$

Gleichung 102

Hier wird die Varianz natürlich über die Werte des Spaltenvektors Y_{ohne_Bias} berechnet.[96]
Die Berücksichtigung der Varianz über die verschiedenen Veranstaltungen kann damit begründet werden, dass auch in diesem Fall bei Vorliegen zwei unterschiedlicher „wahrer“ Leistungen (7 und 1) die durchschnittliche Leistung aufgrund der vollkommenen Fehlleistung minder angesehen werden muss. Diesen Umstand kann man ermessen, wenn Qualität in 100% Umfang gefordert wird, so z. B. für Produkte, die lebenswichtige Funktionen erfüllen. Es scheint evident, dass im einen Fall ein Komplettversagen (1) und eine erwartungsgemäße Auslösung (7) eines Airbags einen gravierenderen Sicherheitsmangel darstellt als eine mangelhafte (Airbag wurde nur zu 50% gefüllt) Auslösung (4) desselben.
Dieser Wert kann wiederum mit dem in Gleichung 100 gefundenen Wert 3 verglichen werden.
Einen ebenfalls interessanten Ansatz zur Bewertung des Ausbildungsprozesses wird am Institut Austauschbau und Messtechnik an der Technischen Universität verfolgt.[97]
„...Zum einen bedient man sich bei der Bewertung von Studenten durch Lehrbeauftragte des üblichen Systems von Prüfungen; zum anderen wird die Bewertung der einzelnen Lehrveranstaltungen durch die Studierenden mittels "Studentenbefragungsbogen" durchgeführt. Die Auswertung wird in Form einer

[96] Auch hier handelt es sich um den nicht erwartungstreuen MLE-Schätzer für die Varianz.
[97] Vgl. [12].

zugeordneten "Veranstaltungsmatrix" vorgenommen, woraus gegebenenfalls Korrekturmaßnahmen abgeleitet werden.
Dabei haben die Studenten die Möglichkeit, eine Lehrveranstaltung mittels Befragungsbögen zu bewerten, wobei die Fragen zu: Umgang mit Teilnehmern, Stoffauswahl, Vortragsweise, Stoffvermittlung und Prüfungsfragen-Auswahl gestellt werden. Zusätzlich können im Rahmen der Beurteilung verbale Kritiken bzw. Verbesserungsvorschläge gebracht werden. Die Fragen werden durch die Studenten in Anlehnung an Schulnoten bewertet, woraus jeweils Durchschnittsnoten errechnet werden.
Welche Zustände können mit der Matrix erfasst werden? Die Matrix ist eine (4x4)-Matrix, wobei die Spalten die Ergebnisse der Endprüfung ("alle Studenten negativ", "sowohl positive als auch negative Beurteilungen", "alle Studenten positiv" und "keine Studenten zur Prüfung angetreten") und die Zeilen die Summennote der Bewertung einer Lehrveranstaltung ("Note 1 bis 1,2", "Note zwischen 1,2 und 4", "Note schlechter als 4" und "keine Teilnehmer an der Veranstaltung") beinhalten. Mit dieser Matrix sind somit sechzehn Zustände definiert, wobei zwischen Normalzustand, Analysezustand und Eingriffszustand unterschieden werden kann. Entsprechend dieses Zustandes richtet sich die weitere Handhabung der jeweiligen Veranstaltung. Damit werden durch das Qualitätsmanagementsystem Tatsachen offenkundig, wie etwa eine "nicht aktuelle Veranstaltung" oder eine "Veranstaltung mit interessantem Inhalt, aber einer schweren Prüfung".
Beim obligatorischen Review wurde festgestellt, dass sich die Studentenbewertung der Lehrveranstaltungen im Notenbereich 1,14 bis 2,35 bewegt. Dagegen liegt die Beurteilung der Studenten durch die Vortragenden bei Durchschnittsnoten im Bereich 1,00 bis 3,40. ...“
Ganz egal aber auf welche Art und Weise Lehr- oder Prüfungsleistungen gemessen werden und welcher Maßstab zu Anwendung kommt, so vollzieht sich ein jeder Qualitätssicherungsprozess in einem Regelkreissystem. Dies wird schon durch den ganz einfachen Umstand bedingt, dass sich sowohl die Notwendigkeit, korrigierende Maßnahmen zu setzen, als auch die Wirksamkeit der gesetzten Maßnahmen sich an den Prozessdaten orientieren müssen.[98]
Dieses Regelkreissystem soll in Abbildung 49 veranschaulicht werden.

[98] Siehe auch die Bemerkungen unter 2.1 zu Der Prozess des Lernens und Lehrens...

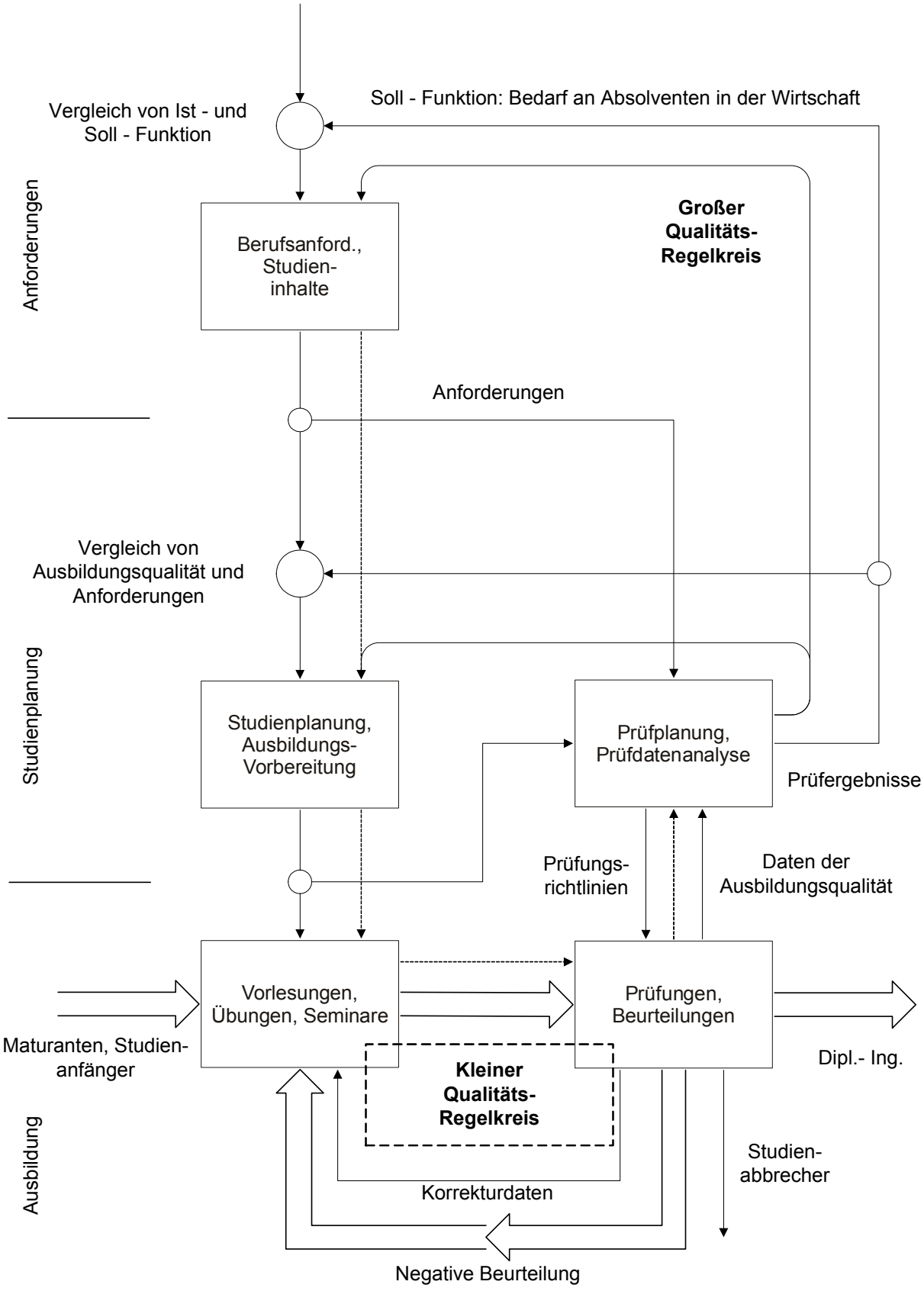


Abbildung 49

Aus dieser Abbildung geht hervor, dass sich der in dieser Arbeit bis jetzt beschriebenen Ansatz zur Qualitätsprüfung auf den „kleinen Qualitätsregelkreises“ bezieht.

Die Verwendung von Prozessdaten zur Qualitätssicherung sprich die ständige Beobachtung der Prozesse und Überprüfung auf Wirksamkeit und hinsichtlich festgesetzter Qualitätsstandards kann als prozessorientierter Ansatz bezeichnet werden.[99]
Bei der Verwendung in einem Qualitätsmanagementsystem betont ein derartiger Ansatz die Bedeutung

- des Verstehens und der Erfüllung von Anforderungen,
- der Notwendigkeit, Prozesse aus der Sicht der Wertschöpfung zu betrachten,
- der Erzielung von Ergebnissen bezüglich Prozessleistung und –wirksamkeit, und
- der ständigen Verbesserung von Prozessen auf der Grundlage objektiver Messungen.

Für die Überprüfung des „großen Qualitätsregelkreises werden die notwendigen Methoden und Ansätze kurz im nächsten Kapitel 5 angerissen.

5. Die Notwendigkeit einer Absolventenstudie zur Überprüfung und Erfahrung des Anforderungsprofils der Akademiker oder die Qualitätskontrolle des großen Qualitätsregelkreises:

Wie schon in einer Sponsionsrede von P. H. Osanna festgestellt wurde, benötigt eine umfassende Qualitätsprüfung von Studium und Lehre erstens eine Orientierung der benötigten Arbeitskräfte bezüglich der Marktsituation und zweitens eine Überprüfung, ob und in welchem Ausmaß die vermittelten Fähigkeiten dem Anforderungsprofil beruflicher Praxis entsprechen.
„...Um funktionierende Qualitätsmanagementsysteme für universitäre Organisations-Einheiten aufzubauen, müssen Rückkoppelungen bzw. Regelkreise vorgesehen werden: einerseits natürlich KLEINE Regelkreise, welche in Form von Prüfungen, Kolloquien oder Seminaren innerhalb der universitären Ausbildung wirken, andererseits GROSSE, welche weit aus dem universitären Bereich hinaus ins wirtschaftliche Geschehen reichen und sich an zu formulierenden Qualitäts-Definitionen orientieren.
In diesem Sinn ist grundsätzlich davon auszugehen, dass insbesondere Universitäten der technischen Wissenschaften nicht ohne Außenwelt leben können, sie sich vielmehr laufend auch den Bedürfnissen der Praxis und vorgegebenen Veränderungen anpassen müssen.
Ich möchte Sie aber auch ausdrücklich darum ersuchen, mit uns, Ihren Lehrern, und mit Ihrer Universität weiter in engem Kontakt zu bleiben. Einerseits sollte dies im Sinne einer "TECHNISCHEN BETREUUNG und WARTUNG" sein, um unsere weitere Unterstützung und entsprechende Weiterbildungsmöglichkeiten in Anspruch zu nehmen. Andererseits können Sie durch entsprechende Rückmeldungen - egal ob positiv oder negativ - mit zur Gestaltung entsprechender REGELKREISE beitragen, ganz im Sinne einer STEIGERUNG DER QUALITÄT DER UNIVERSITÄREN AUSBILDUNG.“[100]

[99] Vgl. [13]
[100] Vgl. [14] und Bemerkungen unter Kapitel 3.4,...Dabei werden 2 Aspekte untersucht...

Dabei ist es notwendig die an die Absolventen gestellten Anforderungen mit den selbst eingeschätzten, direkt nach dem Studienabschluss vorhandenen Fähigkeiten / Kompetenzen gegenüberzustellen.[101]
Diese müssen dann auf signifikante Unterschiede geprüft werden.
Erhoben werden können z.B. folgende Merkmale:

- Lernfähigkeit,
- Fachkenntnisse,
- Kritisches Denken,
- Kreativität,
- Fremdsprachen,
- Teamarbeit,
- Breites Allgemeinwissen,
- Zeiteinteilung,
- Fachübergreifendes Denken,
- Ausdrucksfähigkeit,
- Planen, Koordinieren und Organisieren,
- Anpassungsfähigkeit,
- Wirtschaftliches Denken,
- Verhandlungsgeschick,

Da nicht die bloße Unterschiedlichkeit der abhängigen Stichproben zwischen der Anforderung und der Kompetenz über alle Variablen mit einem multivariaten Ansatz untersucht werden soll, sondern die Analyse dezidiert Auskunft über Unterschiedlichkeiten nicht im summarischen sondern speziell für jede Variable gesondert aufzeigen soll, wählen wir als Methode den „Vergleich von zwei abhängigen Stichproben hinsichtlich ihrer zentralen Tendenz“ oder den Wilcoxon-Test für ordinale Skalenqualität.
Wir wollen davon ausgehen, dass die untersuchte Stichprobe der Absolventen, die in die Berechnung eingehen, größer als 25 ist. Die verwendete Bewertungsskala soll wieder die Endpunkte 1 für völlig unzutreffend bzw. schlechteste Bewertung und 7 für zutreffend bzw. beste Beurteilung besitzen.
Zur Erläuterung mögen sich beispielhaft für die Variable Fremdsprachen folgende selbst eingeschätzte Anforderungen und Kompetenzen über 30 hypothetische Absolventen ergeben haben.

[101] Vgl. [15]

Absolvent:	Fr-Spr.-An.	Fr.-Spr.-Ko	Differenz	Rang	Vorzeichen:	Ränge von positiven Differenzen:
1	4	5	-1	5,5		
2	3	3	0	-		
3	5	3	2	15	+	15
4	7	4	3	22	+	22
5	2	7	-5	26		
6	3	4	-1	5,5		
7	6	6	0	-		
8	6	4	2	15	+	15
9	3	4	-1	5,5		
10	3	5	-2	15		
11	4	5	-1	5,5		
12	5	7	-2	15		
13	7	7	0	-		
14	2	3	-1	5,5		
15	2	4	-2	15		
16	3	4	-1	5,5		
17	5	4	1	5,5	+	5,5
18	5	4	1	5,5	+	5,5
19	5	6	-1	5,5		
20	3	6	-3	22		
21	6	3	3	22	+	22
22	7	5	2	15	+	15
23	7	5	2	15	+	15
24	2	2	0	-		
25	1	5	-4	25		
26	1	4	-3	22		
27	1	4	-3	22		
28	5	3	2	15	+	15
29	4	6	-2	15		
30	4	3	1	5,5	+	5,5
						135,5

Abbildung 50

Die erste Spalte enthält die fiktiven Werte für die beurteilten, fremdsprachlichen Anforderungen in der beruflichen Praxis. Die zweite Spalte enthält die entsprechenden Einschätzungen für die Kompetenz.
Spalte 3 enthält die Differenz bezüglich Anforderung und Kompetenzeinschätzung. Werte die eine 0 in Differenzspalte aufweisen, werden aus der Berechnung ausgeschlossen. Es existieren 4 Übereinstimmungen zwischen Anforderungen und Kompetenz, sodass in die Berechnung n=26 Absolventen (Wert in der dritten Spalte ≠ 0) einfließen.[102]
Wir erhalten in unserem Beispiel $n_{pos.}$ = 10 positive und $n_{neg.}$ = 16 negative Differenzen. In der letzten Spalte werden jene verbundenen Ränge mit der kleineren Anzahl (positive Differenzen, da $n_{pos.} < n_{neg.}$) zum T-Wert zusammengezählt.
Für die Bestimmung der verbundenen Ränge werden die Differenzen nach ihrem Betrag in eine Rangreihenfolge gebracht.[103]

[102] n ist immer noch größer als 25. Siehe Bemerkungen weiter oben.
[103] Siehe Abbildung 51.

Absolvent:	Differenz	Rang		
1	-1	5,5	1	
6	-1	5,5	2	
9	-1	5,5	3	
11	-1	5,5	4	
14	-1	5,5	5	
16	-1	5,5	6	
17	1	5,5	7	
18	1	5,5	8	
19	-1	5,5	9	
30	1	5,5	10	5,5
3	2	15	11	
8	2	15	12	
10	-2	15	13	
12	-2	15	14	
15	-2	15	15	
22	2	15	16	
23	2	15	17	
28	2	15	18	
29	-2	15	19	15
4	3	22	20	
20	-3	22	21	
21	3	22	22	
26	-3	22	23	
27	-3	22	24	22
25	-4	25	25	25
5	-5	26	26	26
2	0	-		
7	0	-		
13	0	-		
24	0	-		

Abbildung 51

Wir erhalten den verbundenen Rang für z.B. die betragsmäßige Differenz von 2, indem wir die Ränge von 11 bis 19 summieren und durch die entsprechende Anzahl, der Ränge dividieren. Also:

$$15 = \frac{11 + ... + 19}{9}$$

Gleichung 103

Wir erhalten die Prüfgröße nach:

$$z = \frac{T - \mu_T}{\sigma_T}$$

Gleichung 104

Wobei der T-Wert aus der Abbildung 50 letzte Spalte und gelb hinterlegt zu entnehmen ist.
μ_T erhalten wir durch:

$$\mu_T = \frac{n*(n+1)}{4} = \frac{26*27}{4} = 175{,}5$$

Gleichung 105

σ_T erhalten wir durch folgende Berechnungsvorschrift:

$$\sigma_T = \sqrt{\frac{n*(n+1)*(2*n+1) - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{2}}{24}}$$

Gleichung 106

Wobei k = der Anzahl der Rangbindungen und t_i = der Länge der Rangbindung i entspricht.
Für unser Beispiel errechnet sich σ_T zu:

$$\sigma_T = \sqrt{\frac{26*27*(2*26+1) - \frac{1}{2}*\left[\left(10^3-10\right)+\left(9^3-9\right)+\left(5^3-5\right)\right]}{24}} = 38{,}89$$

Gleichung 107

Wir haben in unserem Beispiel 3 Rangbindungen nämlich bei den absoluten Differenzen 1, 2 und 3 mit den jeweiligen Längen 10, 9 und 5.
Wir erhalten also schließlich einen z-Wert von:

$$z = \frac{135{,}5 - 175{,}5}{38{,}89} = -1{,}03$$

Dieser wird mit dem kritischen Wert bei einem bestimmten α-Niveau einer Standardnormalverteilung verglichen (z.B. α = 25% und zweiseitigem Test $z_{crit.}$ = ±1,15).[104] Obiges z liegt innerhalb $z_{crit.}$ also innerhalb der Grenzen –1,15< z ≤1,15. Es besteht also in unserem fiktiven Beispiel kein signifikanter Unterschied zwischen der beruflichen, fremdsprachlichen Anforderung und der selbst eingeschätzten Kompetenz über die Stichprobe.
Obige Anforderungsprofile können durch Befragungen der Arbeitgeber verifiziert und validiert werden und zudem den Vorteil bieten Studienordnungs- und Ausbildungsprogramme den wirtschaftlichen Anforderungen und Gegebenheiten im Sinne einer Rückkopplung des „großen Qualitätsregelkreises aktuell zu halten.

6. Schlusswort:

Obige Arbeit hat sich zum Schwerpunkt gesetzt, Transparenz hinsichtlich verwendeter, statistischer Methoden und der zu Grunde liegenden Datenmanipulationen zu schaffen. Mag der Ansatz auch noch so hehr sein, Qualität mittels Indikatoren festhalten und messen zu wollen und damit ein wichtiger und richtiger Schritt hinsichtlich einer Objektivierung der Qualitätsmessung unternommen wird, so muss doch der offensichtlichen Schwäche, verschiedene Berechnungsverfahren und statistische Modelle verwenden zu können, zumindest durch eine vernünftige und begründete Auswahl des betreffenden statistischen Instruments, durch die Festlegung der Berechnungsvorschriften und eine genaue Beschreibung und Argumentation der Datenmanipulation begegnet werden. So bleibt gewährleistet, dass „falsche“ Interpretationen der Ergebnisse zumindest nicht zu Lasten der Unkenntnis der Nachvollziehbarkeit und des Zustandekommens der Ergebnisse resultieren.
Alle Berechnungen können mit Hilfe der Statistikfunktionen von Microsoft Excel mit Leichtigkeit und angenehm vollzogen werden (einzige Ausnahme zur Eigenwertberechnung der entsprechenden Matrizen wurden Mathematikprogramme

[104] Aufgrund der Wunschhypothese (H0) wird wiederum ein vergrößertes α-Niveau von 25% empfohlen, um den β-Fehler indirekt klein zu halten.

benützt- Mathematica / Maple). Dies bietet außerdem den Vorteil, dass der Anwender nicht Besitzer von einschlägigen Statistikprogrammen sein muss, deren Berechnungsprozeduren vielen Anwendern sowieso nicht oder in unzureichendem Maße bekannt sind.
Schließen möchte ich mit einem Zitat von Heiner Rindermann über die Notwendigkeit von studentischer Lehrevaluationen zur Qualitätssicherung von Studium und Lehre: „Studentische Lehrevaluation ähnelt dem Anfertigen eines Porträts. Ein Porträt stellt nie eine exakte Abbildung der äußeren Erscheinung einer Person dar, jedoch ist es genauer und wirklichkeitsgetreuer als jede Vorstellung, die wir uns durch bloßes Hörensagen vom Äußeren einer Person machen können.“[105]
Ich möchte hier einen Schritt weiter gehen: „Für eine „wirklichkeitsgetreue“ Interpretation einer Abbildung ist es nicht nur eminent wichtig zu wissen, was denn „ungefähr“ das Bild vermitteln soll, sondern ebensolche Bedeutung muss auch den eingesetzten Mitteln zur Erzeugung desselben zugebilligt werden. Eine Verwendung von Mitteln unzureichender Dimensionalität und „Tiefe“ zur Darstellung kann zu interpretatorischen Fehlschlüssen ungeübter Interpreten in dem Sinne führen, dass die Betrachtung eines 2-dimensionalen Bildes eines Quadrates auch als solches interpretiert wird, die Wahrhaftigkeit des wirklichen Gegenstandes, nämlich der des Würfels aber verborgen bleibt. Eine Abstraktion und damit einhergehend eine Vereinfachung statistischer Modelle auf eine geringere Dimensionalität ist nur bei entsprechendem Vorwissen über den abzubildenden Gegenstand legitim.“

[105] Vgl. [9].

Literaturliste:

[1] P.H. Osanna und D. Prostrednik: „Quality at the university“. (1)
[2] Herbert Altrichter und Michael Schratz: Qualität von Universitäten / Peter Neudorfer: „Arbeitsberichte der Institutsvorstände und die Performance Indicator-Diskussion“. (3)
[3] Qualität von Studium und Lehre/Dokumente zur Hochschulreform 91/1994. (4)
[4] Bildungsreform Band 4 / Indikatorisierung der Empfehlungen des Forum Bildung von Dr.Isabell van Ackeren und Dr. Gertrud Hovestadt, Dezember 2003. (5, 6, 68)
[5] PISA 2000: Die Studie im Fokus der Statistik. (68)
[6] Multivariate Analysemethoden: Backhaus, Erichson, Plinke, Weiber. (7, 46)
[7] Testaufbau und Testanalyse: Gustav A. Lienert. (8, 9, 10, 12, 15, 19, 20, 25), 4. Auflage 1989
[8] Statistik für Sozialwissenschaftler: Jürgen Bortz. (14, 40, 53, 55, 56, 59, 61, 63, 64, 75, 77, 80, 81), 5. vollständig überarbeitete und aktualisierte Auflage 1999.
[9] Lehrevaluation: Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts, Heiner Rindermann. (21, 26, 31, 39,41, 58, 105)
[10] Welche Faktoren beeinflussen die Evaluation von Lehrkräften? Von Bettina Greimel und Alois Geyer. (21)
[11] Zwischenbericht: Lehrevaluation. Konstruktion eines Fragebogens für Vorlesungen und Seminare von Yvonne Grabbe an der Universität Münster. (21)
[12] QM nach EN/ISO 9000 im universitären Bereich- Erfahrungsbericht bei der konkreten Umsetzung seit 1992 von P. Herbert OSANNA und Thomas MADER- TU-Wien. (2, 97)
[13] Prozessmodell 2000 von P. Herbert OSANNA. (99)
[14] Papier zur Sponsionsrede 1995 von P. Herbert OSANNA. (100)
[15] Dresdner Absolventenstudie 2003 Informatik, wissenschaftliche Leitung Karl Lenz, Verfasser: Andrea Puschmann, Jacqueline Popp, Rene Krempkow. (101)