FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

FACHBEREICH WIRTSCHAFTS-
WISSENSCHAFTEN

Lehrstuhl für Statistik und Ökonometrie

Diskussionspapier
90 / 2013

Lehrverbesserung durch Online-Tests – Effekte der Eigenarbeit von Studierenden

Thomas Pleier
Benedikt Mangold

Lange Gasse 20 · D-90403 Nürnberg

Lehrverbesserung durch Online-Tests – Effekte der Eigenarbeit von Studierenden

Thomas Pleier* Benedikt Mangold†

FAU Erlangen-Nürnberg
Lange Gasse 20
D-90403 Nürnberg

20. November 2013

Abstract

Ein Anreizsystem zur kontinuierlichen Mitarbeit wird als Mittel zur Verbesserung der universitären Lehre vorgestellt. Da Missbrauchsmöglichkeiten bestehen, kommen bei der qualitativen Erfolgsanalyse Zweifel auf, ob die Anreize an der richtigen Stelle greifen. Ziel dieser Arbeit ist es, die vorliegenden Zweifel zu zerstreuen. Dabei steht die Quantifizierung des Effekts der Mitarbeit auf den Lernerfolg im Mittelpunkt. In der konfirmatorischen Analyse des aufgestellten Modells wird der positive Effekt der Lehrverbesserung sichtbar. Das beschriebene Anreizsystem wird, trotz Anfälligkeit, zur Fortsetzung empfohlen – allerdings wird verstärkte Kontrolle nahegelegt.

Schlagwörter: Mitarbeit, Studium, Lernerfolg, Lehrverbesserung, LISREL.

*Thomas.Pleier@fau.de
†Benedikt.Mangold@fau.de

1 Einleitung

Kontinuerliche Mitarbeit der Studierenden während des Semesters ist gerade in einem Fach wie Statistik von zentraler Bedeutung, in dem Lerninhalte sukzessive entwickelt werden und deshalb zwingend aufeinander aufbauen. Wer zu Beginn z.B. die Konzepte der Wahrscheinlichkeit und der Streuung nicht verinnerlicht hat, wird während des gesamten Semesters den restlichen Lernstoff kaum verstehen können. Dies kann dazu führen, dass gerade die Vorlesung nur noch sporadisch besucht wird und sich dort auch kein positives Lernerlebnis einstellen kann. Die Übungen haben nicht die Zielsetzung in der Vorlesung versäumte bzw. nicht verstandenene Lerninhalte in Kurzform zu rekapitulieren. Sie dienen gezielt der Vorbereitung auf die Klausur durch Behandlung spezifischer zu lösender Aufgaben. Letztlich basiert der Erfolg der Übungen auch auf der kontinuierlichen Mitarbeit der Studierenden in allen Lehrveranstaltungen.

Kontinuerliche Mitarbeit steht einer studentischen Lernstrategie entgegen, die den Fokus auf das Lernen kurz vor der Klausur legt. Ohne die Inhalte im laufenden Semester aktiv verfolgt zu haben, ist aber eine Einschätzung des Lernaufwandes kaum möglich, der für das erfolgreiche Bestehen der Klausur nötig ist. Dies führte dazu, dass sich viele Studierende am Tag der Klausur dazu entschieden, unabgemeldet fernzubleiben, was aufgrund der geltenden Prüfungsordnung für Studierende im Erstversuch ohne Konsequenzen möglich war. Dies verursacht einerseits ein logistisches Problem (Räume, Aufsichten, Kopieraufwand etc.), wirkt aber anderseits auch studienzeitverlängernd.

Sogenannte Midterm-Prüfungen, wie an angelsächsischen Universitäten üblich, könnten eine Lösung sein, um die kontinuierliche Mitarbeit während des Semesters sicherzustellen. Sie sind aber im Massenbetrieb einer deutschen Universität mit weit mehr als 1.000 Prüflingen in einer Veranstaltung nicht durchführbar. Eine Alternative zu Midterm-Prüfungen könnte die Entwicklung eines Anreizsystems sein, das aus mehreren automatisch ausgewerteten Tests besteht, deren erfolgreiche Bearbeitung zu Bonuspunkten führt, die zu den regulär in der Klausur erreichten Punkten addiert werden. Voraussetzung für die Durchführbarkeit solcher Tests ist, dass sie automatisch korrigiert und ausgewertet werden können. Dazu bietet sich die den Studierenden ohnehin vertraute Lernplattform StudOn an, über die per elektronischem Zugang den Studierenden Tests zugewiesen und anschließend ausgewertet werden können. Dieses Anreizsystem ist im Wintersemester 2012/13 (WS12/13) für die Veranstaltung der Statistik erstmals etabliert worden. Mittlerweile liegen Aus-

wertungen vor, mit deren Hilfe untersucht werden soll, ob das Instrument erfolgreich im Sinne der Zielerreichung einer kontinuierlichen Mitarbeit und einer geringeren Anzahl an Studierenden, die trotz Anmeldung nicht zur Klausur erschienen sind, eingesetzt werden konnte.

Nach Abschluss des Semester stellte sich die Frage nach dem Erfolg des Projekts und ob es wiederholt werden solle. Die inhaltliche Relevanz und lehrunterstützende Eigenschaft wurde nicht angezweifelt. In einer qualitativen Auseinandersetzung dominierten jeweils ein positives und ein negatives Argument:

+ Das Projekt genoss ein hohes Ansehen unter den Teilnehmern.

– Es ist fragwürdig, inwiefern die im Vorfeld formulierten Ziele erreicht werden konnten.

In dieser Gegenüberstellung nehmen die Zweifel an der Zielerreichung eine gewichtige Position ein. Um die Fortführung des Projekts zu rechtfertigen, wird eine quantitative Analyse der Zielerreichung durchgeführt.

Der Aufbau des Artikels ist wie folgt: In Abschnitt 2 werden die Modalitäten der Teilnahme an den Online-Aufgabenserien mit den zu erreichenden Punkten beschrieben, mögliche Manipulationen des Ergebnisses eines Durchlaufs sowie geeignete Vermeidungsstrategien vorgestellt. In Abschnitt 3 werden die durch Aufgabenserie und Klausur gesammelten Daten quantitativ ausgewertet und die Wirkungen und Zusammenhänge in einer LISREL-Analyse überprüft. Eine abschließende Zusammenfassung und Bewertung der Ergebnisse finden sich im letzten Abschnitt (4).

2 Rahmenbedingungen

Da es sich um die erstmalige Durchführung einer Online-Aufgabenserie handelte, unterschieden sich die Rahmenbedingungen bei den Serien etwas – mit der Zeit wurden Schwachstellen klar, die erst bei späteren Aufgabenserien behoben werden konnten. Gemeinsam war allen Online-Aufgabenserien, dass stets der in der Vorlesung vermittelte Stoff bis ca. eine Woche vor Serienbeginn abgeprüft wurde. Jede Serie bestand aus folgenden Aufgabentypen:

1. Single Choice-Fragen,
2. Multiple Choice-Fragen,

3. Lückentext-Fragen,
4. Freitext-Fragen (für Abfragen von R[1] -Code),
5. Zuordnungsfragen,
6. Image-Fragen (der Bearbeitende musste einen korrekten Bereich auf einem Bild markieren).

Es wurde eine Teilnehmerzahl von etwa 1.500 Studierenden erwartet – einen gleichzeitigen Zugriff hätte die Lernplattform nicht bewältigen können. Daher wurde eine Obergrenze von 100 gleichzeitigen Bearbeitungen der Online-Aufgabenserien festgelegt. Durch die Festlegung eines Start- und Endzeitpunkts auf jeweils 03:00 Uhr morgens wurde diese Obergrenze nur zu Mittagszeiten innerhalb des Bearbeitungszeitraums erreicht.

2.1 Möglichkeiten der Manipulation

Es konnte nicht ausgeschlossen werden, dass einzelne Studierende versuchen würden das Ergebnis der Online-Aufgabenserie zu ihren Gunsten zu manipulieren. In den folgenden Unterpunkten zeigen wir denkbare Fälle auf und beschreiben wie versucht wurde, eine eventuelle Manipulation zu erschweren.

2.1.1 Gruppenarbeit

Zur generellen Vermeidung von Zusammenarbeit wäre es nötig gewesen, alle Teilnehmer die Aufgabenserien unter Aufsicht, z.B. in einem CIP-Pool, bearbeiten zu lassen. Dies war aufgrund der hohen erwarteten Teilnehmerzahl nicht möglich. Während der Bearbeitungszeit konnte auf dem Campus beobachtet werden, dass sich Studierende in Gruppen zusammenfanden um gemeinsam die Aufgabenserien zu bearbeiten. Sofern alle Teilnehmer an derartigen Gruppen bei der Bearbeitung aktiv beteiligt waren, kann dies als eine Art Lerngruppe verstanden werden. Der Effekt der Gruppenbildung war zwar nicht im Sinne der Gestalter der Online-Aufgabenserie – sofern es sich um aktive Teilname eines Studierenden an einer solchen Gruppe handelt erfüllen auch diese Lerngruppen das Ziel der Motivation zur Mitarbeit.

Um die passive Teilnahme an solchen Lerngruppen durch einfaches Kopieren der Ergebnisse anderer Studierender zu erschweren wurden die Aufgaben

[1] www.r-project.de

bei Start eines Serien-Durchlaufs zufällig aus vorher definierten Fragenpools gezogen. Somit wurde eine geringe Wahrscheinlichkeit für zwei identische Serien gewährleistet. Des Weiteren wurden die Antwortmöglichkeiten bei den Fragentypen 1., 2., 3. und 5. bei jedem Teilnehmer zufällig neu angeordnet.

2.1.2 Delegierung

Studierenden ist es untersagt mittels Weitergabe ihrer Login-Daten dritten Personen Zugang zur universitätsinternen Lernplattform zu verschaffen. Dennoch war faktisch nicht auszuschließen, dass sich Teilnehmer externe Hilfe beschafften. Die kontrollierte Teilnahme unter Aufsicht war nicht möglich, ist aber die einzige Möglichkeit, unerwünschte Hilfestellung zu vermeiden. Um zumindest das Vertrauen in die Leistung eines Dritten zu erschweren, war kein Feedback über die erreichte Punktzahl durch das System vorgesehen.

2.1.3 Nicht ernst gemeinte Teilnahmen

Gerade in der ersten Aufgabenserie hatten viele Teilnehmer eine Bearbeitungszeit von unter einer Minute mit niedriger erreichter Punktzahl – dies erweckt den Anschein von nicht ernst gemeinten Versuchen. Diese Versuche könnten durchgeführt worden sein, um Informationen (Absatz 2.1.4) über die in der Aufgabenserie behandelten Inhalte und gestellten Fragen zu sammeln. Um dies zu erschweren wurden den Fragenpools oftmals ähnliche Fragen mit leichten Veränderungen zugeteilt und die verfügbare Bearbeitungszeit knapp gehalten. Zudem erfolgte die Anordnung der Antwortmöglichkeiten zufällig.

2.1.4 Verbreitung der Ergebnisse

Da jeder Studierende die Aufgabenserie auf seinem persönlichen Computer durchführen konnte, überrascht es nicht, dass Screenshots von bereits bearbeiteten Fragen im Internet hochgeladen wurden. Die in 2.1.3 getroffenen Maßnahmen wirken auch diesem Effekt entgegen.

2.2 Rahmenbedingungen der Online-Aufgabenserien

2.2.1 Online-Aufgabenserie 1 (SOTI)

Für die erste Aufgabenserie zugelassen waren alle Studierenden, die Mitglied einer der Vorlesung Statistik zugehörigen Lernplattform-Gruppe waren. Die

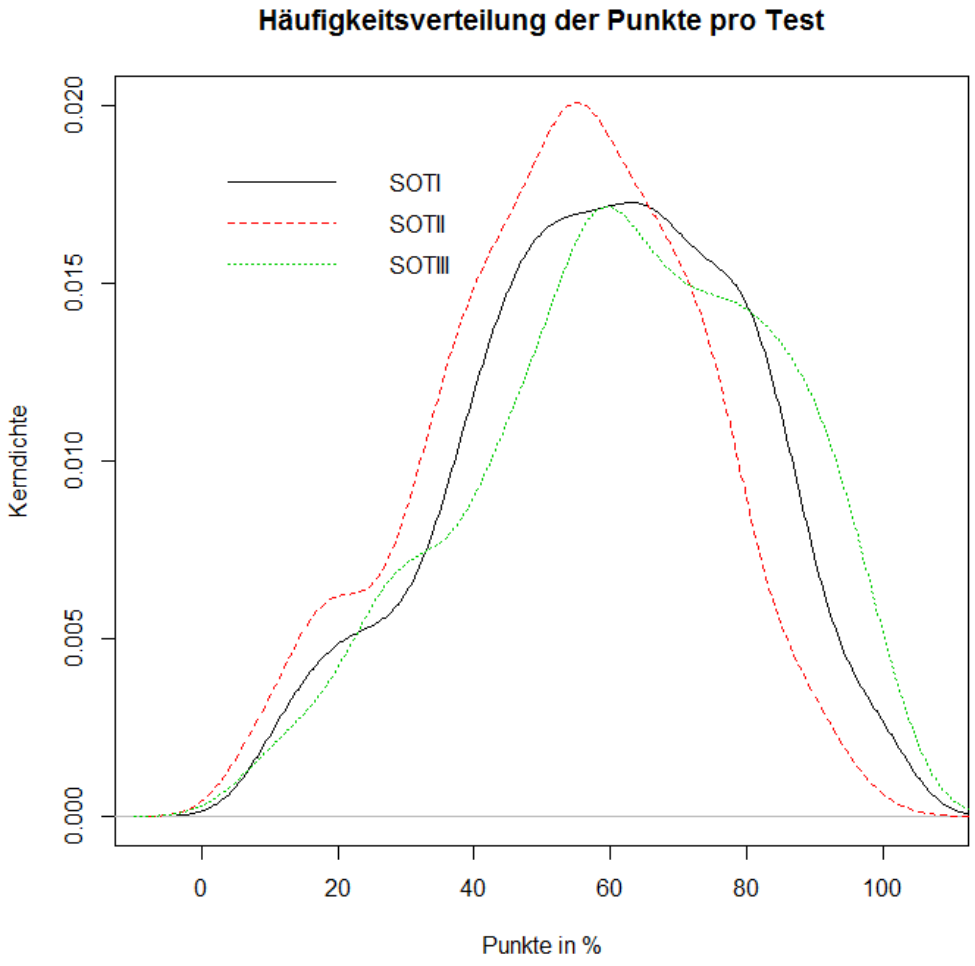

Abbildung 1: Kerndichteplot der erreichten Punkte je Online-Aufgabenserie.

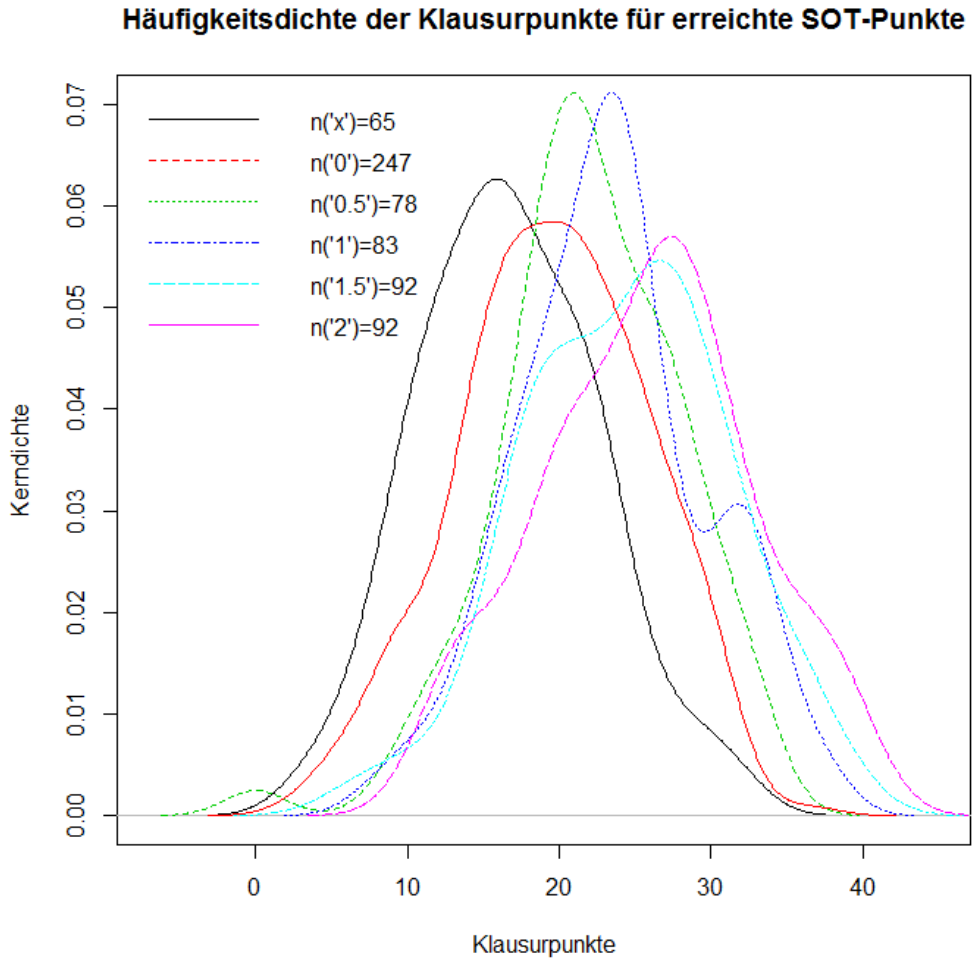

Abbildung 2: Kerndichteplot der erreichten Klausurpunkte (exklusive SOT-Zusatzpunkte) aufgeschlüsselt in die zuvor erreichten SOT-Punkte. Dabei bezeichnet $n(\text{'}\cdot\text{'})$ die Anzahl der Studierenden, die durch die Online-Aufgabenserie '$\cdot$' Punkte bereits im Vorfeld erworben hatten – 'x' bedeutet dabei „Keine Teilname“.

Aufnahme in diese Gruppe erfolgte durch ein in der Vorlesung bekannt gegebenes Passwort, somit war eine Mitgliedschaft jedem Studierenden, der Kenntnis dieses Passworts hatte, möglich. Mitglieder waren nicht nur Studierende, die im Wintersemester 2012/2013 die Veranstaltung „Statistik“ besuchten bzw. planten, die Klausur zu schreiben, sondern auch alle Studierenden aus vorhergehenden Semestern, die sich noch nicht aus der Gruppe abgemeldet hatten. Das in dieser Aufgabenserie abgefragte Themengebiet umfasste sämtliche Themen, die bis eine Woche vor Serienbeginn in der Vorlesung besprochen waren. Dabei ging es primär um Definitionen von Grundbegriffen und Verständnis-Fragen zu diesen.

2.2.2 Online-Aufgabenserie 2 (SOTII)

Zugang zu dieser Aufgabenserie hatten nunmehr Studierende aus der im Absatz 2.2.1 erwähnten Statistik-Gruppe, die auch über ein Buchungssystem zur Abschlussklausur angemeldet waren. Der Grund für diese Einschränkung war, Manipulation in Form explorativer Teilnahme zu verringern (Abschnitt 2.1.3). Zusätzlich zu dem bis dato in der Vorlesung behandelten Stoff wurden in dieser Aufgabenserie Fragen gestellt, die von den Studierenden Transferleistung erforderten - diese bestanden aus der direkten Anwendung erlernter Modelle unter neuen Rahmenbedingungen.

2.2.3 Online-Aufgabenserie 3 (SOTIII)

Die dritte Aufgabenserie fand außerhalb der Vorlesungszeit statt und hatte identische Rahmenbedingungen wie Aufgabenserie 2 (Absatz 2.2.2). Der abgefragte Stoff umfasste den kompletten Inhalt der Vorlesung. Einige Fragen ähnelten späteren Klausuraufgaben.

2.3 Genereller Testaufbau

Für jede Online-Aufgabenserie betrug die Bearbeitungszeit 20 Minuten, wobei die Anzahl der zu bearbeitenden Fragen variierte. Eine Übersicht der in den Aufgabenserien SOTI - SOTIII erreichten Punkte findet sich in Abbildung 1. Die Aufgaben der drei Serien wurden aus zwischen 10 und 15 Fragenpools, bestehend aus 1-11 Fragen, generiert. Die Teilnahme an den Aufgabenserien variierte zwischen 1.123 Teilnahmen von 2.709 potentiellen

Teilnehmern in SOTI, 920 von 1.096 beim SOTII und 742 von 1.096 beim SOTIII.

Pro Aufgabenserie konnten zwischen 0 und 100 Punkten erreicht werden. Lediglich Studierende, die über alle drei Online-Aufgabenserien hinweg mindestens 150 Punkte erreicht haben, wurden bei der Zusatzpunktevergabe berücksichtigt. Die Punktzahl dieser Studierenden wurde der Größe nach aufsteigend sortiert und in vier Untergruppen zusammengefasst. Die untersten 25% erhielten 0,5 Zusatzpunkte für die Klausur, die nächsthöheren 25% 1 Zusatzpunkt, die darauffolgenden 25% 1,5 Punkte und schlussendlich die besten 25% die vollen 2 Zusatzpunkte.

2.4 Klausur

Den Studierenden war es gestattet, neben einem nicht-programmierbaren Taschenrechner auch eine selbst erstellte Formelsammlung mit einem Umfang von 2 DIN-A4 Blättern zur Bearbeitung der Klausur zu verwenden. In 120 Minuten Bearbeitungszeit waren zum Bestehen 16 von 40 Punkten zu erreichen, wobei die per Online-Aufgabenserie erreichten Bonuspunkte den Studierenden zum Zeitpunkt der Klausur nicht bekannt waren. Aufgeteilt war die Klausur in vier Teilaufgaben, in denen jeweils 10 Punkte zu erreichen waren.

Im Vorfeld wurde die Klausur von den Mitarbeitern des Lehrstuhls als gleichwertig zu bisherigen Klausuren eingestuft – dies spiegelte sich auch im üblichen Punkteschnitt wider.

2.5 Zwischenfazit

Die in Abschnitt 1 vermuteten positiven und negativen Auswirkungen der Online-Aufgabenserien auf den Lernerfolg können mit Hilfe der Tabellen 1 bzw. der Abbildung 2 deskriptiv unterlegt werden:

+ Nach Abbildung 2 sieht es aus, als hätten die Online-Aufgabenserien den gewünschten Effekt: Studierende, welche sich bereits im Vorfeld durch Teilnahme an den Online-Aufgabenserien Zusatzpunkte sichern konnten, erreichten offensichtlich eine höhere Punktzahl in der Klausur (exklusive Zusatzpunkte). Diese Beobachtung wird auch durch die stark besetzte Hauptdiagonale in Tabelle 1 unterstützt.

		Erreichte Notenstufe				
		1	2	3	4	5
SOT-Punkte	2	0,239	0,380	0,261	0,065	0,054
	1,5	0,174	0,359	0,239	0,163	0,065
	1	0,167	0,155	0,417	0,179	0,083
	0,5	0,038	0,256	0,359	0,244	0,103
	0	0,004	0,148	0,320	0,260	0,268
	x	0,000	0,060	0,209	0,299	0,433
	↑	0,482	0,528	0,310	0,171	0,090

Tabelle 1: Relative Häufigkeiten der erreichten Notenstufe, bedingt auf die zuvor erhaltenen SOT-Punkte. Die Zeile ↑ beinhaltet den Anteil an Studierenden innerhalb der jeweiligen Notenstufe, die sich wegen der Zusatzpunkte um eine drittel Notenstufe verbessert hatten. **x** bezeichnet dabei die Gruppe der Studierenden, welche nicht am Test teilgenommen hatten. Nähere Erläuterungen zum hellgrau unterlegten Feld in Tabelle 2.

	Erworbene SOT-Punkte					
	2	1,5	1	0,5	0	**x**
Notenstufe 4 durch SOT	5	2	2	1	0	0

Tabelle 2: Nähere Erläuterung der hellgrauen Zelle in Tabelle 1: 17,1% entsprechen 24 Studierenden, die sich in die/innerhalb der Notenstufe 4 verbessert hatten. 10 davon hatten nur durch die Zusatzpunkte die Klausur bestanden – 'x' bedeutet dabei „Keine Teilname an den Online-Aufgabenserien“.

- Fragwürdig bleibt die Besetzung der dunkelgrau unterlegten Felder in Tabelle 2: Immerhin 32 Studierende erreichten in der Klausur lediglich die Notenstufen 4 bzw. 5, obwohl sie im Vorfeld überdurchschnittlich gut bei der Bearbeitung der Online-Aufgabenserien abgeschnitten hatten, siehe Abschnitt 2.3.
  Nicht zu vernachlässigen ist auch eine weitere Aufschlüsselung durch Tabelle 2: Speziell diejenigen Studierenden, welche nur anhand der erworbenen Zusatzpunkte die Klausur überhaupt mit der Note 4 bestanden haben, erreichten überdurchschnittlich viele Punkte bei den Online-Aufgabenserien (sie gehörten zu den 50% besten Bearbeitungen unter denjenigen, die mindestens 50% der Gesamtpunktzahl erreichten). Dies lässt auf die in Abschnitt 2.1 beschriebenen Manipulationsmöglichkeiten schließen.

Um den Einfluss der ersten beiden Effekte gegeneinander abschätzen zu können, sollen diese im folgenden Abschnitt 3 mit Hilfe eines LISREL-Modells gegenübergestellt und ausgewertet werden.

3 Quantitative Analysen

Prinzipiell ist ein gemeinsames Auseinandersetzen mit den Lerninhalten in Form aktiver Gruppenarbeit begrüßenswert. Erhärtete sich jedoch der Verdacht, dass eine große Menge an Teilnehmern bei der Beantwortung ihres Fragebogens nicht beteiligt gewesen war, gälte das erste Ziel als verfehlt:

Das *Hauptziel* der Online-Aufgabenserien war die Qualität der Mitarbeit der Studierenden zu verbessern.

Studierende im Erstversuch können ohne Konsequenzen unentschuldigt der Klausur fernbleiben. Die durch dieses Fernbleiben bedingte mangelhafte Raumauslastung und Überschussdrucke von Angaben verursachen erhebliche unnötige Kosten. Außerdem ist es im Interesse der Studierenden, Prüfungen nicht zu schieben oder zu wiederholen, da sich dies meist studienzeitverlängernd auswirkt.

Das *zweite Ziel* der Online-Aufgabenserien war deshalb das Senken des Anteils an Studierenden, die von der Klausur fernbleiben.

Inwiefern die Zielerreichung gescheitert/gelungen ist, wird im Folgenden untersucht. Der Fokus liegt dabei auf dem Ziel der Verbesserung der Mitarbeit.

3.1 Fernbleiben von der Prüfung

Es wird zunächst über drei Semester hinweg der Anteil an den für die Klausur angemeldeten Studierenden verglichen, die ohne Entschuldigung zur Klausur nicht erschienen sind. Dabei werden nur Studierende im Erstversuch betrachtet, da sich diese bewusst anmelden müssen. Studierende im 2. oder 3. Versuch werden automatisch zum nächstmöglichen Termin angemeldet und können sich nur unter Angabe von zwingenden Gründen von der Klausur abmelden. Von diesen angemeldeten Studierenden wird der Anteil derjenigen bestimmt, die ohne Entschuldigung nicht zur Klausur angetreten ist (p_{NE}). Dies sind die Fälle, die zusätzliche Kosten verursachen, denn sie werden bei der Raumvergabe und dem Druckauftrag mitberücksichtigt. Eine genauere Analyse der dadurch entstehenden Kosten finden sich in (Rybizki, 2013).

Semester	Anteil (p_{NE})
SS12	15,86 %
WS12/13	11,68 %
SS13	14,53 %

Tabelle 3: Anteil an angemeldeten Studierenden, die unentschuldigt nicht zur Klausur erschienen sind.

Tabelle 3 zeigt diesen Anteil über drei vergleichbare Semester hinweg. Ein Anteilswertdifferenzentest bestätigt den signifikanten Rückgang von p_{NE} von SS12 gegenüber WS12/13 (p-Wert 0,0191), gefolgt von einem signifikanten Anstieg des Anteilswerts von WS12/13 auf SS13 (p-Wert 0,0319). Dieser Rückgang fällt mit der Einführung der Online-Aufgabenserien im WS12/13 zusammen. Eine mögliche Erklärung hierfür ist, dass sich in diesem Semester mehr Studierende als sonst gut auf die Klausur vorbereitet fühlten und nicht am Tage der Klausur ohne Entschuldigung fernblieben. Auch ist denkbar, dass die Entscheidung zur Klausur anzutreten oder nicht, wegen der Online-Aufgabenserien früher getroffen wurde, zu einem Zeitpunkt, an dem eine Abmeldung noch möglich war.

3.2 Kontinuierliche Mitarbeit

Das Projekt genießt, wie bereits erwähnt, unter den Studierenden eine hohe Akzeptanz. Um diese als Teil der übergeordneten qualitativen Auswertung, nicht zu gefährden wurde darauf verzichtet, weitere personenbezogene Daten zu erheben. Das bedeutet insbesondere, dass die Datengrundlage der Analyse nicht aus einem statistischen Experiment gewonnen wurde. Stattdessen wird mit den Daten gearbeitet, die ohnehin zur Verfügung stehen. Diese Herangehensweise liefert naturgemäß ein beschränktes Analyseumfeld. Für das formulierte Ziel der vorliegenden Arbeit erweist sich die Grundlage dennoch als ausreichend.

3.2.1 Datengrundlage

Die 1.123 Teilnehmer des ersten Tests können nicht alle berücksichtigt werden. Bisher unerwähnt blieb, da im Speziellen nicht relevant, dass es zwei verschiedene Klausurmodi gibt, die abhängig von Studiumsbeginn und Studiengang zu belegen sind. Teilnehmern beider Klausurmodi wurde Zugang zu den Online-Augabenserien gewährt. Der teilnehmerstärkere und neuere der beiden beschließt das Modul Statistik mit einer Klausur, die 10 ECTS[2] umfasst. Dieses Modul richtete sich beispielsweise an Bachelor-Studierende der Wirtschaftswissenschaften. Im Folgenden werden nur die Ergebnisse der 10 ECTS Klausur betrachtet.

Die nächste Reduktion ergibt sich dadurch, dass nicht alle Studierenden, die an den Online-Aufgabenserien teilgenommen haben, auch an der Klausur teilnahmen. Zudem werden nur bestimmte Ergebnisse in die Analyse aufgenommen: Studierende die in der Klausur eine Punktzahl von maximal 5 Punkten erreichten (12,5% der erreichbaren Gesamtpunktzahl) wurden von der Analyse ausgeschlossen, genauso solche, die nicht in allen Online-Aufgabenserie jeweils mindestens 20% der erreichbaren Punkte erzielten. Die nachfolgenden Ergebnisse reagieren nicht sensitiv auf eine Variation dieser gewählten Grenzen, deshalb werden sie im Folgenden beibehalten.

Im Folgenden werden die Gründe für die genannte Festlegung der Grenzen erläutert: Die Mindestpunktzahlgrenze in der Klausur nimmt diejenigen Studierenden aus, die absichtlich wenige oder keine Teile der Klausur bearbeiteten, z. B. um an der Wiederholungsprüfung im Folgesemester trotz Auslandsaufenthalt teilnehmen zu können. Außerdem kann an der Ernsthaf-

[2]European Credit Transfer System

tigkeit des Versuchs gezweifelt werden, wenn ein Ergebnis von höchstens 5 Punkten erzielt wurde - die Bestehensgrenze beträgt 16 von 40 Punkte. Bei den Online-Aufgabenserien stellt die Selektion sicher, dass Studierende, die ohne ernsthafte Erfolgsabsicht teilnahmen (vgl. Abschnitt 2.1.3) nicht in die Analyse aufgenommen werden. Diese Selektionen sind wichtig, da nur die Ergebnisse betrachtet werden sollen, die auch tatsächlich vorhandene Fähigkeiten von Studierenden messen.

In den ersten beiden Reduktionsschritten gehen viele Instanzen verloren, die verbleibenden Reduktionen sind übersichtlicher. Schließlich bleibt ein Datensatz mit 417 Individuen, zu denen folgende Variablen vorliegen:

1. Geschlecht, Studiengang, Zahl der Semester seit Immatrikulation an der Universität, Zahl der Semester im Studiengang, Anzahl der Fehlversuche in der Klausur Statistik,

2. jeweilige Punktzahl in den vier Teilaufgaben der Klausur Statistik,

3. jeweilige Punktzahl in den drei Online-Aufgabenserien.

Die Variablen aus 2. und 3. stammen direkt aus den Ergebnissen von Klausur und Online-Aufgabenserien. Die Variablen aus 1. werden zu jeder Prüfung vom Prüfungsamt übermittelt.

Aus der Erfahrung vergangener Semester ist bekannt, dass die Variablen aus 1. einen signifikanten Beitrag zu den Klausurergebnissen liefern. Dabei gibt es eine hohe Korrelation von „Zahl der Semester seit Immatrikulation an der Universität“ und „Zahl der Semester im Studiengang“. Deshalb wird meist - auch im Folgenden - nur letztere betrachtet. Generell ist die Variable „Geschlecht“ eine Ausnahme. Innerhalb einzelner Aufgaben gibt es gelegentlich Geschlechterdiskriminierung. Über die Gesamtpunktzahl der Klausur finden sich allerdings in der Vergangenheit keine generellen Vorteile für weibliche oder männliche Studierende. Die Variable wird in der nachfolgenden Analyse dennoch berücksichtigt. Zwar wird das Item „Geschlecht“ für den Einfluss der Voraussetzungen auf den Lernerfolg keine Rolle spielen, für den Einfluss auf die Mitarbeit kann dies nicht geschlossen werden. Im Allgemeinen ist die Vorhersagekraft dieser Modelle eher mäßig. Für das vorliegende Problem bedeutet dieses historische Wissen, dass mit den angesprochenen Variablen aus 1. eine latente Variable gemessen wird, die mit „Voraussetzungen“ benannt wird. Zu beachten ist, dass dies keine individuellen Voraussetzungen sind, die über Items, wie „Abiturnote“ oder „Note in der Grundlagenveranstaltung Mathematik“ abgefragt werden. Es handelt sich um eingebrachte

Voraussetzungen, die die Studiensituation beschreiben, da die Studiengänge unterschiedliche Vorgaben machen und jemand, der im höheren Semester einen Zweitversuch ablegt eine andere Planung vornimmt als jemand im Erstversuch, der probeweise die Klausur in ein früheres Semester vorzieht.

Die Variablen aus 2. betreffen die Klausur, die den Lernerfolg am Ende des Semesters abfragen soll. Dementsprechend werden diese Variablen als Items für die latente Variable „Lernerfolg“ herangezogen.

Die Variablen aus 3. betreffen die Punkte in den Online-Aufgabenserien. Diese manifesten Variablen werden einer latenten Variable zugeordnet, die „Mitarbeit“ genannt werden kann, wenn die beschriebenen Manipulationsmöglichkeiten nicht zu exzessiv genutzt wurden. Die zu klärende Frage ist demnach, ob diese latente Variable die Qualität der eigenständigen Mitarbeit misst oder die einer hinzugezogenen Hilfestellung.

3.2.2 Modell

Abbildung 3 zeigt das Pfaddiagramm des Messmodells, das sich aus den Vorüberlegungen ergibt. Darüber hinaus ist das Strukturmodell abgebildet mit allen Zusammenhängen, die sich aus dem zeitlichen Ablauf ergeben: Alle Teilnehmer kommen mit gewissen Voraussetzungen in die Veranstaltung, während des Semesters arbeiten sie nach eigenem Ermessen und Möglichkeiten mit und schließen die Veranstaltung mit der Klausur ab, deren Ergebnis den Lernerfolg widerspiegelt.

Korrelationen zwischen den manifesten Variablen wurden der Übersicht halber nicht eingetragen. Diese ergeben sich beispielsweise aus struktureller Beschaffenheit einzelner Studiengänge oder inhaltlichen Übereinstimmungen von bestimmten Aufgaben in Klausur und Online-Aufgabenserien. Alle in der Analyse berücksichtigten Korrelationen wurden auf ihren unterstützenden Beitrag zur Modellgüte mittels Modifikationsindizes nach Sörbom (1989) geprüft.

Die Pfade in Abbildung 3 wurden bis auf die im Strukturmodell nicht benannt. Diese spielen in der Analyse die zentrale Rolle:

- Die subjektive Erfahrung der Vergangenheit führt zu der Forderung, dass γ_2 einen positiven Wert annehmen sollte.
- Wenn die Online-Aufgabenserien einen Beitrag zur Lehrverbesserung geliefert haben, sollte β_1 ebenfalls einen positiven Wert annehmen.

Um den Aufwand der Online-Aufgabenserien zu rechtfertigen, wird ein Wert gefordert, der mindestens so groß ist wie γ_2.

- Ein negativer Wert von γ_1 spräche dafür, dass Studierende mit ungünstiger Ausgangssituation mehr Punkte in den Online-Aufgabenserien erzielten, als solche mit günstiger Ausgangssituation. Da dieses Projekt neu initiiert wurde und alle Studierenden gleichermaßen hohes Interesse an dieser Verbesserungsmöglichkeit zeigten, scheint es unplausibel davon auszugehen, dass bessere Studierende bei der Teilnahme weniger konzentriert arbeiteten. Es wird ein positiver Wert gefordert.

- Um Verzerrung durch weitere[3] Auslassungen zu vermeiden, wird γ_3 mitgeschätzt. Im Sinne der Lehrverbesserung durch Eigenarbeit ist das Vorzeichen dieses Parameters nicht festgelegt. Das Vorzeichen wird stattdessen zeigen, welche Gruppe von den Online-Aufgabenserien besonders profitieren konnte. Dieser geschätzte Wert liefert somit eine Zusatzinformation.

Ausgehend von der Annahme, dass Studierende mehrheitlich ehrliche Prüflinge sind und im Status quo ante keine Manipulationsmöglichkeiten nutzten, ist die zu prüfende Nullhypothese verbal formuliert:

„Die Manipulationsmöglichkeiten wurden nur in Einzelfällen genutzt und die Online-Aufgabenserien haben die Qualität der Eigenarbeit der Studierenden gemessen."

In der numerischen Analyse wird daraus:

$$\mathrm{H}_0: \quad \gamma_1 > 0,\ \beta_1 > \gamma_2 > 0{,}3 \tag{1}$$

Die Grenze von 0,3 wurde nach folgender Überlegung festgelegt: Hauptaugenmerk liegt auf der Forderung $\beta_1 > \gamma_2$, da daraus die Bedeutung des Faktors „Mitarbeit" abgeleitet wird und insbesondere das Projekt als Lehrverbesserung angesehen werden kann. Um die Größe von β_1 mit γ_2 vergleichen zu können, werden die latenten Variablen auf eine Varianz von 1 standardisiert[4]. Bei standardisierten Variablen liefert ein linearer Zusammenhang von

[3] Zumindest individuelle Fähigkeiten spielen vermutlich eine wesentliche Rolle, die nur bedingt über die latente Variable „Voraussetzungen" aufgefangen werden kann.

[4] und nicht, wie üblich, den Einfluss der jeweils führenden manifesten Variable

0,3 einen beträchtlichen Erklärungsanteil und ist dabei noch nicht so groß, dass kein Spielraum mehr für weitere Zusammenhänge verbleibt.

Für die Aufnahme des Moderatoreneffektes wurde die Übersicht für entsprechende Herangehensweisen von Cortina et al. (2001) herangezogen. Als Kompromiss aus Umsetzbarkeit und Leistungsfähigkeit wurde der Ansatz von Jöreskog and Yang (1996) ausgewählt. Die Auswertung erfolgte in `R` unter Verwendung der Pakete `lavaan` von Rosseel (2012) und `MVN` von Korkmaz (2013).

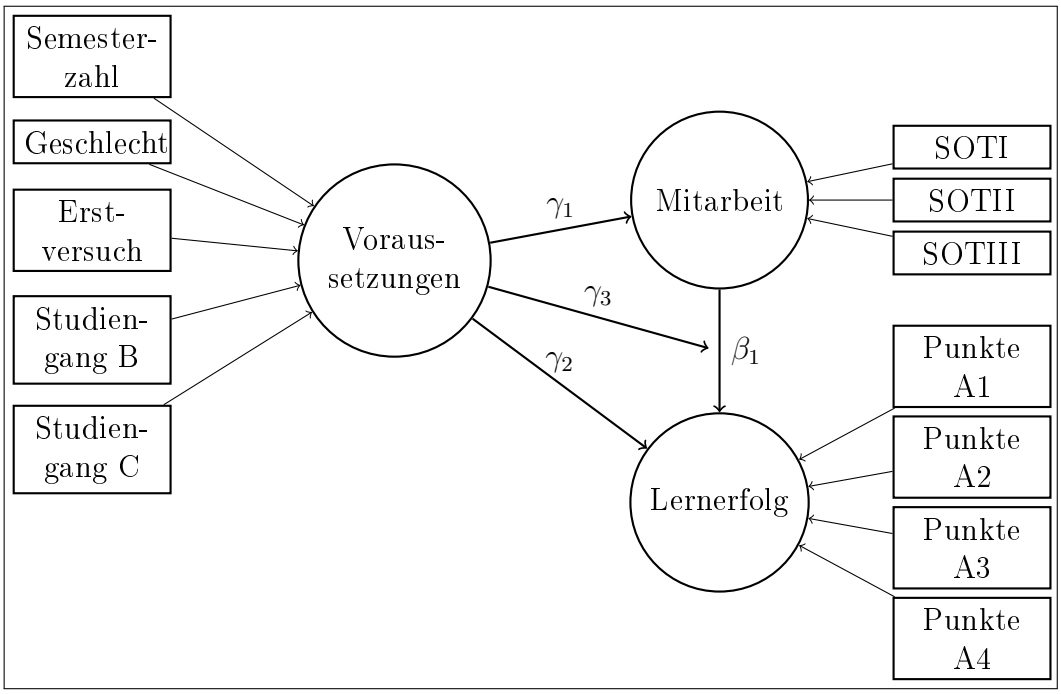

Abbildung 3: Pfaddiagramm des unterstellten Gesamtmodells ohne Korrelationen.

3.2.3 Auswertung

Der LISREL-Analyse wird die Frage nach multivariater Normalverteilung der verwendeten Daten vorangestellt. Eine mögliche Herangehensweise wurde von Mardia (1974) vorgestellt. Dabei wurden Tests konstruiert, die auf Schiefe und Kurtosis basieren. Beide Tests liefern zur Nullhypothese der

Anpassungsmaß		Anpassungsgüte
RMSEA	0,05	gut
p-Wert (RMSEA $<$ 0,05)	0,47	gut
90%KI	0,04 0,06	akzeptabel
SRMR	0,05	gut
CFI	0,95	akzeptabel

Tabelle 4: Beurteilung der Anpassungsgüte mittels Root Mean Square Error of Approximation, Standardized Root Mean Square Residual und Comparative Fit Index.

Normalverteiltheit der vorliegenden Daten einen p-Wert von 0,000. Ein weiterer Test auf multivariate Normalverteiltheit stammt von Henze and Zirkler (1990), auch hier wird die Nullhypothese abgelehnt (p-Wert $< 2{,}2\, e^{-16}$).

Für die Anpassung des Modells wird der Standardschätzer (sogenannter ML-Schätzer) verwendet, der eine likelihoodbasierte Diskrepanzfunktion minimiert. Die Standardabweichungen werden dabei robust geschätzt - auch gebootstrapte Standardabweichungen (50.000 Bootstrap-Stichproben) werden angegeben um der fehlenden Normalverteilungsannahme gerecht zu werden. Damit wird die Anpassungsgüte des Standardschätzers bei geringen Beobachtungszahlen ($<$ 1.000) trotz fehlender Begründung der Normalverteiltheitsannahme, kombiniert mit der verlässlichen Schätzung der Standardabweichungen (vgl. Punkt **Multivariate Normality** in Cortina et al. (2001) und dort zitierte Literatur).[5]

Modellanpassung

Die betrachteten Anpassungsmaße werden interpretiert nach den empfohlenen Daumenregeln von Schermelleh-Engel et al. (2003).
Weiterhin wurde keine Variablenvarianz negativ geschätzt und auch die Kovarianzen haben die erwarteten Vorzeichen. Die Ladungen der manifesten auf die latenten Variablen sind mäßig. Während die Punkte in den Klausuraufgaben akzeptabel laden ($\sim$ 0,6), gilt dies bei den Punktzahlen der Online-

[5]Im `R`-Paket `lavaan` kann mittels `GLS` oder `WLS` geschätzt werden. Die entsprechenden Schätzungen liefern Ergebnisse, die so wenig von der vorliegenden Schätzung abweichen, dass die Interpretation gleich bleibt. Daher werden hier die Standard-Ergebnisse diskutiert.

	Schätzer	Std.-Fehler (robust)	Std.-Fehler (BS)
Mitarbeit ∼			
Voraussetzungen	$\widehat{\gamma}_1 =$ 0,220	0,079	0,084
Lernerfolg ∼			
Voraussetzungen	$\widehat{\gamma}_2 =$ 0,419	0,081	0,095
Mitarbeit	$\widehat{\beta}_1 =$ 0,458	0,112	0,131
Voraussetzungen:Mitarbeit	$\widehat{\gamma}_3 =$ −0,157	0,062	0,067

Tabelle 5: Ergebnisse der Parameterschätzung und Standardfehler für das Strukturmodell (s. Abbildung 3).

Aufgabenserien nur für SOTI. Schlusslicht bildet das Geschlecht, das nur mit 0,1 auf die „Voraussetzungen“ lädt. Insgesamt wird damit die Anpassung noch als hinnehmbar eingestuft.

Parameterschätzung

Die Schätzung erfolgt mit der likelihoodbasierten Diskrepanzfunktion auf Grundlage der Korrelationsmatrix und mit normierten latenten Variablen. Es werden nicht die Standardschätzer für die Standardabweichung angegeben. Diese liegen etwas unterhalb der angegebenen robusten Schätzer. Das Paket `lavaan` bietet mehrere Möglichkeiten, wobei die klassischen Schätzer verwendet wurden.[6] Die Schätzwerte von γ_1, γ_2 und β_1 liegen deutlich in dem gewünschten Bereich. Werden jedoch aus den Standardfehlern einseitige 95%-Konfidenzintervalle zu den Schätzwerten von γ_2 und β_1 gebildet, liegt die Untergrenze unterhalb der 0,3 - bei 90%-Konfidenzintervallen etwas darüber. Dies spricht zunächst gegen die zu prüfende Hypothese der Nichtnutzung von Manipulationsmöglichkeiten bzw. die Fähigkeit der Online-Aufgabenserien, die Qualität der Eigenarbeit der Studenten zu messen. Dieses sequentielle Vorgehen entspricht dabei nicht dem exakten Vorgehen zum Testen der beschriebenen Nullhypothese, sondern liefert ein Indiz zur Einschätzung der Ergebnisse. Ein Test, der sich zur Beantwortung dieser Fragestellung eignet,

[6] Damit werden die Likelihoodbasierten Schätzer bezeichnet, die mit den Funktionsparametern `MLM, MLMV, MLMVS` aufgerufen werden.

	Typ B -Test ($H_0^{>}$)	Typ A - Test ($\overline{H}_0^{=}$)
LR statistic	0,00	18,31
Adjusted p-value	0,60	0,02

Tabelle 6: Wert der Teststatistik und p-Wert zur Nullhypothese (1).

folgt.

Das Ergebnis der Schätzung von γ_3 ist ebenfalls interessant. Der negative Wert (auch das 95%-Konfidenzintervall bleibt negativ) spricht dafür, dass es die als schwächer eingestuften Studenten sind, die von den Tests überproportional profitieren konnten. Es scheint, sie konnten die Erfolge in den Online-Tests vermehrt in einen Lernerfolg umsetzen.

Nullhypothese mit Ungleichungsbedingungen

Die zu prüfende Hypothese (1), mit der die Zielsetzung der vorliegenden Arbeit geprüft wird, wird nicht mit einzelnen Konfidenzintervallen getestet, wie das oben angedeutet wurde. Stattdessen schlagen Van de Schoot et al. (2010) einen Bootstrap-Test auf Basis der Likelihooddifferenzen genesteter Modelle vor. Anstelle des vorgesehenen parametrischen Bootstraps wird eine Methode nach Bollen and Stine (1992) verwendet, um der fehlenden Normalverteiltheit Rechnung zu tragen.

Verwendet wurden 50.000 Bootstrap-Züge. Dabei ergab sich: Es wird zum einen die Nullhpothese H_0 getestet, indem dem durch die Ungleichungen restringierten Modell das unrestringierte Modell gegenübergestellt wird – der Typ B-Test. Das Hypothesenpaar kann nicht umgedreht werden, da resampling basierend auf dem unrestringierten Modell nicht möglich ist. Stattdessen wird ein Modell gebildet, indem die Ungleichungen zu Gleichungen umgeschrieben werden. Auf diesem Modell basierend ist resampling möglich und ein Vergleich mit dem ungleichungsrestringierten Modell liefert den Typ A-Test.

Der Typ A-Test gibt an, wie weit die geschätzten Parameter in den Ungleichungsrestriktionen liegen. Der niedrige p-Wert spricht bereits für H_0. Mehr Gewicht liegt auf den Typ B-Test, der direkt die formulierte Nullhypothese H_0 testet und für gängige Testniveaus nicht ablehnt.

Da dieser Test speziell für derartig gestaltete Nullhypothesen entwickelt wurde, wiegt dieses Testergebnis schwerer als die aus den Konfidenzintervallen abgeleiteten Einzeltestentscheidungen. H_0 wird als nicht abgelehnt betrachtet.

3.2.4 Ergebnis

Ein schwerwiegendes Problem stellen die teilweise niedrigen Ladungen der Online-Aufgabenserien dar. Insgesamt sind die Ergebnisse dennoch verwertbar.

Die Analyse stärkt die Hoffnung, dass der vorhandene Missbrauch der Lehrverbesserung eine deutlich geringere Rolle spielt als der tatsächlich generierte Lernerfolg. Die geschätzten Parameter nehmen Werte an, die eine plausible Interpretation zulassen. Die eigenständige Mitarbeit hat einen positiven Einfluss auf den Lernerfolg, der stärker ist als die studiumsbedingten Voraussetzungen. Gute Voraussetzungen beeinflussen auch die Mitarbeit positiv. Interessant ist, dass Studierende mit tendenziell schlechten Voraussetzungen überproportional von der Teilnahme an den Online-Aufgabenserien profitieren.

Dieser Moderatoreffekt muss wegen der großen Varianz vorsichtig interpretiert werden, allerdings spricht das negative Vorzeichen deutlich dagegen, dass schwache Teilnehmer effektiv auf externe Hilfe zurückgreifen konnten. Sonst wäre es ihnen schließlich nicht gelungen, hohe Scores in der Mitarbeit in hohe Scores beim Lernerfolg zu übertragen.

3.3 Ausblick

In diesem Abschnitt möchten die Autoren einen kurzen Überblick über das Verbesserungspotential geben. Wie zu Beginn der Abschnitte 1 und 2 erwähnt war die Durchführung ein Pilotprojekt und daher auch ein Lernprozess. Die gesammelten Erfahrungen sollten in einem vergleichbaren Projekt während eines anderen Semesters angewandt werden um ein homogeneres Aufgabenserien-Design und damit einen homogeneren Datensatz zu erhalten. Um die in Abschnitt 3.1 aufgestellte Vermutung zu überprüfen müsste bei einer etwaigen Wiederholung der Online-Aufgabenserien der zeitliche Verlauf der Abmeldungen erfasst werden, damit Aussagen über eventuelle kausale Zusammenhänge getroffen werden können. Weiter können neue Strategien zur Vermeidung von Manipulationsmöglichkeiten etabliert und deren Wirk-

samkeit in einem weiteren Durchlauf während eines Semesters untersucht werden, um die qualitative Verifizierung aus Abschnitt 2.5 abzusichern.

4 Zusammenfassung

Es wurde ausführlich ein Anreizsystem zur Verbesserung der individuellen Mitarbeit beschrieben, sowie auf dessen potentielle Schwachstellen hingewiesen. Mit diesem Anreizsystem kann auch die Lehre von Fächern verbessert werden, bei denen die Teilnehmerzahl so groß ist, dass eine persönliche Kontaktaufnahme nicht möglich ist.

Die durchgeführten Analysen liefern zwar diskussionswürdige Ergebnisse (Datenbasis bei beiden Zielen gering, bei Hauptziel zusätzlich: niedrige Ladungen, hohe Standardfehler), spricht nach Ansicht der Autoren dennoch dafür, dass die formulierten Ziele weitgehend erreicht wurden. Insbesondere scheint die Mehrheit der hier betrachteten Studierenden ohne Manipulation von den Online-Tests profitiert zu haben. Die eingangs beschriebenen Zweifel an der Sinnhaftigkeit einer Fortführung des Projekts können damit so weit zerstreut und damit das Gewicht in der qualitativen Gegenüberstellung der Argumente so weit reduziert werden, dass die positiven Argumente überwiegen.

Deshalb spricht aus Sicht der Autoren nichts dagegen, das Projekt in vergleichbarer Form fortzusetzen oder zu imitieren. Allerdings gilt diese Empfehlung nur bei Kontrolle der Ergebnisse und wird ergänzt um die Empfehlung, Kontrollmechanismen einzubauen bzw. auszuweiten.

Literatur

Bollen, K. A. and Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods and Research*, 21:205–229.

Cortina, J. M., Chen, G., and Dunlap, W. P. (2001). Testing interaction effects in lisrel: Examination and illustration of available procedures. *Organizational Research Methods*, 4(4):324–360.

Henze, N. and Zirkler, B. (1990). A class of invariant consistent tests for mul-

tivariate normality. *Communications in Statistics - Theory and Methods*, 19(10):3595–3618.

Jöreskog, K. G. and Yang, F. (1996). Nonlinear structural equation models: The kenny-judd model with interaction effects. In Marcoulides, G. and Schumaker, R. E., editors, *Advanced structural equation modeling: Issues and techniques*, pages 57–87. Mahwah, NJ: Lawrence Earlbaum Associates.

Korkmaz, S. (2013). *MVN: Multivariate Normality Tests*. R package version 1.0.

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis for testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics*, 36:115–128.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Rybizki, L. (2013). Construction of cost sensitive binary classification rules accounting for measurenment and uncertain misclassification costs. Dissertation, *Lehrstuhl für Statistik und Ökonometrie, FAU Erlangen-Nürnberg*.

Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2):23–74.

Sörbom, D. (1989). Model modification. *Psychometrika*, 54(3):371–384.

Van de Schoot, R., Hoijtink, H., and Dekovic, M. (2010). Testing inequality constrained hypotheses in sem models. *Structural Equation Modeling*, 17:443–463.