FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

FACHBEREICH WIRTSCHAFTS-
WISSENSCHAFTEN

Lehrstuhl für Statistik und Ökonometrie

Diskussionspapier
91 / 2014

Verbesserung des Lernverhaltens durch Online-Tests – ein Jahr später

Benedikt Mangold
Thomas Pleier
Christoph Brug
Jan Nolzen
Johannes Stübinger

Lange Gasse 20 · D-90403 Nürnberg

Verbesserung des Lernverhaltens durch Online-Tests – ein Jahr später

Benedikt Mangold* Thomas Pleier† Christoph Brug‡
Jan Nolzen§ Johannes Stübinger¶

FAU Erlangen-Nürnberg
Lange Gasse 20
D-90403 Nürnberg

3. November 2014

Abstract

In Pleier und Mangold (2013) wurde ein Pilotprojekt zur Verbesserung des Lernverhaltens untersucht, welches im Wintersemester 2012/2013 durchgeführt wurde. Studierende sollten dabei semesterbegleitend an Online-Tests teilnehmen, um sich frühzeitig auf die abschließende Klausur vorzubereiten. Als Anreiz zur Teilnahme diente die Aussicht auf Bonuspunkte, welche für jene Klausur im Vorfeld erworben werden konnten. Nun, ein Jahr später, soll untersucht werden, ob sich die Ergebnisse auf das Wintersemester 2013/2014 übertragen lassen und welchen Einfluss mögliche Manipulationen durch die Studierenden besitzen.

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

*Benedikt.Mangold@fau.de
†Thomas.Pleier@fau.de
‡Christoph.Brug@fau.de
§Jan.Nolzen@fau.de
¶Johannes.Stuebinger@fau.de

1 Einleitung

Ziel dieses Artikels ist es, die Effekte der Nacharbeit[1] des in der Vorlesung behandelten Stoffes auf den Lernerfolg[2] zu messen. Im Rahmen der Veranstaltung *Statistik* des Lehrstuhls für Statistik und Ökonometrie der Friedrich-Alexander-Universität Erlangen-Nürnberg wurde bereits im Wintersemester 2012/2013 (WS 12/13) ein Projekt gestartet. Bei diesem konnten Studierende an drei Terminen im laufenden Semester an Online-Tests teilnehmen, in denen der aktuelle Wissensstand abgefragt wurde. Als Anreiz teilzunehmen dienten Bonuspunkte für die Klausur am Semesterende, die die Teilnehmenden je nach Erfolg in den Online-Tests ansammeln konnten. Analysen und deren Ergebnisse finden sich in Pleier und Mangold (2013). In der vorliegenden Arbeit wird die Motivation zu einem solchen Projekt hervorgehoben und die Daten ausgewertet, die ein Jahr später, im Wintersemester 2013/2014 (WS 13/14), erhoben wurden. Darüber hinaus werden die Änderungen bewertet, die nötig wurden, um die Schwachstelle des ersten Durchlaufs - Manipulatiosversuche beim Online-Test von Studierenden - zu reduzieren.

In der Literatur wird zwischen zwei Typen von Übungen unterschieden (Bourne und Ekstrand, 2005, S. 180):

> „Typ I ist die wiederholende Übung um zu verhindern, dass ein Inhalt aus dem Kurzzeitgedächtnis verschwindet. Er wird dabei in keinster Weise modifiziert und bleibt unverändert, bis man ihn verwendet.[..] Erst die Einübung vom Typ II (modifizierende Einübung) verbindet neuen Inhalt mit bereits im Langzeitgedächtnis gespeichertem Wissen."

Das Ziel universitärer Veranstaltungen ist Wissensfestigung und Wissenseinübung vom Typ II. Die Erfahrungen der Vergangenheit zeigen jedoch, dass Studierende eher zum Einübungsprozess vom Typ I neigen. Ein Ziel des Projekts war deshalb, die Studierenden zu einem anderen Lernverhalten zu motivieren und durch abgewandelte Fragestellungen in den Online-Tests direkt zu unterstützen.

Auch Mietzel (1993) (S. 184) geht speziell auf die Einprägungsarbeit des Lernenden (unterstützt von geeignetem Lernmaterial, das bereits einen op-

[1] das eigenverantwortliche, kontinuierliche Nach- und Aufbereiten der Vorlesungsinhalte

[2] damit ist nicht gemeint die Prüfung lediglich zu bestehen, sondern die Inhalte in weiterührenden Veranstaltungen bzw. im späteren Berufsleben anwenden zu können

timalen Organisationsgrad aufweist), ein:

> „Je länger und intensiver sich der Lernende mit dem einzuprägenden Material auseinandersetzt, desto größer ist die Wahrscheinlichkeit, dass zu einem späteren Zeitpunkt ein schneller und sicherer Abruf aus dem Gedächtnis erfolgen kann. – Eine Einprägungsphase ist somit als effektiv zu kennzeichnen, wenn der Lernende motiviert ist, dem (möglichst gut organisierten) Gedächtnismaterial und seiner Verarbeitung hohe Aufmerksamkeit entgegenzubringen."

In der abschließenden Klausur ist Zeit meist ein wichtiger Faktor – nur wenn die Inhalte gut aufbereitet und verstanden worden sind, können sie schnell abgerufen werden.

Artelt (2000) vertritt die These, dass die Intensität von Beschäftigung mit einem Inhalt auf die Fähigkeit diesen im Gedächtnis zu behalten wirkt und bereits von der Herangehensweise an den Inhalt beeinflusst wird.

Eigenverantwortliche Nacharbeit unterstüzt den Prozess des Einübens beider Typen und verlängert die Zeit der Auseinandersetzung mit den Inhalten. Somit kann die Nacharbeit als wichtiger Einfluss auf den Lernerfolg angesehen werden. Dabei war es nicht das Ziel des Projekts das eigenverantwortliche Lernen zu ersetzen – vielmehr sollten die Studierenden immer wieder zurück an den Schreibtisch gelockt werden, um von der positiven Wirkung der Nacharbeit auf den Lernerfolg zu profitieren.

Der Großteil der Artikel zu diesem Thema beschäftigt sich mit verschiedenen Anreizen: Wann nehmen Studierende/Schüler an Online-Tests teil? Kremer et al. (2009) untersuchten die Wirkung einer Schulgeldbefreiung bei gutem Abschneiden im Examen und finden positive Effekte bei Schülerinnen, die in einem vorherigen Examen schlecht abgeschnitten hatten.
Fryer (2011), der die Wirkung finanzieller Anreize auf den Lernerfolg untersucht, kommt zu dem Schluss, dass diese Art von Anreiz nur dann den Lernerfolg steigert, wenn für bereits im Vorfeld erbrachte Leistung eine Belohnung erhalten wird – ist die Belohnung an den Lernerfolg selbst gekoppelt findet er keine Effekte.
Kibble (2007) untersuchte im Rahmen eines Physiologie-Kurses die Effekte der unbeaufsichtigten Teilnahme an einem Onlinequiz auf den Lernerfolg bei Medizinstudenten. Auch hier schnitten Studierende, welche sich zur Teilnahme an diesen Online-Tests entschlossen hatten, in den Examina besser ab.

Allerdings rät Kibble von einer zu hohen Belohnung bei Teilnahme ab, da dann auch die Bereitschaft des *Cheating* zunimmt.
Neben weiteren Arbeiten über finanzielle Anreize (Angrist et al. (2007), Patel und Richburg-Hayes (2012), u.a.) variieren Luehrmann et al. (2013) die Form der Anreize und untersuchen unter anderem die Wirkung des Einsatzes nicht-finanzieller Belohnung auf die Steigerung des Lernerfolges. Sie kommen zu dem Schluss, dass diese Belohnung effektiv die Nacharbeit und damit den Lernerfolg der Studierenden steigert.
Andere Artikel, wie die von Angus und Watson (2009) oder Woit und Mason (2003), thematisieren die Wirkung von Online-Tests und zeigen einen positiven Effekt.
Zu erwähnen ist noch der Artikel von Rowe (2004) – hier wird der Effekt von *Cheating* bei Online-Tests untersucht, was auch Bestandteil der vorliegenden Untersuchung ist (Abschnitt 3.4). Er warnt vor der Vielzahl von Möglichkeiten des *Cheating*, die durch den Einsatz von Online-Tests entstehen und beklagt die nur unbefriedigenden Gegenmaßnahmen.

Diesen Erkenntnissen folgend wurden die Tests auf einer Onlineplattform erstellt, welche die Pflichtveranstaltung Statistik in WS 12/13 und WS 13/14 begleiteten. Die dabei erzielbaren Bonuspunkte, die zu der regulär erreichten Punktzahl der Klausur addiert wurden, errechneten sich aus einer vorher erbrachten Leistung - den Online-Tests.

Nach dem Durchlauf im WS 12/13 befürworteten Pleier und Mangold (2013) die Fortsetzung des Projekts. Schlüsselargument war eine *quantitative* Analyse, da die *qualitative* Analyse eine gewisse Ambivalenz aufzeigte. Zum einen genoss das Projekt ein hohes Ansehen unter den Teilnehmern, was vor allem an dem positiven Feedback im Freitext der Vorlesungsevaluation zu erkennen war. Andererseits lieferte die freie Bearbeitung den Teilnehmern zu leicht die Möglichkeit zum *Cheating*[3] etwa dadurch, dass geeignete Kommilitonen die Bearbeitung der Online-Tests übernahmen.

Das Ausmaß der Manipulationen war unbekannt. Damit diesem Argument nicht zu viel - also unberechtigte - Bedeutung beigemessen wird, wurden die Daten nach Auffälligkeiten durchsucht. Dazu wurde beispielsweise die Verteilung der Bonuspunkte innerhalb der Notenstufen und die Anzahl an Teilnehmern, welche sich allein durch die Bonuspunkte auf die Notenstufe 4.0 (Bestehensgrenze) retten konnten, betrachtet. Ein weiteres bedeutendes

[3]im Folgenden als *Manipulation* bezeichnet

Argument war eine konfirmatorische Analyse. Dabei wurde überprüft, ob in den Ergebnissen der Teilnehmer ein positiver Effekt der Nacharbeit auf den Lernerfolg gefunden werden kann. Wäre dieser nicht erkennbar gewesen, hätte dies dafür gesprochen, dass das Ziel der intensivierten Nacharbeit nicht erreicht wurde, weil z.B. die Teilnehmer in großem Maße manipulierten.

Die Resultate dieser Analysen suggerierten, dass die Nacharbeit tatsächlich gefördert wurde. Außerdem zeigte sich ein starker Rückgang der kurzfristigen Abmeldungen von der Prüfung. Diese Ergebnisse führten dazu, die große Beliebtheit des Projekts als Argument in der qualitativen Analyse stärker zu gewichten als das Gegenargument (Manipulationsmöglichkeiten und die befürchtete Neigung der Teilnehmer diese zu Nutzen). Es wurde also die Fortsetzung des Projekts empfohlen, obwohl nach wie vor Zweifel über das Ausmaß an Manipulationen bestand. Aufgrund dieser Zweifel sollten bei einer erneuten Durchführung der Online-Tests in einem späteren Semester Kontrollmechanismen implementiert werden.

Diese Kontrollmechanismen wurden bei der Wiederholung im WS 13/14 bei nahezu gleichen Rahmenbedingungen umgesetzt: Eine zufällige Auswahl an Studierenden hatte nun den Test in einem Computerlabor unter Aufsicht durchzuführen. Damit konnte, zumindest teilweise, die Teilnahme unter Kontrolle gestellt werden. Abweichungen von den Ergebnissen der beaufsichtigten Gruppe mit kontrollierten Bedingungen und den übrigen Studierenden erlauben es, Rückschlüsse auf das Ausmaß an Manipulationen jenseits der Kontrolle zu ziehen.

Die Auswertung der Projektdaten lieferte ernüchternde Ergebnisse: Im Gegensatz zum WS 12/13 gab es kein positives Feedback. In einer kurzen Umfrage wurde klar, dass die Teilnehmer nur die Bonuspunkte erhalten möchten. An eine Förderung der Nacharbeit durch das Projekt konnten oder wollten sie nicht glauben und der Erfolg des Projektes insgesamt wurde angezweifelt. Außerdem berichteten Teilnehmer von den Manipulationsmethoden ihrer Kommilitonen. Dadurch fühlten sie sich benachteiligt, da sie ihre Durchgänge bei den Online-Tests, wie vorgesehen, ohne Manipulationen durchführten. Die Hoffnung mancher Teilnehmer, in den Online-Tests eine spezielle Vorbereitung auf die Klausuraufgaben zu erhalten - etwa: selbe Frage, andere Zahlen - wurde auch enttäuscht. Darüber beschwerten sich viele nach der Klausur. Insgesamt nahmen auch weniger Studierende an der Maßnahme teil als noch im Jahr zuvor. Zusammengefasst ergibt sich damit folgende Situation – alle Argumente sprechen gegen die Fortsetzung des Pro-

jekts:

- Das Ansehen des Projekts unter den Teilnehmern geht im Vergleich zum Vorjahr stark zurück.
- Der konkrete Sinn der Online-Tests als Mittel zur Vorbereitung auf die Klausur wird angezweifelt.
- Gemeldete Methoden von Manipulationen der Testergebnisse bestätigen diesen Verdacht aus dem Vorjahr.

Bereits an diesem Punkt der Analyse ist klar, dass das Projekt nicht fortgesetzt wird.

Während die inhaltliche Frage nach der Weiterführung des Projekts verneint werden kann, bleibt noch zu klären, wie sich die Änderung der Bedingungen auf das Lernverhalten ausgewirkt hat und in wie weit die analytischen Ergebnisse des Vorjahres reproduziert werden können.

Der Aufbau des Artikels ist wie folgt: In Abschnitt 2 werden die Unterschiede der Modalitäten der Teilnahme an den Online-Aufgabenserien sowie der Mechanismus zur Kontrolle auf Manipulationen vorgestellt und kurz auf den Aufbau der Aufgabenserien eingegangen. In Abschnitt 3 werden die durch Aufgabenserie und Klausur gesammelten Daten quantitativ ausgewertet und die Wirkungen und Zusammenhänge in einer LISREL-Analyse überprüft. Eine Zusammenfassung und Bewertung der Ergebnisse finden sich im schließenden Abschnitt 4.

2 Rahmenbedingungen

Im Gegensatz zum WS 12/13 waren Aufbau und Inhalt der Online-Aufgabenserien in WS 13/14 homogener, da bereits die Schwachstellen klar waren und aus Fehlern gelernt wurde. Dies resultierte in einer weitestgehend ähnlichen Verteilung der erreichten Punkte pro Online-Aufgabenserie (vgl. Abbildung 1). Insgesamt wurden ausgehend von den Teilnehmerzahlen aus WS 12/13 mit ca. 1.500 Studierenden gerechnet – tatsächlich fiel die Zahl wesentlich geringer aus: Lediglich 1.007 Studierende nahmen an der Klausur teil. Breite und Lage der Zeitslots der Online-Tests waren nahezu identisch zu denen im WS 12/13.

2.1 Möglichkeiten der Manipulation

Erwartet wurde, dass sich die Bereitschaft der Studierenden sich zu Lerngruppen zusammenzuschließen bzw. Fremdleistung zu akquirieren erhöht, da sich Informationen bezüglich Erfolg und Vorgehensweise herumgesprochen haben könnten. Diese Befürchtungen spiegelten sich in den Ergebnissen wider. So schrieben viele Studierende, die im Vorfeld die ausgezeichnete Leistung der vollen 2 Bonuspunkten erreichten, in der Klausur eine tendenziell schlechte Note (graue Felder in Tabelle 1, näher erläutert in Tabelle 2).

Um dieses Vorgehen zu untersuchen wurden bei jedem Test eine Anzahl zufällig ausgewählter Studierender festgelegt – diese mussten den Test unter Aufsicht in den örtlichen CIP-Pools[4] durchführen. Die Auswertung der so generierten Daten finden sich im Abschnitt 3.2.

2.2 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, insbesondere der aus WS 12/13, eingestuft – dies spiegelte sich auch im üblichen Punkteschnitt wider, vgl. Abbildung 2.

[4] dabei handelt es sich um Computer-Labore an der Universität, die im Rahmen von Lehrveranstaltungen benutzt werden und in der übrigen Zeit den Studierenden zur Verfügung stehen

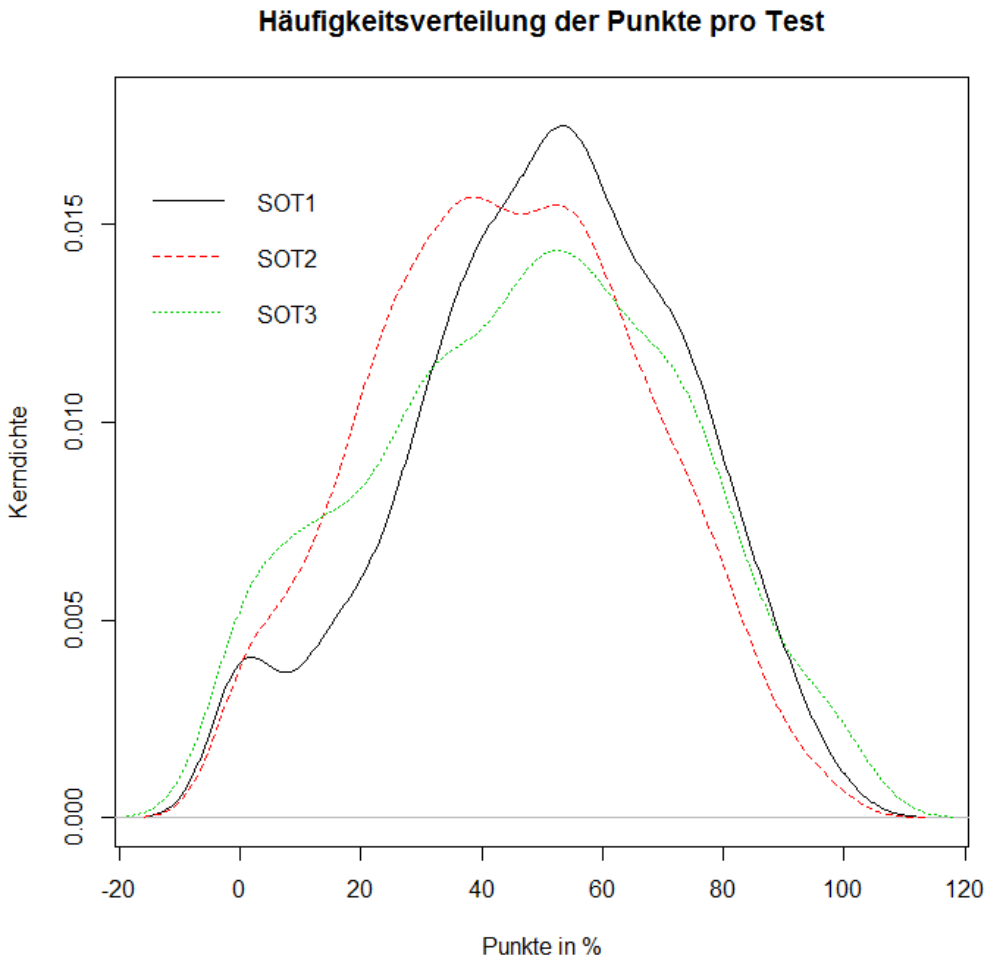


Abbildung 1: Kerndichteplot der erreichten Punkte je Online-Aufgabenserie.

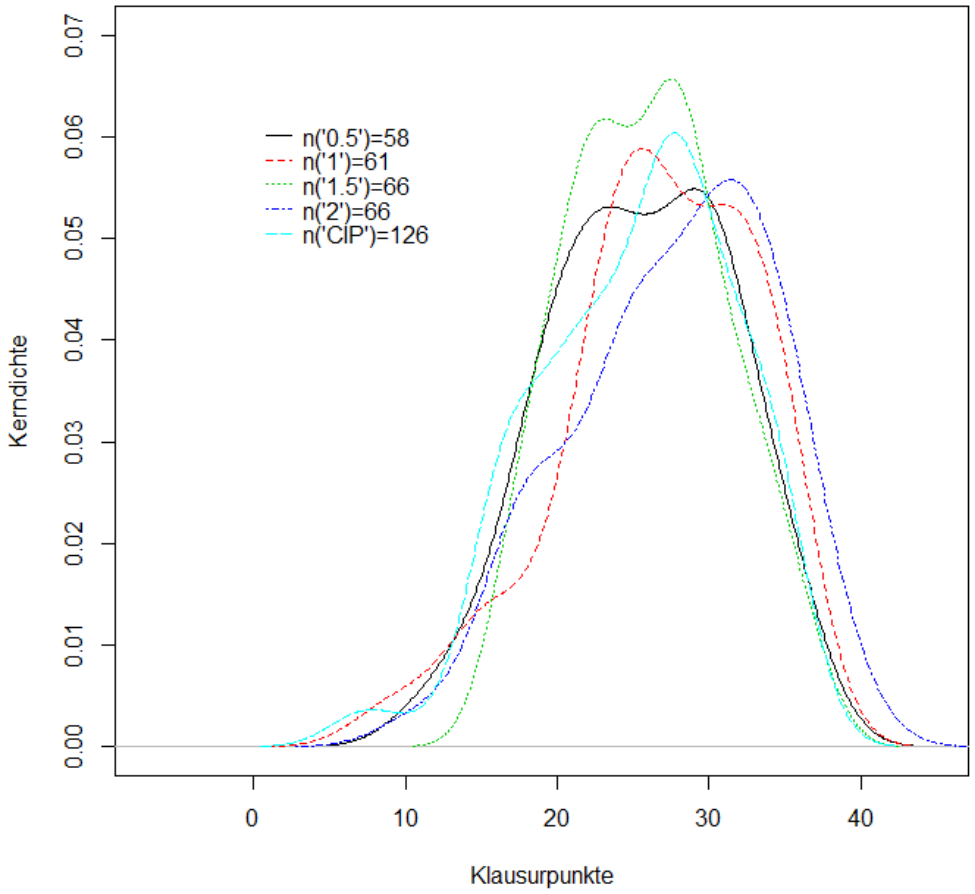


Abbildung 2: Kerndichteplot tatsächlich erreichter Klausurpunkte aufgeschlüsselt in zuvor erreichte Online-Test-Zusatzpunkte. $n('\cdot')$ bezeichnet die Anzahl der Studierenden, die sich '$\cdot$' Online-Test-Zusatzpunkte verdient hatten. 'CIP' zeigt die tatsächlich erreichten Klausurpunkte derer, die mindestens einmal in einem CIP-Pool arbeiten mussten.

		Erreichte Notenstufe				
		1	2	3	4	5
Online-Test Zusatzpunkte	2	0.106	0.082	0.044	0.024	0.004
	1,5	0.051	0.094	0.090	0.024	0.000
	1	0.070	0.079	0.063	0.020	0.008
	0,5	0.032	0.086	0.063	0.035	0.012
	0	0.000	0.000	0.004	0.004	0.008
	↑	0.422	0.229	0.140	0.117	

Tabelle 1: Relative Häufigkeiten der erreichten Notenstufe, aufgeschlüsselt in die zuvor erhaltenen Online-Test-Zusatzpunkte.
Die Zeile ↑ beinhaltet den Anteil an Studierenden innerhalb der Notenstufe der jeweiligen Spalte, die sich aufgrund der Zusatzpunkte entweder in die jeweiligen Notenstufe oder innerhalb der Notenstufe um eine Teilnote verbessert haben. Nähere Erläuterungen zum dunkelgrau unterlegten Feld in Tabelle 2.

	Online-Test-Zusatzpunkte			
	2	1,5	1	0,5
Notenstufe 4 wegen Zusatzpunkten	8	10	4	9

Tabelle 2: Nähere Erläuterung der dunkelgrauen Zelle in Tabelle 1: 11.7% entsprechen 31 Studierenden, die sich in die/innerhalb der Notenstufe 4 verbessert hatten. 12 davon hatten nur durch die Zusatzpunkte die Klausur bestanden.

3 Quantitative Analysen

Dieser Abschnitt behandelt die Auswertungen der Daten, die im Zusammenhang mit der Durchführung und der Kontrolle der Online-Tests erhoben wurden. Hauptaugenmerk liegt dabei auf der Bestätigung der Ergebnisse aus dem Vorjahr aus Pleier und Mangold (2013) – gleichzeitig soll dabei auch auf Unterschiede zwischen freier und kontrollierter Teilnahme eingegangen werden. Dazu wurden auch neue Datenquellen herangezogen anhand derer die Häufigkeit der Nutzung online bereitgestellter Materialien gemessen wurde.

3.1 Nutzung bereitgestellter Materialien

Um die Nacharbeit zu unterstützen, wird den Studierenden eine Vielzahl von Unterlagen zum Herunterladen bereitgestellt. Für das WS 13/14 stehen die täglichen Downloadzahlen zur Verfügung, welche als Näherung der Nutzung dieser Hilfsmittel zum jeweiligen Zeitpunkt interpretiert werden.

Es handelt sich um sich Zeitreihen auf Tagesbasis. Sie weisen unterschiedliche Besonderheiten auf, sodass unterschiedliche Modelle herangezogen werden. In den stabilisierten Zeitreihen überwiegen manchmal ARMA-Effekte, manchmal GARCH-Effekte. Um eine homogene Analyse durchführen zu können, wurde der gemeinsame ARMA-GARCH-Ansatz verwendet, den das `R`-Paket `rugarch`[5] bereitstellt.

In den Zeitreihen werden Auffälligkeiten gesucht, insbesondere die Phasen vor den Tests werden nach den Vorgehen aus Chang et al. (1988), Doornik und Ooms (2005) und Ardelean (2014) nach Ausreissern durchsucht.

3.1.1 Klausurensammlung

Diese Zeitreihe beinhaltet die tägliche Anzahl an Zugriffen auf online zur Verfügung gestellten Altklausuren. Abgesehen von vereinzelten Zugriffen findet der Großteil aller Zugriffe kurz vor der Klausur statt. Dies ist nicht überraschend und es werden keine Besonderheiten in der zeitlichen Umgebung der Online-Tests festgestellt.

[5] s. Ghalanosl (2014)

3.1.2 Hilfsmittel

Hilfestellungen zur Nutzung und Installation von in der Vorlesung benutzter Software und Tabellensammlungen werden hauptsächlich zu Beginn des Semesters und kurz vor der Klausur heruntergeladen. Die einzige Auffälligkeit in der Zeitreihe findet sich im Vorfeld des ersten Tests. Womöglich wurde mit Beginn der Online-Tests die Notwendigkeit einer ersten Auseinandersetzung mit den Hilfsmitteln erkannt.

3.1.3 Übungsblätter

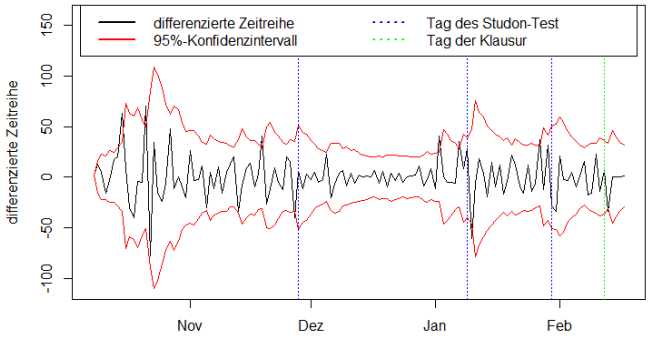
Auch die zur wöchentlichen Bearbeitung herausgegebenen Übungsblätter konnten so analysiert werden. Während hier die Weihnachtsferien den offensichtlichsten Effekt auf die Downloadhäufigkeit haben, finden sich bei der Analyse deutlich Ausreißer vor dem 1. und 2. Online-Test – Details in den Abbildungen 3a und 3c.

3.1.4 Video-Aufzeichnungen

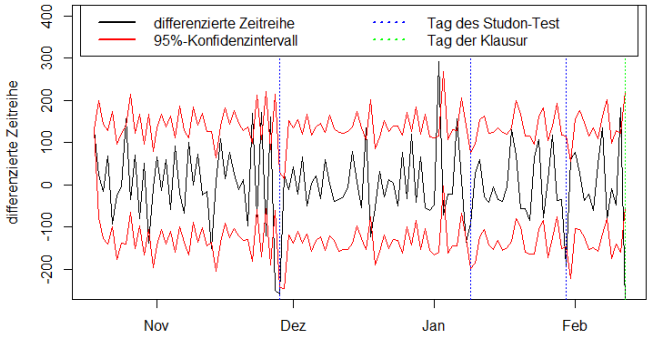
Als besonderes Entgegenkommen an die Studierenden, die in Massenveranstaltungen durch die entstehende Unruhe der Veranstaltung nur bedingt folgen können, werden die entsprechenden Veranstaltungen mitgefilmt. Diese Video-Aufzeichnungen werden auch zum Herunterladen bereitgestellt. Auch in dieser Zeitreihe finden sich Ausreißer vor den Tests 1 und 2 deutlich erkennbar – Details in den Abbildungen 3b und 3d.

3.1.5 Ergebnis

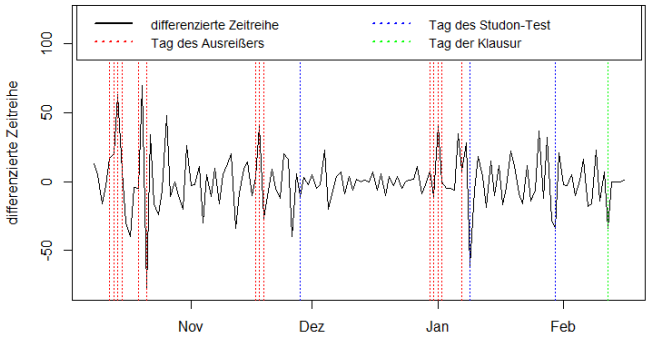
Die betrachteten Zeitreihen sprechen dafür, dass die Online-Tests die Studierenden dazu gebracht haben sich mit den bereitgestellten Unterlagen auseinanderzusetzen. Dabei kann auch eine Bewertung zu der Wahl der Zeitpunkte abgegeben werden: der erste und der zweite Test scheinen Einfluss auf die Aktivität der Studierenden gehabt zu haben. Der dritte Online-Test war zu nahe an der Klausur, um einen Einfluss neben den ohnehin laufenden Klausurvorbereitungen erkennen zu können. Da solche Daten erstmalig erhoben wurden und daher ein Vergleich mit älteren Semestern nicht möglich ist, muss bei dieser Interpretation bedacht werden, dass der erste und der zweite Online-Test in die Anfangsphase des Semesters und in die Phase nach den Weihnachtsferien fällt. Diese Effekte konnten nicht herausgerechnet werden.



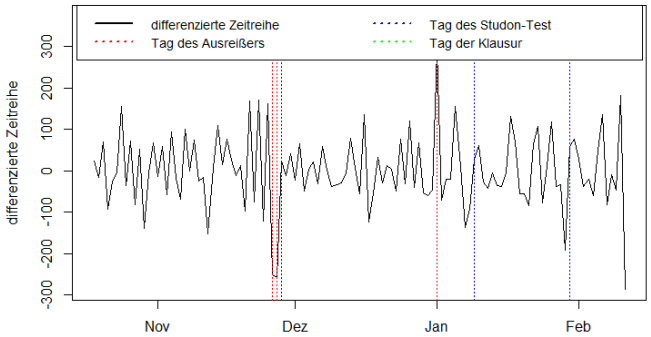
(a) GARCH(1,1)-Modell für Downloadzahlen der Übungsblätter.



(b) ARMA(1,1)-Modell für Zugriffszahlen auf die Video-Vorlesung.



(c) GARCH(1,1)-Modell für Downloadzahlen der Übungsblätter.



(d) ARMA(1,1)-Modell für Zugriffszahlen auf die Video-Vorlesung.

Abbildung 3: Downloadzahlen ausgewählter Materialien im Semesterverlauf. Hier wurde getestet, ob es zu Anomalien (Ausreißern) im Datensatz während des Zeitraums der Online-Tests kommt. Diese wurden in 3c und 3d auf Grundlage eines Likelihood-Verhältnistests und in 3a und 3b auf Grundlage von 95% Konfidenzintervallen identifiziert.

3.2 Beaufsichtigung der Teilnahme

Ziel der folgenden Analyse ist die Untersuchung, inwiefern sich die Verteilungen der Testergebnisse unterscheidet, wenn die unbeaufsichtigten (zu Hause) und die beaufsichtigten (im CIP-Pool) Ergebnisse getrennt betrachtet werden. Den Studierenden, die zufällig für die beaufsichtigte Bearbeitung ausgewählt wurden, war es beispielsweise nicht möglich, die Tests im Team zu lösen oder unerlaubte Hilfsmittel in Anspruch zu nehmen. Unter der Annahme, dass auch die Studierenden, die frei teilnehmen konnten, derartige Hilfestellungen nicht nutzten, sollten zu jedem Testdurchlauf keine Verteilungsunterschiede zwischen den freien und den beaufsichtigten Testergebnissen erkennbar sein.

Wenn die Manipulationen bei der Bearbeitung übergreifende Größenordnung annehmen, sollte sich hier ein Unterschied bemerkbar machen.

3.2.1 Deskriptive Auswertung

Die Daten unterteilen sich in insgesamt sechs Stichproben, da im Laufe des Semesters drei Online-Tests durchgeführt wurden und zu jedem Test die Daten der frei und der unter Aufsicht arbeitenden Gruppen vorliegen. **Test 1-3** beschreiben im weiteren Verlauf die Ergebnisse der drei Tests, die zu Hause durchgeführt wurden und **CIP 1-3** die dazugehörigen Ergebnisse der unter Aufsicht gelösten Tests. Somit liegen drei Paare (**Test 1**/**CIP 1**, **Test 2**/**CIP 2** und **Test 3**/**CIP 3**) vor.

	n	Mittelwert	Median	Unteres Quartil	Oberes Quartil
Test 1	985	0.5019	0.5167	0.3583	0.6667
Test 2	774	0.4442	0.4364	0.2727	0.6090
Test 3	686	0.4729	0.48	0.29	0.67
CIP 1	66	0.4140	0.4667	0.2083	0.65
CIP 2	15	0.5188	0.5727	0.3818	0.6545
CIP 3	45	0.4542	0.43	0.28	0.64

Tabelle 3: Deskriptive Maßzahlen der Ergebnisse (Erreichte Punkte in %) in den Online-Aufgabenserien 1-3 frei (**Test**) und unter Aufsicht (**CIP**).

In der ersten Online-Aufgabenserie liegt der Mittelwert der Ergebnisse unbeaufsichtigter Bearbeitung **Test 1** deutlich über denen der beaufsichtigten in **CIP 1**. Bei der dritten Online-Aufgabenserie ist der Unterschied geringer und bei der zweiten ist das Größenverhältnis genau entgegengesetzt. Zu beachten ist dabei die geringe Fallzahl in **CIP 2**, die durch geringe Raumverfügbarkeit und nur geringe Auswahlhäufigkeit unter den bearbeitenden Studierenden zustande kam.

3.2.2 Inferenz

Normalverteiltheit der Daten ist bei mehreren Standardanalysen eine fundamentale Annahme, kann aber nicht durch entsprechende Tests bestätigt werden (p-Werte für die Normalverteilungstests von Jarque-Bera bzw. Shapiro-Wilk sind jeweils kleiner als `5e-4`). Hier werden deshalb die Ergebnisse nichtparametrischer Methoden betrachtet.

Die Nullhypothese, alle 6 Stichproben kämen aus derselben Verteilung, wird mit dem Kruskal-Wallis Test überprüft und abgelehnt (p-Wert < 0.01). Da die Unterschiede durch den sich ändernden Schwierigkeitsgrad der Online-Tests erklärt werden könnten, sind die Paarvergleiche **Test/CIP** innerhalb der Online-Aufgabenserien besonders interessant. Bei Anwendung des Wilcoxon Rangsummentests ergeben sich folgende p-Werte

Online-Aufgabenserie 1: 0.032
Online-Aufgabenserie 2: 0.152
Online-Aufgabenserie 3: 0.61

Während in den Online-Aufgabenserien 2 und 3 der Unterschied in der Lage der Verteilung durch den Zufall erklärt werden kann, ist der Lage-Unterschied in der ersten Online-Aufgabenserie zumindest am 5%-Niveau signifikant von 0 verschieden.

3.2.3 Ergebnis

Es konnten keine deutlichen Signale gefunden werden, dass sich die Verteilung durch die kontrollierten Bedingungen änderte[6]. Das muss zwar nicht unbedingt bedeuten, dass generell keine Manipulationen vorliegen – sie scheint bei dieser Analyse jedoch nicht ins Gewicht zu fallen. Das Ausmaß scheint also

[6]Im Übrigen liefern ANOVA-Analyse und t-Tests vergleichbare Ergebnisse.

nicht so groß zu sein, wie befürchtet. Problematisch ist bei dieser Analyse, dass bestimmte Verzerrungen nicht ausgeschlossen werden können:

- Wer manipulieren wollte, kam womöglich schlicht nicht zu dem kontrollierten Test (Geringe Fallzahl macht Untersuchung unmöglich)

- Die Organisation der kontrollierten Tests machte es notwendig, deren Termin zeitlich **nach** dem des regulären Online-Tests anzusetzen. Wer an den kontrollierten Tests teilnahm, konnte deshalb die akkumulierten Informationen, die andere Teilnehmer bereitgestellt hatten[7], zur Vorbereitung nutzen.

Da die Fallzahlen bei den kontrollierten Teilnahmen gering sind, darf diesem Ergebnis nicht zu viel Bedeutung beigemessen werden. Dennoch kann festgehalten werden, dass die Unterschiede deutlich geringer ausgefallen sind als im Vorfeld erwartet wurde.

3.3 Kontinuierliche Nacharbeit – Vergleich mit Vorjahr

Im WS 12/13 stand eine konfirmatorische Analyse der Beziehungen zwischen den latenten Variablen Voraussetzungen, Nacharbeit[8] und Lernerfolg im Mittelpunkt der Analysen. Dabei waren die erwarteten linearen Wirkungen bestätigt worden. Die Folgerung daraus war, dass die intensivierte Nacharbeit, gemessen durch die Ergebnisse der Online-Tests, in eine bessere Prüfungsleistung übertragen werden konnte. Diese Ergebnisse führten schließlich dazu, dass in der zu Beginn erwähnten qualitativen Analyse die hohe Beliebtheit des Projekts stärker wog, als die befürchtete Möglichkeit zu manipulieren.
In der zweiten Durchführungsperiode überwogen die negativen Argumente von vornherein. Daher war bereits vor der Durchführung der Analyse mit den Daten aus dem WS 13/14 klar, dass deren Ergebnisse keinen Einfluss auf die letztendliche Abwägung der Argumente über den Fortbestand des Projektes haben würden. Allerdings wollte man untersuchen, inwiefern die Analyse der Daten aus dem WS 13/14 die bereits erzielten Ergebnisse reproduzieren kann bzw. mögliche Unterschiede zwischen den beiden Perioden gefunden werden können.

Die Variation der Bedingungen in WS 13/14 wurde in einem erweiterten Modell ebenfalls berücksichtigt (Abschnitt 3.4).

[7] z.B. durch persönlichen Austausch, Organisation in Foren, Gruppen sozialer Netzwerke
[8] irrtümlich als Mitarbeit bezeichnet, gemeint war auch dort: Nacharbeit

3.3.1 Datengrundlage

Um die Ergebnisse der quantitativen Analysen vergleichbar zu halten, wird die Datengrundlage nach demselben Muster erzeugt wie in Pleier und Mangold (2013). Von den Teilnehmern der 10 ECTS Klausur werden diejenigen ausgeschlossen, die in der Klausur 5 Punkte oder weniger erzielten oder in den Online-Tests die 20%-Grenze unterschritten haben.

Im WS 13/14 ist die Fallzahl deutlich geringer, da generell weniger Studierende die Online-Tests nutzten und zusätzlich diejenigen aus der vergleichenden Analyse ausgeschlossen wurden, die zur Durchführung in die CIP-Pools der Universität geladen wurden.

Mit 171 verbleibenden Beobachtungen ist die Mindestzahl von 100 zumindest nicht unterschritten. Das standardisierte Prüfungsprotokoll liefert dieselbe Datenstruktur, sodass die Datengrundlage zwar weniger Beobachtungen, aber die gleichen Variablen zur Verfügung stellt.

3.3.2 Modell

Abbildung 4 zeigt das Pfaddiagramm des Messmodells, das gegenüber dem in Pleier und Mangold (2013) verwendeten keine Unterschiede aufweist. Es müssen lediglich einige Korrelationen zwischen den Items angepasst werden. Die Pfade im Strukturmodell standen im Mittelpunkt der Analyse vom WS 12/13 und die Überlegungen behalten auch für das WS 13/14 ihre Gültigkeit:

- γ_2 soll einen positiven Wert annehmen,
- β_1 soll mindestens so groß sein wie γ_2,
- für γ_1 wird ein positiver Wert gefordert,
- γ_3 wird der Vollständigkeit halber mitgeschätzt.

Zum Testen der Annahme, dass die Online-Tests in positiver Weise die Nacharbeit beeinflussten und damit die Befürchtung weitgehender Manipulationen abschwächten, wurde eine konkrete Hypothese formuliert:

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

Diese steht im vergleichbaren Analyseblock zum WS 13/14 wieder im Mittelpunkt.

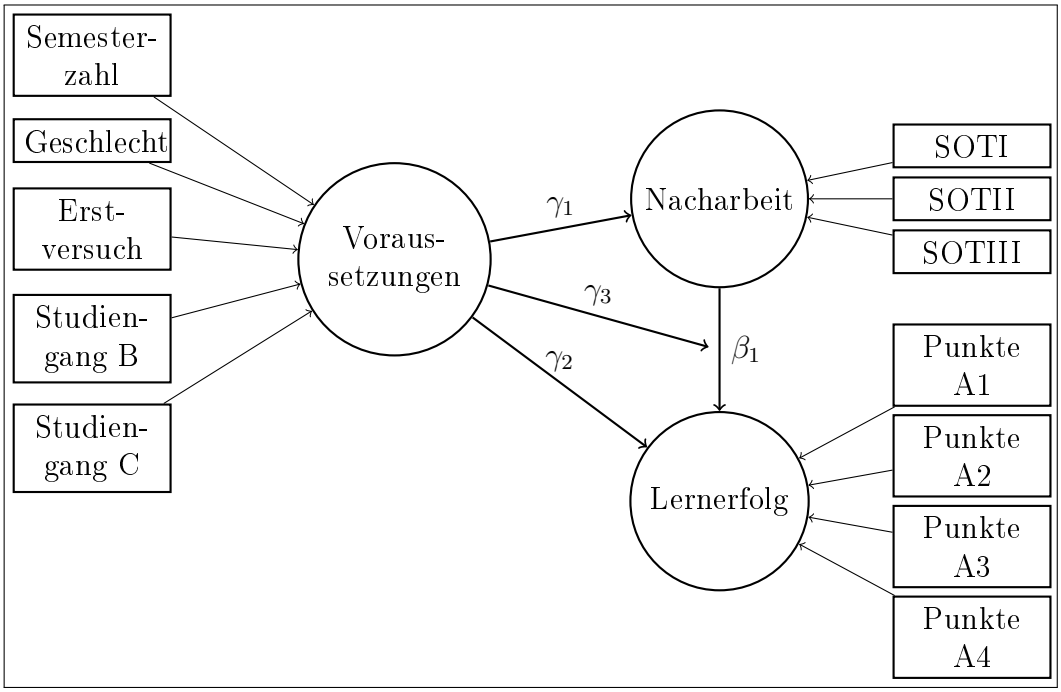


Abbildung 4: Pfaddiagramm des unterstellten Gesamtmodells ohne Korrelationen – können die Ergebnisse vom WS 12/13 aus Pleier und Mangold (2013) im WS 13/14 reproduziert werden?

3.3.3 Auswertung

Um der verletzten Annahme der Normalverteilung bei gleichzeitig geringer Fallzahl gerecht zu werden, wurde im WS 12/13 der Standardfehler mittels Bootstrap bestimmt. Dabei ergab sich eine akzeptable Modellanpassung, eher mäßige Ladungen der Items, Punkt- und Intervallschätzer der interessierenden Größen lagen im erwarteten Bereich und die Nullhypothese (1) wurde nicht abgelehnt.

Die analogen Analyseergebnisse zum WS 13/14 werden mit den gleichen Algorithmen erzeugt.

Modellanpassung

Die Anpassungsmaße in Tabelle 4 nach Schermelleh-Engel et al. (2003) deuten eine gute Anpassung an, es werden keine negativen Varianzen ausgegeben und die Kovarianzen zeigen in die erwarteten Richtungen. Die Punkte in

Anpassungsmaß		Anpassungsgüte
RMSEA	0,00	gut
p-Wert (RMSEA $<$ 0,05)	1,00	gut
90%KI	0,00 0,03	gut
SRMR	0,05	gut
CFI	1,00	gut

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

	Schätzer	Std.-Fehler (robust)	Std.-Fehler (BS)
Nacharbeit $\sim$			
Voraussetzungen	$\widehat{\gamma}_1 =$ 0,235	0,114	0,125
Lernerfolg $\sim$			
Voraussetzungen	$\widehat{\gamma}_2 =$ 0,366	0,109	0,118
Nacharbeit	$\widehat{\beta}_1 =$ 0,280	0,138	0,148
Voraussetzungen:Nacharbeit	$\widehat{\gamma}_3 = -$0,209	0,098	0,107

Tabelle 5: Ergebnisse der Parameterschätzung und Standardfehler für das Strukturmodell (s. Abbildung 4).

den Klausuraufgaben laden akzeptabel, die der Online-Tests etwas schlechter (angeführt vom dritten Test) und die Items zur Messung der Voraussetzungen laden durchmischt (Semesterzahl: ~ 0.97, Geschlecht: ~ 0.20). Damit ist die Anpassung insgesamt etwas besser als im WS 12/13.

Parameterschätzung

Die Schätzung per likelihoodbasierter Diskrepanzfunktion auf Grundlage der Korrelationsmatrix und normierten latenten Variablen liefert Schätzwerte die etwa in der Größenordnung liegen, wie die aus dem Vorjahr. Auch stimmen die Vorzeichen überein. Es sieht jedoch so aus, als haben sich die

	Typ B -Test ($H_0^{>}$)	Typ A - Test ($\overline{H}_0^{=}$)
LR statistic	0,04	4,89
Adjusted p-value	0,46	0,09

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

Wirkungen abgeschwächt, da die Werte teilweise kleiner sind bei größeren Standardfehlern.

Die in R[9] erfolgte Schätzung (mittels der Pakete `lavaan` und `MVN`[10]) bestärken die Ergebnisse aus dem Vorjahr (vgl. Tabelle 5). Auch diesmal sieht es so aus, als könnten die schwächeren Studierenden von einer vertieften Nacharbeit überproportional profitieren. Es fällt jedoch auf, dass der Koeffizient $\widehat{\beta}_1$ deutlich kleiner ist als im Vorjahr, was u.a. durch eine Zunahme der Manipulationen erklärt werden könnte.

Nullhypothese mit Ungleichungsbedingungen

Das Kriterium, das letztlich den entscheidenden Ausschlag gab, die Fortführung des Projekts zu empfehlen, war der Bootstrap-Test der Nullhypothese (1) nach Van de Schoot et al. (2010) mit einer Methode aus Bollen und Stine (1992) aufgrund der fehlenden Normalverteiltheit. Dieser Test wird auf dem neuen Datensatz ebenfalls durchgeführt.

Es stellt sich heraus, dass basierend auf den vorliegenden Daten die Hypothese (1) nicht abgelehnt werden kann (Typ B - Test). Werden die Ungleichheitsbedingungen aus (1) als Alternative den entsprechenden Gleichheitsbedingungen als Nullhypothese gegenübergestellt (Typ A - Test) kann die Nullhypothese am 5%-Signifikanzniveau nicht mehr abgelehnt werden. Dies ist ein Unterschied zu den Ergebnissen in WS 12/13, der für einen verminderten Lerneffekt spricht (vgl. Tabelle 6).

[9] R Core Team (2014)
[10] Rosseel (2012)

3.3.4 Ergebnis

Bis auf einige geringer ausfallende Ladungen der Items liegt eine gute Modellanpassung vor, sodass die Analyse verwertbare Ergebnisse bereitstellt.

Insgesamt ist entscheidend, dass die Nullhypothese (1) nicht abgelehnt werden kann. Damit gilt die Befürchtung starker Manipulationen als abgemildert.

Da die grundsätzliche Interpretation in beiden Datensätzen aus WS 12/13 und WS 13/14 vergleichbar ist, ist die Verlässlichkeit der Analyse aus Pleier und Mangold (2013) gestärkt.

3.4 Beaufsichtigung der Teilnahme

Forderung bei der empfohlenen Projektfortsetzung war es, zusätzliche Kontrollmechanismen einzusetzen. Der Rahmen des Projekts erlaubte dabei nur, eine Zufallsauswahl von Teilnehmern in die CIP-Pools zu setzen und nicht etwa alle. Die Zufallsauswahl sollte glaubwürdig sein, weshalb sie nicht von den Mitarbeitern, sondern direkt von der Lernplattform durchgeführt wurde.

Besonderes Augenmerk wird darauf gelegt, wie sich die Ergebnisse des angepassten Modells von denen aus dem Abschnitt 3.3 unterscheiden.

3.4.1 Datengrundlage

Es handelt sich um die Datengrundlage der vorherigen Analyse aus Abschnitt 3.3, ergänzt um die Teilnehmer die die Kriterien erfüllen und mindestens einmal im CIP-Pool teilnahmen - mit mindestens 20% der erreichbaren Punkte. Insgesamt verbleiben zur Schätzung 188 Beobachtungen.

3.4.2 Modell

Die Stochastik bei dem beschriebenen Vorgehen brachte es mit sich, dass nicht alle verfügbaren Plätze in den CIP-Pools belegt wurden (die Anzahl der an der Online-Aufgabenserie teilnehmenden Studierenden war unbekannt, die Anzahl der Plätze im CIP-Pool beschränkt). Darum stehen nur wenige Beobachtungen zu den erreichten Punktzahlen in den CIP-Pools zur Verfügung. Für ein eigenständiges Modell, in dem zwei unterscheidbare Formen der Nacharbeit, etwa unbeaufsichtigte zu Hause und beaufsichtigte Nacharbeit in den CIP-Pools, untersucht werden können, reichten die Fallzahlen nicht.

Stattdessen laden alle 6 Items (3 Online-Tests × 2 Teilnahmemodi) auf eine latente Variable Nacharbeit.

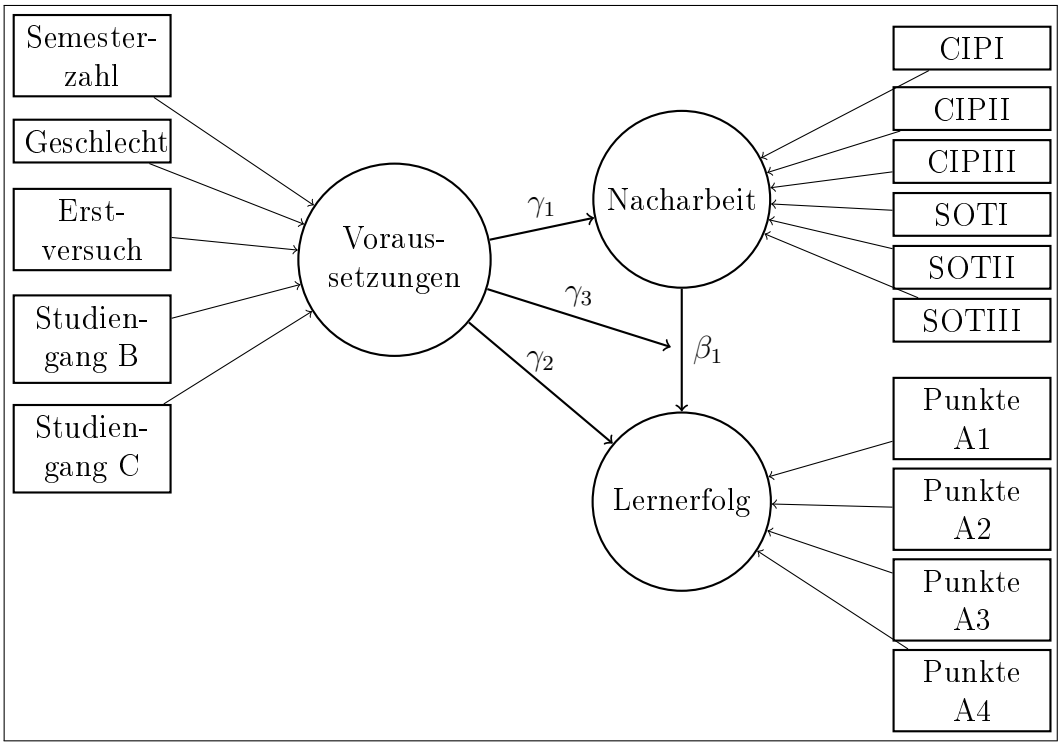


Abbildung 5: Pfaddiagramm des unterstellten Gesamtmodells ohne Korrelationen – gleiches Modell wie in Pleier und Mangold (2013), jedoch zusätzliche Items (erreichte Punktzahl im CIP-Pool).

3.4.3 Auswertung

Auch für die Auswertung des erweiterten Modells wurden die gleichen Algorithmen wie in Abschnitt 3.3 verwendet. Lediglich das Modell wurde angepasst (s. Abbildung 5 und die Korrelationen zwischen den Items).

Anpassungsmaß		Anpassungsgüte
RMSEA	0,00	gut
p-Wert (RMSEA $<$ 0,05)	1,00	gut
90%KI	0,00 0,03	gut
SRMR	0,05	gut
CFI	1,00	gut

Tabelle 7: Maße der Anpassungsgüte: Root Mean Square Error of Approximation, Standardized Root Mean Square Residual und Comparative Fit Index.

Modellanpassung

Die Anpassungsmaße (vgl. Tabelle 7) unterscheiden sich erst in der dritten Nachkommastelle von denen in der vorherigen Analyse, die Varianzen sind nicht negativ und die Korrelationen verhalten sich erwartungsgemäß. Die Ladungen sind ähnlich wie im vorherigen Modell und die neuen Items laden etwas schwächer als die Ergebnisse in der freien Bearbeitung. Damit ist die Anpassung insgesamt vergleichbar mit der in der vorherigen Analyse.

Parameterschätzung

Die Schätzwerte (vgl. Tabelle 8) liegen nun wieder näher an denen der Analyse des Vorjahres. Die Vorzeichen stimmen überein. Die Standardfehler sind wieder etwas kleiner (dies ist wohl zum Teil auf die größere Datengrundlage zurückzuführen). Es deutet sich an, dass in dieser Analyse die Manipulationen wieder weniger Einfluss ausüben.

Nullhypothese mit Ungleichungsbedingungen

Auch der Test zu Hypothese (1) wurde durchgeführt (vgl. Tabelle 9). Die Daten sprechen dafür, dass man sich wieder tiefer in Nicht-Ablehnungsbereich der Nullhypothese befindet. Auch das ist ein Hinweis, dass in diesen Daten Manipulationen einen geringeren Einfluss ausüben.

	Schätzer	Std.-Fehler (robust)	Std.-Fehler (BS)
Nacharbeit $\sim$			
Voraussetzungen	$\widehat{\gamma}_1 =$ 0,355	0,112	0,123
Lernerfolg $\sim$			
Voraussetzungen	$\widehat{\gamma}_2 =$ 0,304	0,100	0,110
Nacharbeit	$\widehat{\beta}_1 =$ 0,335	0,125	0,137
Voraussetzungen:Nacharbeit	$\widehat{\gamma}_3 = -0{,}165$	0,080	0,087

Tabelle 8: Ergebnisse der Parameterschätzung und Standardfehler für das Strukturmodell (s. Abbildung 5).

	Typ B -Test ($H_0^{>}$)	Typ A - Test ($\overline{H}_0^{=}$)
LR statistic	0,15	11,73
Adjusted p-value	0,76	0,01

Tabelle 9: Wert der Teststatistik und p-Wert zur Nullhypothese (1) unter Beaufsichtung der Teilnahme.

3.4.4 Ergebnis

Bis auf einige gering ausfallende Ladungen der Items liegt eine gute Modellanpassung vor, sodass die Analyse verwertbare Ergebnisse bereitstellt.

Die Nullhypothese (1) wurde nicht abgelehnt. Entscheidender bei dieser Analyse sind die Unterschiede zwischen den Modellen aus Abschnitt 3.3 und 3.4. Dabei scheint sich herauszustellen, dass in den Daten, bei denen zumindest teilweise die Teilnahme beaufsichtigt wurde, der Einfluss von Manipulationen geringer ist. Damit wird die Forderung nach Ausbau von Kontrollmechanismen untermauert – gleichzeitig zeigen sich dabei jedoch auch die Grenzen des Anreizsystems auf.
Das Projekt kommt seinen Zielen insgesamt ziemlich nahe. Nur lassen sich über 1.000 Studierende in so einer Maßnahme nicht komplett beaufsichtigen.

4 Zusammenfassung

Eine Weiterführung des Projekts ist nicht sinnvoll. Diese Entscheidung basiert im Wesentlichen auf den Erfahrungen, die aus dem Feedback der teilnehmenden Studierenden gewonnen wurden.

Erwähnenswert sind die erneuten positiven Befunde der quantitativen Analysen. Es gibt Hinweise darauf, dass die Online-Tests die Studierenden dazu gebracht haben, sich während des Semesters verstärkt mit dem Inhalt der Veranstaltung zu beschäftigen. Es finden sich unterstützende Argumente, dass auch während dieser zweiten Durchführung Studierende durch die Teilnahme an den Online-Tests eine intensivierte Nacharbeit in gute Prüfungsergebnisse umsetzen konnten. Die Verteilung der erreichten Punkte in den Online-Tests hängt weniger von der Kontrolle ab als erwartet. Insgesamt scheinen diese Ergebnisse zu suggerieren, dass die beobachteten und berichteten Manipulationen ein eher geringeres Ausmaß annehmen als befürchtet und die Mehrheit an Studierenden die Online-Tests ehrlich bearbeitet haben. Bei dem durchgeführten Projekt hatte also das *Cheating* wenig Einfluss auf die Ergebnisse der durchgeführten Analyse. Allerdings schaden solche Einzelfälle dem Projekt. Denn einerseits kann die notwendige Kontrolle nicht gewährleistet werden - andererseits führt ein gutes Ergebnis in den Online-Tests auf der Grundlage von Manipulation zu Unzufriedenheit unter den ehrlich arbeitenden Studierenden. Dies schadet der Akzeptanz des Projekts.

Sollte man sich zu einer Neuauflage des Projektes entschließen, sollten

auch die Erkenntnisse von Fryer (2011) berücksichtigt werden: Die Wirkung eines Anreizes wird verstärkt, wenn die Belohnung unmittelbar erfolgt. Dies könnte durch direktes Feedback nach Beantwortung der Frage in Kombination mit einem Punktestand der aktuell verdienten Bonuspunkte für die Klausur verwirklicht werden.

Insgesamt folgt der Schluss, dass die Lernzielerreichung der Veranstaltung durch die Online-Tests positiv unterstützt wurde. Die Nacharbeit der Studierenden wurde durch das Projekt gefördert und damit hoffentlich auch der langfristige Lernerfolg.

5 Danksagung

Die Autoren möchten sich bei der Hans-Frisch-Stiftung für die Ermöglichung dieser Forschungsarbeit bedanken.

Literatur

Angrist, J., Lang, D. W., Oreopoulos, P. (2007). Incentives and Services for College Achievement: Evidence from a Randomized Trial. IZA Discussion Papers 3134, Institute for the Study of Labor (IZA).

Angus, S. D., Watson, J. (2009). Does regular Online Testing enhance Student Learning in the Numerical Sciences? Robust Evidence from a large Data Set. *British Journal of Educational Technology*, 40(2):255–272.

Ardelean, V. (2014). Identifikation von Ausreißern in (V)ARMA- und (M)GARCH-Prozessen. Dissertation, *Lehrstuhl für Statistik und Ökonometrie, FAU Erlangen-Nürnberg*.

Artelt, C. (2000). *Strategisches Lernen*. Waxmann Münster.

Bollen, K. A., Stine, R. A. (1992). Bootstrapping Goodness-of-Fit Measures in Structural Equation Models. *Sociological Methods and Research*, 21:205–229.

Bourne, L. E., Ekstrand, B. R. (2005). *Einführung in die Psychologie*. Klotz Verlag GmbH, 4 edition.

Chang, I., Tiao, G. C., Chen, C. (1988). Estimation of Time Series Parameters in the Presence of Outliers. *Technometrics*, 30(2):pp. 193–204.

Doornik, J. A., Ooms, M. (2005). Outlier Detection in GARCH Models. Economics Working Papers W24, Nuffield College.

Fryer, R. (2011). Financial Incentives and Student Achievement: Evidence from Randomized Trials. *Quaterly Journal of Economics*, 126(4):1755–1798.

Ghalanosl, A. (2014). *rugarch: Univariate GARCH Models.*

Kibble, J. (2007). Use of Unsupervised Online Quizzes as Formative Assessment in a Medical Physiology Course: Effects of Incentives on Student Participation and Performance. *Advances in Physiology Education*, 31(3):253–260.

Kremer, M., Miguel, E., Thornton, R. (2009). Incentives to Learn. *The Review of Economics and Statistics*, 91(3):437–456.

Luehrmann, M., Chevalier, A., Dolton, P. (2013). Making it count: Evidence from a Field Experiment on Assessment Rules, Study Incentives and Student Performance. Number C07-V3 in Beiträge zur Jahrestagung des Vereins für Socialpolitik 2013: Wettbewerbspolitik und Regulierung in einer globalen Wirtschaftsordnung - Session: Effort Compensation. ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft.

Mietzel, G. (1993). *Psychologie in Unterricht und Erziehung: Einführung in die pädaogische Psychologie für Pädagogen und Psychologen.* Verl. für Psychologie Hogrefe, Göttingen, 4 edition.

Patel, R., Richburg-Hayes, L. (2012). Performance-Based Scholarships: Emerging Findings from a National Demonstration. Policy Brief. *MDRC.*

Pleier, T., Mangold, B. (2013). Lehrverbesserung durch Online-Tests: Effekte der Eigenarbeit von Studierenden. *Diskussionspapiere, Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Statistik und Ökonometrie*, 90.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2):1–36.

Rowe, N. C. (2004). Cheating in Online Student Assessment: Beyond Plagiarism. *Online Journal of Distance Learning Administration*, 7(2).

Schermelleh-Engel, K., Moosbrugger, H., 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.

Van de Schoot, R., Hoijtink, H., Dekovic, M. (2010). Testing Inequality Constrained Hypotheses in SEM Models. *Structural Equation Modeling*, 17:443–463.

Woit, D., Mason, D. (2003). Effectiveness of Online Assessment. In *ACM SIGCSE Bulletin*, volume 35, pages 137–141. ACM.