



## IPSDS WORKING PAPER SERIES #2

# **Datenqualitätsprobleme bei der umfragebasierten Messung von studentischem Workload: Eine Mixed-Methods-Studie auf Grundlage von Learning Analytics und kognitiven Interviews**

Evgenia Samoilova, Tobias Wolbring, Florian Keusch

8 April, 2019

University of Mannheim

SPONSORED BY THE



Federal Ministry  
of Education  
and Research

The project on which this report is based was funded by the Federal Ministry of Education and Research under the number [16OH22064]. Responsibility for the content of this publication lies with the author.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

## **Abstract**

Diese Studie untersucht die Qualität studentischer Selbstangaben zum zeitlichen Aufwand anhand von Learning Analytics und kognitiven Interviews im Rahmen eines berufsbegleitenden online-basierten Fernstudiengangs. Die Selbsteinschätzung des Zeitaufwands wurde mittels einer wöchentlichen webbasierten Umfrage über ein ganzes Semester hinweg erhoben. Die Online-Lernumgebung des Studienprogramms erlaubte eine nichtreaktive Messung des Workloads mittels Videoansicht-Logs in Echtzeit und deren Verknüpfung mit den Umfragedaten. Die Ergebnisse zeigen, dass sich die Schätzungen auf Grundlage von Logs und Umfragedaten deutlich unterscheiden. Im Durchschnitt liegen die Selbstangaben deutlich über den Log-basierten Werten. Die Korrelation zwischen beiden Messungen ist sehr schwach und bivariate Regressionsmodelle deuten darauf hin, dass die Umfrage-basierten Selbstangaben ein statistisch signifikanter, jedoch schwacher Prädiktor des Workloads sind. Die kognitiven Interviews zeigen ergänzend dazu, dass Erinnerungs- und Urteilsfehler die zentrale Ursache für Verzerrungen in den studentischen Angaben darstellen.

## **Einleitung**

Der zeitliche Aufwand, den Studierende für ihr Studium erbringen, steht im engen Zusammenhang mit Belastungsfaktoren wie Stress und wirkt sich auf den erzielten Lernerfolg aus (Wang et al. 2015). Daher sollte die geplante und tatsächliche Zeitinvestition der Studierenden bei der Gestaltung von Lehrveranstaltungen und der Konzeption von Studiengängen unbedingt berücksichtigt werden. In Anbetracht der zunehmenden Verbreitung von E-Learning-Angeboten erscheint die Frage nach der zeitlichen Belastung besonders wichtig, da solche Formate intensive Selbstlernphasen implizieren und ein hohes Maß an Selbstständigkeit und Gewissenhaftigkeit der Studierenden erfordern. Es bleibt daher gerade bei E-Learning-Angeboten unklar, wie viel Zeit tatsächlich in die Vor- und Nachbereitung der Lehre investiert wird. Obwohl Workload wichtige Auswirkungen auf die Lehrpraxis und deren Bewertung hat, bleibt die Frage nach der adäquaten Messung des studentischen Workloads in der Literatur weitgehend unbeantwortet. Insbesondere die Selbsteinschätzung in herkömmlichen Evaluationsfragebögen zum erbrachten Zeitaufwand könnte unter verschiedenen Messproblemen leiden, die bei der umfragebasierten Datenerhebung häufig auftreten, wie z.B. Erinnerungsfehler, Verzerrungen aufgrund von Non-Response und Effekte sozialer Erwünschtheit (siehe Groves et al. 2009 sowie Berger und Baumeister in diesem Band).

Vor diesem Hintergrund untersuchen wir auf der Basis von zwei Kursen eines berufsbegleitenden online-basierten Fernstudiengangs die Qualität studentischer Selbstangaben zum zeitlichen Aufwand (reduziert auf das Ansehen von aufgezeichneten Videovorlesungen) anhand von Learning Analytics (LA) und kognitiven Interviews. Die Online-Lernumgebung des Studienprogramms erlaubte eine nichtreaktive Messung des Workloads mittels Videoansicht-Logs in Echtzeit. Die Selbsteinschätzung des Zeitaufwands wurde in einer wöchentlichen webbasierten Umfrage unter

16 (Kurs 1) und 13 Studierenden (Kurs 2) über ein ganzes Semester hinweg erhoben und konnte (nach Einverständniserklärung der Studierenden) mit den Log-Dateien im Längsschnitt verknüpft werden (n=192 und n=143). Neben dem Vergleich von Befragungsdaten und Learning Analytics präsentieren wir Ergebnisse kognitiver Interviews mit Studierenden (n=13). Kognitive Interviews stellen einen qualitativen Ansatz dar, der sich auf die kognitiven Prozesse bei dem Verständnis und der Beantwortung von Surveyfragen konzentriert (Willis 2005). Da eine Störung im kognitiven Beantwortungsprozess zu Messfehlern führen kann, werden kognitive Interviews sowohl in der Pretestingphase als auch nach der finalen Datenerhebung verwendet, um zu beurteilen, ob die Frage tatsächlich misst, was ursprünglich beabsichtigt wurde. Darüber hinaus präsentieren wir Ergebnisse zu einem dritten Kurs, für welchen wir aufgrund unserer empirischen Befunde zu Kurs 1 und 2 die Messung des Workloads im Fragebogen für Kurs 3 von Stunden auf Minuten geändert haben.

Die Ergebnisse zeigen, dass die Logs zur Videonutzung eine deutlich feinkörnigere Messung des Workloads erlauben, während sich die studentischen Selbstangaben in der Befragung auf fokale Punkte (z.B. 30 Minuten) konzentrieren. Zudem ergeben bivariate Zusammenhangsanalysen, dass sich die Schätzungen auf Grundlage von Logs und Umfragedaten deutlich unterscheiden und nur schwach miteinander korrelieren. Auch bivariate Regressionsmodelle zeigen, dass die Umfragebasierten Selbstangaben ein schwacher Prädiktor für die Zeit des Ansehens von aufgezeichneten Videovorlesungen sind. Daher sollten die mittels Befragung ermittelten Durchschnittswerte der Arbeitsbelastung mit Vorsicht betrachtet werden (z.B. bei der Umrechnung in ECTS). Die kognitiven Interviews zeigen ergänzend dazu, dass Erinnerungs- und Urteilsfehler die zentrale Ursache für Verzerrungen in den studentischen Angaben darstellen. Zudem konnten mittels kognitiver Interviews zwei verschiedene Schätzstrategien identifiziert werden, welche Studierende angesichts der Erinnerungsprobleme bei der Antwortfindung anwenden.

## **1. Wie kann studentischer Workload gemessen werden?**

### *Traditionelle Methoden der Workload-Messung: Befragung und Tagebuch*

Eine verbreitete Methode zur Workload-Messung besteht darin, die Studierenden zu befragen, wie viel Zeit sie für bestimmte Lernaktivitäten aufgewendet haben. Die Frage zur Selbsteinschätzung des Workloads kann offen gestellt werden (d.h. Angabe des Zeitumfangs als Zahl) oder eine Reihe von geschlossenen Antwortoptionen beinhalten (z.B. Kategorien „bis zu 1 Stunde“, „1-2 Stunden“ etc.). Trotz der breiten Anwendung solcher Messungen ist bekannt, dass solche Selbsteinschätzungen angesichts von Erinnerungs- und Urteilsfehlern problematisch sein können, wenn es darum geht, den tatsächlichen Zeitaufwand für eine Aktivität zu messen (Tourangeau et al. 2000). Erinnerungsfehler entstehen, wenn Menschen Schwierigkeiten haben, sich daran zu erinnern, was sie im jeweiligen Zeitraum getan haben, insbesondere wenn die betreffende Tätigkeit nicht täglich durchgeführt wird. Urteilsfehler beziehen sich auf Fehler, die bei der Summierung des

zeitlichen Aufwands für einzelne Aktivitäten (z.B. Lesen, Skripte wiederholen, Aufgaben bearbeiten), die einer gemeinsamen Kategorie (z.B. Vor- und Nachbereitung) angehören, auftreten.

Weitere Nachteile der Verwendung von Selbsteinschätzungen zur Zeitmessung liegen in möglichen Verzerrungen aufgrund von sozialer Erwünschtheit, *Satisficing* und Non-Response (Massey und Tourangeau 2013; Tourangeau und Yan 2007; Krosnick 1991). Verzerrungen durch soziale Erwünschtheit ergeben sich in Situationen, in denen die Befragten ihre Antworten auf sensible Fragen systematisch ändern, um Verlegenheit zu vermeiden oder sich selbst im besseren Licht darzustellen. Im Falle der Workloadmessung ist soziale Erwünschtheit denkbar, wenn die Studierenden wegen ihrer (subjektiv) geringen Zeitinvestition in Verlegenheit geraten und daher höhere Werte berichten. *Satisficing* entsteht, wenn die Beantwortung von Fragen in aufmerksamer und umfassender Weise einen beträchtlichen kognitiven Aufwand erfordert und die Befragten sich, um kognitiven Aufwand zu reduzieren, für vereinfachte (anstelle der optimalen) Antwort entscheiden. Wenn es beispielsweise in einer geschlossenen Frage mehrere Antwortmöglichkeiten gibt, können die Studierenden die erste sinnvolle (anstelle der am besten geeigneten) Antwort wählen. Non-Response führt zu Verzerrungen, wenn sich Befragte und Nicht-Befragte systematisch in der interessierenden Variable oder Faktoren, die diese bedingen, unterscheiden. Beispielsweise zeigten einige Studien, dass bei MOOCs (massive open online courses) ein Zusammenhang zwischen der Teilnahme an der Umfrage vor dem Kurs und dem Engagement im Kurs besteht (Samoilova et al. 2018). Diese Beziehung zwischen Studienmotivation und Teilnahmebereitschaft besteht vermutlich auch für andere Arten von Kursen.

Die Nutzung von Tagebüchern ist eine weitere Möglichkeit zur Messung von Workload (Berger und Baumeister 2016). Von dieser Methode wird erwartet, dass sie Daten mit höherer Validität erzeugt als traditionelle Befragungen, da die Respondenten laufend ihren Workload mitprotokollieren können. Tagebuchstudien können daher Probleme mit Erinnerungs- und Urteilsfehler reduzieren, wenn die Teilnehmenden die aufgewendete Zeit sofort nach einer Aktivität protokollieren. Einige Studien zeigen jedoch, dass manche Teilnehmende ihr Verhalten nicht direkt nach dem Ereignis dokumentieren, sondern alle Ereignisse einmal am Tag oder noch seltener protokollieren (Ohly, Sonnentag, Niessen und Zapf 2010). Damit kommen die genannten Vorteile hinsichtlich der Reduktion von Erinnerungs- und Urteilsfehlern in der Praxis weniger stark zum Tragen. Zudem ist der hohe Antwortaufwand der Befragten zu bedenken, da die Teilnehmenden bereit sein sollten, sich zu einer regelmäßigen (zeitintensiven) Protokollierung zu verpflichten. In solchen Situationen werden oft (monetäre oder nicht-monetäre) Incentives genutzt, die jedoch zusätzliche Kosten mit sich bringen. Darüber hinaus können die Studierenden ihr Tagebuch verlieren oder auch vergessen, den Workload zu protokollieren, was dann zu Item- oder Unit-Nonresponse führen kann.

#### *Learning Analytics als neue passive Datenquelle*

Laut der *1st International Conference on Learning Analytics and Knowledge* umfasst die Definition von LA die Messung, Erhebung, Analyse und Berichterstattung von Daten, die während der Lernprozesse entstehen (zitiert nach Long und Siemens 2011, S. 34). Eine wichtige Ergänzung zu

dieser Definition ist, dass sich LA auf bereits vorhandene, maschinenlesbare Daten beziehen (Ferguson 2012, S. 305), die in anderen Forschungsgebieten auch als *found* (Japac et al., 2015), *organic* (Groves 2011) oder *trace data* bezeichnet werden (Golder und Macy 2014; Stier et al. 2019). Darüber hinaus kann man LA als passive, nichtreaktive Messung betrachten, die der Workloadforschung neue Potenziale eröffnet (Webb et al. 2000). Insbesondere eröffnen LA die Möglichkeit, Daten zur Workload passiv zu erheben, ohne die Studierenden zu belasten und sich mit oben genannten Problemen befassen zu müssen. Dies macht es zwingend erforderlich, diese neue Datenquelle zu erforschen und ihre Möglichkeiten und Grenzen, gerade auch im Vergleich zu klassischen Verfahren, auszutesten.

#### *Studien zum Vergleich passiver Datenerhebung mit Befragungsdaten*

Die Nutzung der passiven Datenerhebung wird mit der Verbreitung des Internets und mobiler Endgeräte immer weitreichender. Obwohl es unseres Wissens keine empirischen Untersuchungen zum Thema „studentische Arbeitsbelastung“ mittels LA gibt, können wir aus Studien lernen, die Befragungen und passive Datenerhebungen in anderen Bereichen bereits verglichen haben. Wie die im Folgenden referenzierten Studien zeigen, besteht im Bereich des Online-Verhaltens noch kein Konsens darüber, ob Selbsteinschätzungen der Befragten über- oder unterschätzt sind (im Vergleich zu automatisch getrackten Daten) und wie groß diese Unterschiede sind. In Bezug auf die Messung der Zeitnutzung fand Junco (2013) beispielsweise heraus, dass in einer Stichprobe von Studierenden die Selbsteinschätzungen über die auf Facebook verbrachte tägliche Zeit höher war als die durch Logs angezeigte Zeit. Dennoch hatten die beiden Messungen eine starke positive Korrelation. Jedoch gibt es auch Studien, die ein geringes Maß an Übereinstimmung zwischen den Selbsteinschätzungen und anderen passiven Datenquellen feststellen, wenn es um die Zeit für körperliche Betätigung (Gilbert und Calderwood 2017) oder die berichtete Häufigkeit der Mobiltelefonnutzung geht (Boase und Ling 2013). In der Studie über körperliche Betätigung junger Menschen (14 Jahre alt) im Vereinigten Königreich (n=19000) stellten Gilbert und Calderwood (2017) fest, dass die befragten Jugendlichen ihre körperliche Aktivität im Vergleich zu den Daten der verwendeten Aktivitätstracker unterschätzt haben. In ihrer Studie zur Häufigkeit der Mobiltelefonnutzung zeigten Boase und Ling (2013), dass Selbstangaben nur eine moderate Korrelation mit den für die Validierung verwendeten Log-Daten hatten.

Empirische Untersuchungen zur Verwendung passiver Daten deuten damit darauf hin, dass ihr Verhältnis zu traditionell verwendeten Umfragedaten stark mit dem Kontext und einer bestimmten Zielgruppe zusammenhängt. Daher lassen sich aus anderen Studien, die Selbsteinschätzungen und passive Daten vergleichen, nur erste Indizien zur Belastbarkeit von Umfragedaten und LA zum studentischen Workload ziehen. Aus diesem Grund ist es wichtig, mit entsprechenden LA-Daten zu arbeiten und die Messungen mit selbstberichteten Angaben abzugleichen, um das Potenzial von LA für die Workloadforschung zu verstehen, aber auch Probleme bei deren Nutzung zu erkennen.

Nach unserem Kenntnisstand gibt es keine empirische Literatur zur Brauchbarkeit von LA zur Messung des Workloads von Studierenden. Gibt es deutliche Unterschiede zwischen Umfragedaten

und LA? Wenn ja, inwieweit sind diese Unterschiede auf ein schlechtes Fragebogendesign zurückzuführen? In einem früheren Beitrag (Samoilova et al. 2017) haben wir erste Ergebnisse einer Pilotstudie in einem zwölfwöchigen Onlinekurs vorgestellt, in der wir Ergebnisse aus zwei Datenquellen verglichen haben: LA und einer wöchentlichen Umfrage. In diesem Beitrag vervollständigen wir diese ersten Ergebnisse, indem wir uns zwei zusätzliche Kurse ansehen. Darüber hinaus präsentieren wir Ergebnisse von kognitiven Interviews und betrachten die Workloadmessung für einen dritten Kurs, bei dem das Antwortformat im Fragebogen von Stunden auf Minuten umgestellt wurde.

## **2. Daten und Methoden**

Im Mittelpunkt der Studie standen drei Kurse, die im Rahmen des vom Bundesministerium für Bildung und Forschung geförderten Online-Programms „International Program in Survey and Data Science“ (IPSDS) für Berufstätige an der Universität Mannheim stattfanden. IPSDS ist darauf ausgerichtet, die wachsende Nachfrage nach Fachkräften für Datenerhebung und -analyse durch ein berufsbegleitendes Studium zu decken. Die ersten zwei Kohorten von 31 Studierenden aus 15 Ländern starteten im Februar 2016 und 2017. Das Programm findet auf Englisch statt (Kreuter et al. 2018).

Die Hauptdatenerhebung basiert auf einer Kohorte von Studierenden (n=16), die zwei Einführungskurse belegt haben. Der Kurs „Fundamentals of Survey and Data Science“ (Kurs 1) fand im Frühjahrssemester 2016 statt und dauerte 12 Wochen. Der Kurs „Data Collection Methods“ (Kurs 2) folgte direkt darauf im Sommer 2016, wurde von 13 der ursprünglichen 16 Studenten besucht und dauerte 11 Wochen. Die 13 Studierenden, die Kurs 2 belegten, nahmen an kognitiven Interviews vom 1. September bis 13. Oktober 2016 teil. Zusätzlich zu den Kursen 1 und 2 der ersten Kohorte wurden Daten von 15 weiteren Studierenden der zweiten Kohorte von IPSDS im Kurs „Fundamentals of Survey and Data Science“ (Kurs 3) im Frühjahrssemester 2017 erhoben.

Die Kurse wurden dabei so entwickelt, dass die Arbeitsbelastung – bei einer erwarteten Arbeitsbelastung von 10 Stunden pro Woche – gleichmäßig auf die jeweils 11 (Kurs 2) bzw. 12 Wochen (Kurs 1 und 3) verteilt ist. Das Kursmaterial umfasste zuvor aufgezeichnete Vorlesungsvideos, wöchentliche (Kurs 1 und 3) bzw. zweiwöchentliche Aufgaben (Kurs 2) sowie wöchentlich erforderliche und empfohlene Lektüre. In Kurs 2 fand die Kommunikation mit den Dozierenden innerhalb des Online-Forums statt, in den Kursen 1 und 3 wurden wöchentlich synchrone Online-Meetings mittels des Online-Videokonferenzsystems BlueJeans (ca. 50 Minuten pro Woche) durchgeführt.

Während in der ersten Kohorte die Mehrheit der Studierenden Frauen sind (10 Frauen und sechs Männer), ist in der zweiten Kohorte der Anteil der Männer und Frauen fast gleich (sieben Männer und acht Frauen). Tabelle 1 stellt weitere Merkmale der Studierenden in den beiden Kohorten dar.

Alle Studierenden waren teil- oder vollzeitbeschäftigt. Alle Studierenden mussten erst den Einführungskurs „Fundamentals of Survey and Data Science“ besuchen. Danach konnten sie Kurse ihrer Wahl aus unterschiedlichen Modul-Blöcken belegen. Um das Hintergrundwissen der Studierenden zu messen, wurden sie vor Beginn des Programms gebeten, ihre Vertrautheit mit den Inhalten des Einführungskurses zu berichten. Wir haben sie auch gebeten, ihr erwartetes Zeitbudget anzugeben, um sicherzustellen, dass es für sie möglich ist, an dem Kurs teilzunehmen.

**Tab 1:** Merkmale der Studierenden basierend auf einer Vorkursbefragung.

	Kohorte 2016 (Kurs 1 und 2), n=16			Kohorte 2017 (Kurs 3), n=15		
	Mittelwert/ %	Median	s	Mittelwert/ %	Median	s
Arbeitszeit (pro Woche)	40.75	41	12.19	41.07	43	9.76
Stunden/Woche, die voraussichtlich für den Einführungskurs verwendet werden	8.69	8	3.02	8.8	10	2.96
Vertrautheit mit den Inhalten des Einführungskurses:						
-Überhaupt nicht vertraut						
-Ein wenig vertraut	0%			0%		
-Etwas vertraut	25%			13%		
-Sehr vertraut	44%			73%		
	31%			13%		

Anmerkung: Die Kurse 1 und 2 wurden von den gleichen Studierenden besucht (Kohorte 2016). Der Kurs 3 wurde von der zweiten Kohorte (Kohorte 2017) absolviert.

### *Befragung*

Die selbstberichteten Daten stammen aus 11 (Kurs 2) bzw. 12 (Kurse 1 und 3) wöchentlichen Onlinebefragungen. Für die erste Kohorte (Kurse 1 und 2) wurden Fragebögen in der Befragungssoftware EFS 10.9 von Unipark programmiert. Der Webfragebogen für die zweite

Kohorte (Kurs 3) wurde in Qualtrics programmiert. Da das Programm auf Englisch durchgeführt wurde, erfolgte die gesamte Datenerhebung auf Englisch.

Am Ende der jeweiligen Woche wurden den Studierenden nach Ablauf der Frist für die Einreichung der wöchentlichen Hausaufgaben Einladungen zu den Befragungen mit individualisierten URLs per E-Mail zugesandt. Durch Anklicken der URL in der E-Mail-Einladung wurden die Studierenden automatisch zum Webfragebogen weitergeleitet. Die Rücklaufquote für Befragungen in den Kursen 1 und 2 der ersten Kohorte betrug in allen Wochen 100%. Für die Befragungen in Kurs 3 der zweiten Kohorte ähnelte die Rücklaufquote mit 53% eher einer realistischen Situation. Nur acht der 15 Studierenden beantworteten die Befragungen über den gesamten Kurs hinweg vollständig.

Der wöchentliche Fragebogen enthielt vier Fragen zum Workload, definiert als Zeitaufwand für individuelle Lernaktivitäten (siehe Abbildung 1), zur Lernmotivation, gemessen anhand von drei Items aus der von Keller (2009) entwickelten ARCS3-Motivationsskala für webbasierten Unterricht, zur Zufriedenheit mit den Lernmaterialien sowie zum wahrgenommenen Stress. Die Operationalisierung des Workloads wurde zwischen den Befragungen der Teilnehmenden der ersten Kohorte im Jahr 2016 (Kurse 1 und 2) und der Befragung im Jahr 2017 (Kurs 3) angepasst. Während 2016 noch nach dem Workload offen in Stunden gefragt wurde, wurde die Frage 2017 auf offene Angaben in Minuten umgestellt. Ziel der Änderung der Operationalisierung war es, die Feinkörnigkeit der selbst-berichteten Daten zu erhöhen und besser mit den LA vergleichen zu können.

**During the past week, how much time did you spend (in hours) on the activities below?**  
If you don't know precisely, then please provide your best estimate.

Watching pre-recorded lecture videos	<input type="text"/>
Doing required readings	<input type="text"/>
Doing recommended readings	<input type="text"/>
Completing course assignments	<input type="text"/>
Discussing course topics with other participants outside of the BlueJeans meetings	<input type="text"/>
Other course-related work	<input type="text"/>
Paid Work	<input type="text"/>
Household chores	<input type="text"/>
Child care	<input type="text"/>
Leisure	<input type="text"/>

**Abb. 1:** Question on workload in the weekly survey instrument

### *Learning Analytics*

LA wurden für die Nutzung der Vorlesungsvideos erhoben. Die Videos wurden durch die Video-Plattform Mediasite in die Lernplattform Moodle eingebettet, so dass die Videos nur per Streaming angesehen werden konnten. Für die Studierenden war es möglich, ein Video anzuhalten, das Video vorwärts und rückwärts zu spulen, es (bzw. Teile davon) erneut abzuspielen und die



Geschwindigkeit des Videos zu ändern (es war sowohl möglich die Geschwindigkeit zu erhöhen als auch diese zu verringern) (siehe Abbildung 2). Die Daten wurden für jedes Video und jede Person einzeln erhoben.

Jede Woche umfasste etwa eineinhalb Stunden Videomaterial. Die Vorträge wurden in mehrere kürzeren Videos unterteilt (zwischen vier und elf Videos pro Woche). Für die Erfassung und den Export der Daten wurde ein von Mediasite bereitgestellter Standard-Analyseberichte verwendet. Nach Abschluss des Kurses wurden die Daten als XML-Datei exportiert und mit R version 3.5.3 (R Core Team 2018) bereinigt. Zusätzlich zur Anzahl der Aufrufe pro Video beinhalteten die Daten Informationen dazu, wieviel Prozent eins bestimmten Videos angesehen und wie lange das Video abgespielt wurde (zur Erinnerung: Videos konnten schneller oder langsamer abgespielt und wiederholt angesehen werden). Die Gesamtdauer und der prozentuale Anteil der Videonutzung sind stark miteinander korreliert (0,85, 0,87, und 0,89 für die drei Kurse). Für dieses Papiers konzentrieren wir uns ausschließlich auf die Gesamtdauer (in Minuten) der Videonutzung.

Sowohl für die Nutzung der LA als auch die Befragungsdaten wurde die Einwilligung der Teilnehmenden eingeholt. Um die Selbsteinschätzungen aus der Befragung mit den LA zu verknüpfen, haben wir jeder Person eine ID zugewiesen.

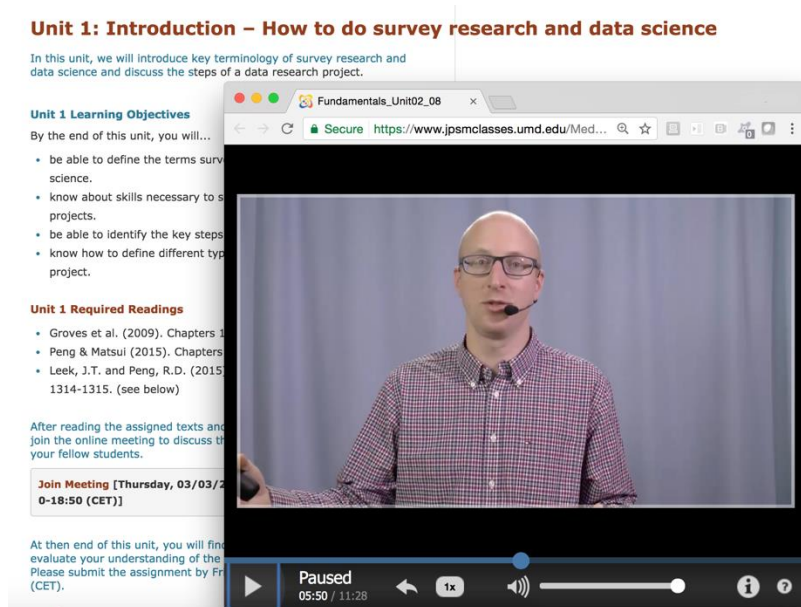


Abb. 2: Videoplayer eingebettet in die Lernplattform Moodle.

### *Kognitive Interviews*

Qualitative semi-strukturierte Interviews wurden mithilfe der Online-Konferenzsoftware BlueJeans nach dem Kurs 1 und 2 auf Englisch durchgeführt. Sowohl die Interviewerin als auch die Befragten hatten ihre Webcam während des gesamten Interviews eingeschaltet. Nach der Einwilligung der Befragten wurden die Interviews aufgezeichnet.

Zu Beginn des kognitiven Interviews wurden zwei allgemeine Fragen zu den Erfahrungen mit den Kursvideos und darauf basierten Selbsteinschätzungen zum Zeitumfang gestellt. Die erste Frage betraf die Art und Weise, wie die Studierenden mit den Videovorlesungen umgegangen sind. Insbesondere interessierte uns, ob sie die Videos pausieren, um Notizen zu machen oder das Thema andernorts zu recherchieren, und wenn ja, wieviel Zeit das in Anspruch nimmt. Diese Frage ist wichtig, da LA nur die Zeit erfassen, in der ein Video abgespielt wird. Wenn die Studierenden die Videovorträge häufig unterbrechen, jedoch während der Unterbrechung anderen Kurs-relevanten Aktivitäten nachgehen, sind Unterschiede in den Befragungsdaten und den LA zu erwarten. Zudem wären Abweichungen dann eher hinsichtlich der Reichweite der Informationen aus den LA problematisch als für die Befragungsdaten.

Im zweiten Schritt wurden die Studierenden nach ihren Erfahrungen bei der Beantwortung der Fragen zum Workload in den letzten beiden Kursen gefragt: „During the spring and summer semester, you were invited to participate in the weekly evaluation survey where we asked you (among others) about time spent on watching pre-recorded lecture videos. [Reading aloud the question: “During the past week, how much time did you spend (in hours) on watching pre-recorded lecture videos?”]. How hard was it for you to answer that question?” Nach diesen allgemeinen Fragen folgten Nachfragen zum Frageverständnis und zur Informationsbeschaffung: „How did you arrive at the answer?“; „What do you understand under *watching pre-recorded lecture videos*?“ Ziel dieses Schrittes war es, herauszufinden, ob die Teilnehmenden während der betreffenden Kurse Schwierigkeiten mit der Workloadfrage hatten.

Als letzten Schritt haben wir die Technik des lauten Denkens<sup>1</sup> genutzt, um Schätzstrategien zu identifizieren, welche die Teilnehmenden verwenden, um zu einer Zeitschätzung zu gelangen. Erst wurde das laute Denken mit den Teilnehmenden geübt. Wir haben die dazu übliche Übungsfrage verwendet: “Try to visualize the place where you live and think about how many windows are there. As you count up the windows, tell what you are seeing and thinking about.” Die Studierenden hatten keine Schwierigkeiten mit dieser Trainingsaufgabe. Als nächstes wurde ihnen die Frage nach ihrem Videoworkload im aktuellen Kurs, den sie gerade besuchten, gestellt: „During the past week, how much time (in hours) did you spend on watching pre-recorded lecture videos?“ Die Befragten wurden angewiesen, bei der Beantwortung der Frage laut zu denken.

## *Analyse*

---

<sup>1</sup> Die Technik des lauten Denkens wurde ursprünglich von Psychologen entwickelt, um Gedankenprozesse bei Problemlösungsaufgaben zu erfassen (Ericsson und Simon 1980). In kognitiven Interviews (siehe Willis 2005) wird der Begriff think-aloud verwendet, um eine ganz bestimmte Methode zu beschreiben, bei der die Probanden ausdrücklich angewiesen werden, bei der Beantwortung der Umfragefragen laut zu denken. Die Interviewerin ist bei dieser Technik primär passiv. Laut Willis (1999), hat diese Methode sowohl Stärken als auch Schwächen. Zu den Hauptvorteilen dieser Methode gehören: 1. relative Freiheit von Interventionen des Interviewers und 2. die Befragten können durch das offene Format unvorhersehbare Informationen liefern. Der Hauptmangel besteht darin, dass die befragte Person üben muss, laut zu denken, da diese Methode keine übliche Tätigkeit ist. Ein weiteres Problem ist, dass die Person dazu gebracht werden könnte, mehr zu denken als im eigentlichen Prozess der Beantwortung einer Frage.

Im nächsten Abschnitt präsentieren wir die Ergebnisse basierend auf der Umfrage, den LA und den kognitiven Interviews. Zur Anonymisierung und weil die Analyse es nicht direkt erforderte, haben wir die Daten aus den kognitiven Interviews nicht mit den Befragungsdaten und LA verknüpft. Um LA- und Umfragedaten verknüpfen zu können, haben wir die Einwilligung der Studierenden eingeholt.

Für die Befragungs- und LA-Daten berichten wir deskriptive Statistiken wie Mittelwerte, Mediane und Standardabweichungen sowie bivariate Zusammenhänge mittels Korrelationen. Betrachtet werden hierbei sowohl Korrelationen innerhalb von als auch zwischen Individuen. Zur Berechnung der entsprechenden Korrelationen wurden zunächst personenspezifische Mittelwerte der Workloadmessungen gebildet, um Korrelationen zwischen Personen zu schätzen. Anschließend wurden die individuellen wöchentlichen Beobachtungen um den personenspezifischen Mittelwert zentriert, um personenspezifische Unterschiede in zeitkonstanten Merkmalen zu beseitigen und Korrelationen innerhalb von Individuen zu berechnen.

Zusätzlich präsentieren wir Ergebnisse von bivariaten Random-Effects- und Fixed-Effects-Modellen, wobei die umfragebasierten Selbstangaben als Prädiktor für LA-Messung fungieren. Random Effects Modelle bieten ein flexibles Werkzeug, da sie sowohl feste als auch zufällige Effekte beinhalten. Da die Workload-Werte innerhalb von Personen, Wochen und Kursen genestet und damit nicht unabhängig sind, haben wir diese drei Variablen als Zufallseffekte in einem längsschnittlichen Mehrebenenmodell modelliert. LA werden als fester Effekt modelliert. Das Modell ist gegeben durch:

$$Y_{ijk} = X_{ijk}\beta + Z1_{,ijk}b_1 + Z2_{,ijk}b_{ij} + Z3_{,ijk}b_{ijk} + \epsilon_{ijk}.$$

Dabei bezeichnet  $Y$  den beobachtbaren Zufallsvektor der LA der Person  $i$  in einer Kurswoche  $t$ ,  $X$  eine bekannte Modellmatrix der festen Effekte des selbsteingeschätzten Workloads,  $Z1$  eine bekannte Designmatrix der zufälligen Effekte der Personen,  $Z2$  der zufälligen Effekte der Wochen,  $Z3$  der zufälligen Effekte der Kurse, und  $\epsilon$  den nicht-beobachtbaren Zufallsvektor der Fehler. Die Analyse wurde mittels R Pakets lme4 (Bates et al. 2015) durchgeführt.  $i, j, k$  bezeichnen Personen, Wochen und Kursebenen.

Um sicherzustellen, dass wir die individuelle Heterogenität der Studierenden erfasst haben, wurde auch ein Fixed-Effects-Modell berechnet:

$$Y_{it} = \beta X_{it} + \alpha_i + e_{it}.$$

Dabei bezeichnet  $Y$  die abhängige Variable LA der Person  $i$  in einer Kurswoche  $t$ ,  $X$  die unabhängige Variable selbsteingeschätzter Workload der Person  $i$  in einer Kurswoche  $t$ ,  $\alpha$  den unbekanntem Achsenabschnitt für jede Entität und  $\epsilon$  den idiosynkratischen Fehler. Die Auswertung erfolgte mit dem R Paket plm (Croissant und Millo 2008).

Da während des Interviews eine Person aus Kurs 1 erwähnte, dass es möglich sei, die Vorlesungsvideos herunterzuladen (obwohl sie/er es nicht getan hatte), haben wir die Robustheit der Modelle gegenüber dieser Art von Messfehlern mittels LA überprüft. Die Online-Plattform und die Programmordnung sahen eindeutig kein Herunterladen der Vorlesungsvideos vor. Allerdings lässt sich davon ausgehen, dass man das bei einem gewissen Maß an technischem Fachwissen tun könnte. Da die Downloads in der Regel relativ schnell sind, haben wir die ursprünglichen Modelle mit einer reduzierten Stichprobe verglichen, bei der wir Beobachtungen mit LA von weniger als fünf Minuten ausgeschlossen haben. Die im Folgenden präsentierten Befunde zeigten sich auch in diesem Robustheitstest.

Die Interviews wurden mit Hilfe von *cognitive coding* (Willis 2015) analysiert. Die Ansätze von *cognitive coding* verwenden Codes, die eng mit dem vierstufigen Modell des Frage-Antwort-Prozesses von Tourangeau (1984) verbunden sind. Das 4-Stufen-Modell bezieht sich auf folgende Prozesse: Verständnis der Frage, Abruf relevanter Informationen, Integration der Informationen und Entscheidung über Antwortmöglichkeiten, Zuordnung der Antwortmöglichkeiten auf formale Antwortkategorien. Zunächst wurden für alle Teilnehmenden Textzusammenfassungen erstellt. Anschließend wurden die zusammengefassten Texte bearbeitet und jedem Element ein oder mehrere kognitive Codes zugeordnet (z.B. Verständnis, Erinnerung).

### 3. Ergebnisse

Tabelle 2 enthält deskriptive Statistiken darüber, wie die geschätzte Arbeitsbelastung (in Minuten) je nach Messmethode variiert. Während für die Befragungsdaten in 2016 die Zeit von Stunden in Minuten umgerechnet wurden, hat die 2017er Kohorte die Zeit in Minuten angegeben. Für alle drei Kurse ist der durchschnittliche selbstberichtete Workload aus der Befragung deutlich höher als der Workload gemäß LA.

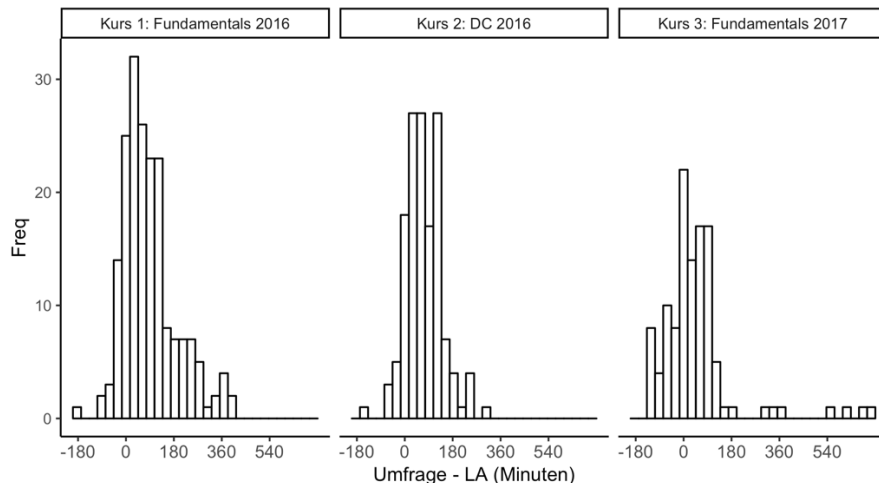
**Table 2:** Deskriptive Statistik für das Ansehen von aufgezeichneten Videovorlesungen (in Minuten/pro Woche) für jeden Kurs.<sup>a</sup>

	LA (Minuten pro Woche)				Befragung (Minuten pro Woche)			
	n <sup>b</sup>	Mittelwert	Median	s	n <sup>c</sup>	Mittelwert	Median	s
<b>Kurs 1: Fundamentals 2016</b>	192	73.54	74.00	46.0 4	192	161.25	120	100.5 7

<b>Kurs 2: Data Collection 2016</b>	143	51.00	51.33	46.6	4	142	121.71	120	67.07
<b>Kurs 3: Fundamentals 2017</b>	180	64.08	50.50	66.4	3	126	116.00	100	139.5

Anmerkungen: <sup>a</sup> Für die Kurse 1 und 2 wurden die Umfrageergebnisse in Stunden in Minuten umgerechnet. <sup>b</sup> Gesamte Stichprobe. <sup>c</sup> Stichprobe, nachdem fehlende Werte ausgeschlossen wurden.

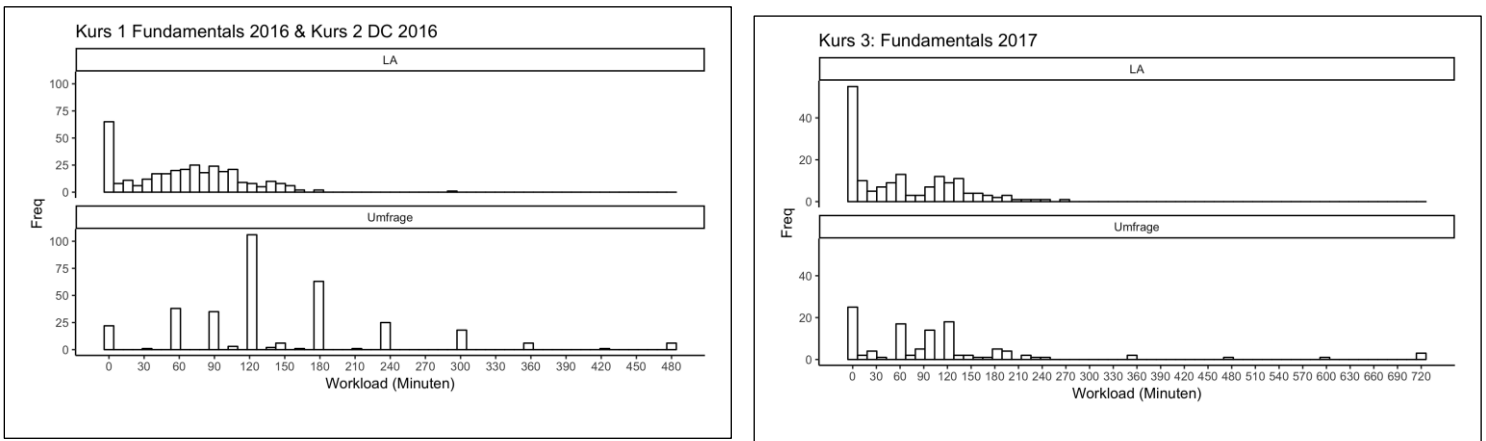
Abbildung 3 zeigt die Verteilung der individuellen Differenzen in Minuten zwischen den Befragungsdaten und den LA. Bei allen drei Kursen ist die Verteilung deutlich linkssteil. Die Verteilung sieht für Kurs 3 etwas anders aus, was auf einen höheren Prozentsatz an fehlenden Werten zurückzuführen sein könnte. Es ist deutlich zu erkennen, dass die Befragungswerte fast durchweg merklich höher sind als die LA. Selbstberichte des Workloads führen also vermutlich zu einer deutlichen Überschätzung der zeitlichen Belastung. Diese Differenz gilt unabhängig vom Zeitpunkt über alle Wochen (siehe Appendix 1).



**Abb. 3:** Histogramme der individuellen Differenz zwischen Umfragedaten (Selbsteinschätzungen) und LA in Minuten

Abbildung 4 visualisiert den Unterschied in der Feinkörnigkeit der LA-Daten im Vergleich zu den selbstberichteten Daten. Während die Verteilung für Kurs 3 mit der Workloadmessung in Minuten etwas von der Verteilung für die Kurse 1 und 2 abweicht und feinkörnigere Werte aufweist, so ist dennoch weiterhin zu beobachten, dass die Befragten – so legen es auch die qualitativen Interviews

nahe – abschätzen und runden. In beiden Fällen sind zwei Stunden der am häufigsten verwendete Wert, unabhängig davon, ob die Workload in Minuten oder Stunden abgefragt wurde.



**Abb. 4:** Histogramme für Befragungsdaten und LA nach Kursen. Die Kurse 1 und 2 wurden von den gleichen Studierenden besucht und die Zeit in Stunden gemessen wurde. Für Kurs 3 wurde die Zeit in Minuten gemessen.

Um die Beziehung zwischen LA und Befragungsdaten weiter zu untersuchen, unterscheiden wir zwischen Korrelationen innerhalb von und zwischen Individuen. Korrelationen zwischen Individuen decken generelle Unterschiede und Assoziationen auf Personenebene auf (z.B. generelle Unterschiede in der Studienmotivation und Lernbereitschaft) und zeigen, zu welchem Grad Personen, die zum Beispiel im Durchschnitt eine hohe Workload in der Befragung angeben, auch gemäß LA im Durchschnitt eine hohe Workload aufweisen. Im Gegensatz dazu liefern Korrelationen innerhalb der Individuen Informationen darüber, inwieweit wöchentliche Schwankungen im Workload laut Befragung für ein und dieselbe Person mit wöchentlichen Schwankungen im Workload laut LA einhergehen. Tabelle 3 enthält entsprechende Korrelationen innerhalb und zwischen Personen von LA und Befragungsdaten. Wie man sieht, sind LA und selbst-berichteten Daten auf Personenebene für zwei der drei Kurse nur sehr schwach miteinander korreliert. Einzig für Kurs 3 fällt die Korrelation zwischen den Teilnehmenden deutlich stärker aus, was jedoch auch mit fehlenden Werten und einer selbstselektierten Stichprobe zusammenhängen könnte. Für Kurs 1 und 2 fallen die Korrelationen innerhalb von Personen deutlich stärker aus und sind auch auf dem 1%-Niveau signifikant. Personen, die also in einer Woche eine überdurchschnittliche (unterdurchschnittliche) Workload in der Befragung angeben, erzielen tendenziell auch höhere (niedrigere) Werte bei den LA. Allerdings ist dieser Zusammenhang nur mittel bis schwach. Für Kurs 3 ergibt sich zudem kein entsprechender Zusammenhang, was jedoch erneut mit den Datenlimitationen zusammenhängen könnte.

**Table 3.** Korrelationen innerhalb von und zwischen Individuen

	<b>innerhalb von Individuen</b>	<b>zwischen Individuen</b>
--	---------------------------------	----------------------------

	n <sup>a</sup>	Korrelation (LA/Umfragedaten)	n <sup>a</sup>	Korrelation (LA/Umfragedaten)
<b>Kurs Fundamentals/ 2016</b> Kohorte	14	0.09	168	0.23***
<b>Kurs Data Collection/ 2016</b> Kohorte	12	0.05	132	0.38***
<b>Kurs Fundamentals/ 2017</b> Kohorte	8	0.44	93	0.06

Signifikanzniveaus: \*\*\*  $p \leq 0.01$  \*\*  $p \leq 0.05$  \*  $p \leq 0.1$

Anmerkungen: <sup>a</sup> c Stichprobe, nachdem Ausreißer ausgeschlossen wurden.

Tabelle 4 zeigt die Ergebnisse der Random- und Fixed-Effects-Modelle. Die Resultate zeigen, dass in beiden Modellen der Effekt der Umfragedaten positiv und statistisch signifikant ist. Dennoch ist der Regressionskoeffizient eher klein, bedenkt man dass LA und Umfrage eigentlich das gleiche Konstrukt messen sollten. Auch ist in beiden Modellen das  $R^2$  sehr klein, was darauf hindeutet, dass die Umfragedaten nur einen kleinen Teil der Variation in den LA erklären können.

Tabelle in Appendix 2 zeigt die Ergebnisse derselben Modelle, jedoch für die reduzierte Stichprobe. Um mögliche Verzerrungen der LA aufgrund des Herunterladens der Videos zu berücksichtigen und die Robustheit der Ergebnisse zu überprüfen, wurden Zeilen mit LA-Werten unter 5 Minuten entfernt. Die Modelle ergaben ähnliche Schätzungen mit Konfidenzintervallen zwischen 0.04 und 0.16 für das Random-Effects-Modell und 0.08 und 0.21 für das Fixed-Effects-Modell.

**Table 4.** Lineare Regressionen zum studentischen Workload: Umfragedaten als Prädiktor der LA

Modell	Random Effects Modell		Fixed Effects Modell	
	Koeffizient	95% CI	Koeffizient	95% CI
Umfragedaten	0.17***	0.11, 0.23	0.18***	0.12, 0.25
Konstante	45.21***	26.39, 64.23	-	-
$R^2$ ( <i>marginal</i> für Random-	0.09		0.08	

Effects-Modell)		
Log Likelihood für Random-Effects-Modell/ F für Fixed-Effects	-1971.29	28.83***
n	382	382

Signifikanzniveaus: \*\*\*  $p \leq 0.01$  \*\*  $p \leq 0.05$  \*  $p \leq 0.1$

Die Ergebnisse der kognitiven Interviews zeigen, dass das Hauptproblem bei der Frage höchstwahrscheinlich nicht das Verständnis der Frage war, sondern eine Schwierigkeit bei der Erinnerung und Einschätzung des Workloads bestanden. Keiner der Befragten dachte daran, dass das Ansehen von Videovorlesungen bedeuten könnte, Videos für Notizen oder für andere Zwecke anzuhalten. Sechs Teilnehmende haben explizit erwähnt, dass es schwierig war, sich an die Zeit zu erinnern. Zwei Studierende sagten, sie führten ein Tagebuch. Fünf Befragte erklärten, dass es nur am Anfang schwierig war, da sie später eine Strategie entwickelt haben, um das Problem der Erinnerung und Einschätzung zu lösen. Doch ihre Strategien sorgten nach ihrem Dafürhalten nicht unbedingt für die optimale Antwort. Wie eine der fünf Personen erklärte: „It was difficult at the beginning, but got easier with time when I found a strategy how to come up with an answer. But it could be questioned, whether that was the best strategy. I would actually try to come up with a number of hours I had at my disposal given how much I worked.“

Das Beispiel der Strategie, die diese Person beschrieben hat, ist insofern wichtig, als es in eines der beiden Muster fällt, die wir während des Think-aloud-Verfahrens entdeckt haben. Die Teilnehmenden schienen eine bestimmte Strategie zu haben, mit der sie den Workload abgerufen und geschätzt haben. Diese beiden Strategien könnten als „ereignisbasierte Schätzung“ und „Nachdenken über Anzahl und Länge der Videos“ bezeichnet werden. Die erste Strategie bedeutet, dass Menschen ihre Zeiteinschätzung auf fokale Ereignisse zurückführen. Sie schätzen auch die Zeit dieser Ereignisse, indem sie sich an bestimmte Regelmäßigkeiten und Routinen erinnern, z.B.: „I remember it was exactly one hour, because I had one hour before the meeting started. And on Wednesday, also I came home at I think it was 6:30 and I had to leave quarter past eight, so I had about 2 hours. It was easy, because video watching time was framed by other things I had to do.“ Die zweite Strategie bezieht sich auf die Länge und Anzahl der Videos. Man kann davon ausgehen, dass für diejenigen, die mindestens ein Video vollständig angesehen haben, diese Strategie einfacher sein könnte. Doch es war nicht klar, ob die Studierenden die Geschwindigkeit, mit der sie die Videos angesehen haben, in Betracht gezogen haben. Dies illustriert ein Textauszug aus einem



der Interviews: "It is really hard to remember, but I guess it was Wednesday, no it was Sunday, now I remember. And I watched all the videos once, so it was around 90 minutes. I remember that there were 6 or 7 and all of them lasted about 8-15 minutes. Well, then it is not such a good estimate. I don't know 90 minutes or 2 hours, that's what popped up in my head, but if you think it through I don't think I spent as much time as 90 minutes. I think it was less."

Unabhängig von den genauen Strategien ist deutlich zu erkennen, dass Menschen Näherungswerte und Rundungen verwenden. Diese Rundungen sind eine Erklärung für die Unterschiede zwischen der selbstberichteten Zeit und der mittels LA gemessenen Zeit. Unklar ist dabei einerseits, ob die Strategie von Anfang an gleich geblieben ist, ob kontextabhängig von einer Strategie zur anderen gewechselt wurde und inwieweit Lerneffekte im Sinne einer Optimierung der Strategie aufgetreten sind. Andererseits ist auch von zukünftiger Forschung zu klären, ob Unterschiede in diesen Strategien unterschiedliche Auswirkungen auf die Qualität der Daten haben können.

Obwohl die Studierenden in den kognitiven Interviews soziale Erwünschtheit nicht direkt als Grund für die Vermeidung zu geringer Workloadangaben angesprochen haben, sind dadurch verzerrte Antworten nicht auszuschließen. Denn die Frage nach dem Workload könnte für manche Teilnehmende sensibel sein; gerade wenn davon ausgegangen wird, dass Lehrende diese Information auf Individualebene oder auch in aggregierter Form zugreifen können.

#### **4. Fazit**

Das Ziel dieser Studie war es, Aufschluss über die Nutzung von LA für die Workloadforschung zu geben. Der Workload der Studierenden spielt eine zentrale Rolle für die Unterrichtspraxis, aber auch für den studentischen Lernerfolg, Stressempfinden und die Work-Life-Balance. Daher ist es wichtig, den Workload bei der Gestaltung von Studium und Lehre zu berücksichtigen und geeignete Instrumente zu Workloadmessung zu entwickeln. Da die Umfrage das am häufigsten verwendeten Instrument zur Messung von studentischem Workload ist, haben wir Umfragedaten mit einer möglichen Alternative – Learning Analytics (LA) – verglichen.

Die Ergebnisse zeigen, dass die LA-Logs und die wöchentlichen Befragungen unterschiedliche Schätzungen für die Zeit liefern, die für das Ansehen von voraufgezeichneten Vorlesungsvideos aufgewendet wird. Darüber hinaus ist die Korrelation zwischen diesen beiden Datenquellen auf Personenebene sehr schwach. Regressionsmodelle, die Mehrebenenstruktur und den längsschnittlichen Charakter der Daten berücksichtigen, bestätigten eine schwache Beziehung zwischen den Umfragedaten und LA.

Die Ergebnisse aus kognitiven Interviews zeigten, dass die Umfragefrage über den Workload wie beabsichtigt verstanden wurde. Daher kann LA in diesem Fall für die Überprüfung der Datenqualität der Umfrage verwendet werden. Da sowohl LA- als auch Umfragedaten eigentlich das gleiche

Konstrukt messen sollten, sind die Unterschiede zwischen den beiden Messungen als stark zu bewerten.

Angesichts des Forschungsstands zu potenziellen Problemen bei Befragungen zum Lernverhalten ist es nicht überraschend, solche Unterschiede zu sehen. Deren Stärke ist jedoch durchaus überraschend. Kognitive Interviews unterstützen unsere These, dass die Ursache dafür Erinnerungs- und Urteilsfehler sind und nicht das Verständnis der Frage über den Workload. Es ist daher auch unwahrscheinlich, dass sich das Problem durch die Verbesserung eines Umfrageinstruments beheben lässt. Auch eine noch feinkörnigere Messung mit mehr als einer Messung pro Woche, um Erinnerungsproblem zu reduzieren, ist kaum realistisch und dürfte schon alleine an der mangelnden Teilnahmebereitschaft der Studierenden scheitern.

Wir hoffen daher, dass unsere Studie eine kritische und reflektierte Nutzung von befragungsbasierten Workloaddaten anregt. Unsere Studie deutet darauf hin, dass die durch die Umfrage ermittelten Workload-Werte (insbesondere Durchschnittswerte) mit großer Vorsicht betrachtet werden sollten (z.B. bei der Umstellung auf ECTS oder Änderung des Kursdesigns). Es bedarf weiterer Forschung, um noch besser zu verstehen, wie und für welche Fragestellungen sie verwendet werden können.

Unsere Ergebnisse zeigen aber auch, dass LA nicht nur dazu verwendet werden können, das Engagement der Studierenden (inclusive Lernverhalten, siehe Järvelä und Renninger 2014) zu erforschen (wofür es oft verwendet wird), sondern auch, um den studentischen Workload zu messen. Wenn möglich, könnte LA als Ersatz für Umfragedaten verwendet werden oder Befragungsergebnisse zumindest ergänzen. Beispielsweise könnte man LA verwenden, um die durchschnittliche Zeit, die Studierende für einzelne Vorlesungsvideos verwenden, vorherzusagen und diese Daten sowohl zur Evaluierung des Kurses als auch zur Unterstützung neuer Studierende mit Informationen über ihre potenzielle Zeitinvestition zu verwenden.

Dennoch hat unsere Studie eine Reihe von Einschränkungen, die berücksichtigt werden müssen und die weitere Forschung erfordern. Mit einer größeren Stichprobe könnten wir die Beziehung zwischen den Umfragedaten und LA viel besser modellieren. Unter anderem könnten wir multivariate Modelle anwenden und Umfragedaten zur Vorhersage der tatsächlichen Zeit in unterschiedlichen Untergruppen untersuchen. In unserer Studie haben wir uns auf eine bestimmte Art von Aktivität konzentriert und nicht auf die gesamte Arbeitsbelastung. Da wir gelernt haben, dass Studierende in ihren Strategien einen einzigen Bezugspunkt für die gesamte wöchentliche Kursarbeit verwenden, könnte man untersuchen, ob die Unterschiede bestehen bleiben, wenn man sich auf den gesamten Kurs-Workload und nicht auf einzelne Aktivitäten konzentriert. Eine solche Studie setzt jedoch voraus, dass LA aussagekräftige Daten über alle Workload-Aktivitäten liefern.

Zusätzlich zu einer weiteren empirischen Untersuchung der oben genannten LA-Daten erfordert die Nutzung von LA für die Workload-Forschung auch eine Diskussion über Ethik- und

Datenschutzbelange. Die LA-Community hat bereits begonnen, diese Themen aktiv zu diskutieren (Drachslar und Greller 2016). Eine interessante Frage betrifft zum Beispiel die nach den Effekten der Einwilligung zur Messung auf das studentische Verhalten. Wie empfinden die Studierenden die Einwilligung im Rahmen der Erfassung ihrer LA-Daten und deren möglicher Verknüpfung mit anderen Daten? Welche sind die besten Praktiken, um die Einwilligung einzuholen, und was wissen wir über die damit verbundenen Selektionseffekte? Während diese Fragen im Rahmen der passiven Datenerhebung behandelt werden, müssen wir sie auch im unmittelbaren Kontext der LA-Forschung betrachten.

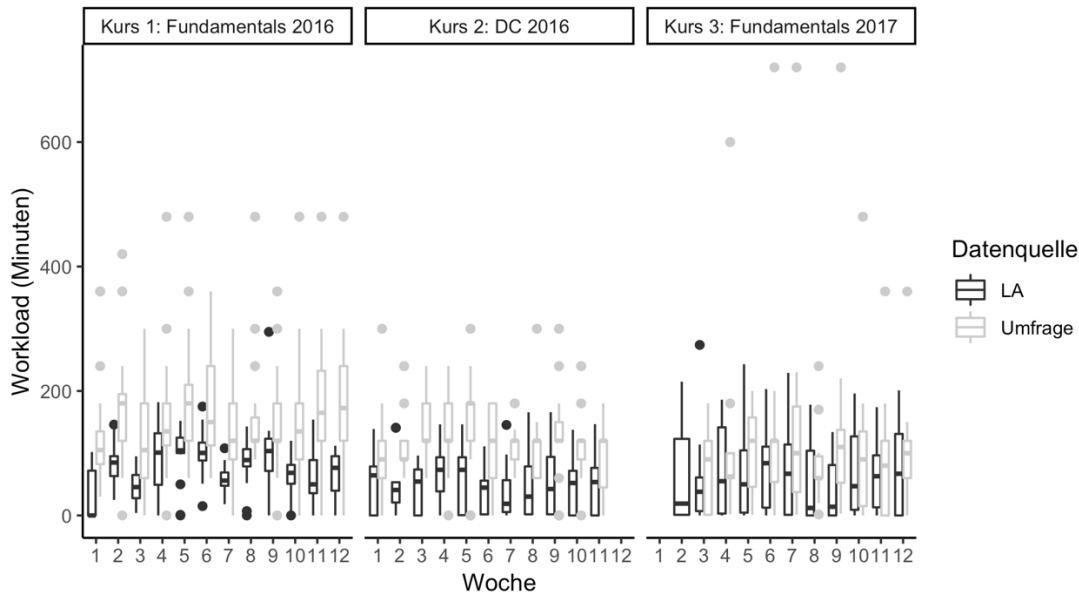
## Bibliographie

- Bates, D., Maechler, M., Bolker, B., und S. Walker. 2015. „Fitting Linear Mixed-Effects Models Using lme4.“ *Journal of Statistical Software*, 67(1): 1-48.
- Berger, R., und B. Baumeister. 2016. „Messung von Studentischem Workload: Methodische Probleme Und Innovationen.“ In *Evaluation von Studium Und Lehre*, Hrsg. T D. Großmann und T. Wolbring. Wiesbaden: Springer.
- Boase, J., und R. Ling. 2013. „Measuring Mobile Phone Use: Self-Report versus Log Data.“ *Journal of Computer- Mediated Communication* 18: 508–519.  
<https://doi.org/doi:10.1111/jcc4.12021>.
- Croissant, Y., und G.. Millo. 2008. Panel data econometrics in R: The plm package. *Journal of Statistical Software* 27 (2): 1–43.
- Drachsler, H., und W. Greller. 2016. Privacy and analytics — it’s a DELICATE issue: A checklist to establish trusted learning analytics. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge*: 89-96.
- Ericsson, K.A., und H. A. Simon. 1984: *Protocol Analysis: Verbal Reports as Data*. Cambridge: MIT Press.
- Ferguson, R. 2012. „Learning Analytics: Drivers, Developments and Challenges.“ *International Journal of Technology Enhanced Learning* 4 (5/6): 304–17.
- Gilbert, E., A. Conolly, S. Tietz, L. Calderwood, und N. Rose. 2017. „Measuring Young People’s Physical Activity Using Accelerometers in the UK Millennium Cohort Study.“ *Centre for Longitudinal Studies Working Paper*.
- Golder, S. A., und M. W. Macy. 2014. „Digital Footprints: Opportunities and Challenges for Online Social Research.“ *Annual Review of Sociology* 40 (1): 129–52.  
<https://doi.org/10.1146/annurev-soc-071913-043145>
- Groves, R.M. 2011. „Three Eras of Survey Research.“ *Public Opinion Quarterly* 75 (5): 75(5), 861–71.
- Groves, R.M., Floyd J. Fowler, M. Couper, J.M. Lepkowski, E. Singer, und R. Tourangeau. 2009. *Survey Methodology*. 2nd ed. Hoboken: Wiley.
- Japac, L., F. Kreuter, M. Berg, P. Biemer, P. Decker, C. Lampe, J. Lane, C. O’Neil, und A. Usher. 2015. „Big Data in Survey Research. AAPOR Task Force Report.“ *Public Opinion Quarterly* 79 (4): 839–880.
- Järvelä, S., und K. A. Renninger. 2014. Designing for learning: Interest, motivation, and engagement. In *The Cambridge Handbook of the learning sciences*, Hrsg. R. K. Sawyer. New York: Cambridge University Press.
- Junco, R. 2013. „Comparing Actual and Self-Reported Measures of Facebook Use.“ *Computers in Human Behavior* 29: 626–631.

- Kreuter, F., Keusch, F., Samoiloa, E., and K. Frößinger. 2018. „International Program in Survey and Data Science.” In *Big Data. Chancen, Risiken, Entwicklungstendenzen*, Hrsg. C. König, J. Schröder, und E. Wiegand. Wiesbaden: Springer 27-41.
- Krosnick, Jon A. 1991. “Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys.” *Applied Cognitive Psychology* 5: 213–236.
- Long, P., und G. Siemens. 2011. “Penetrating the Fog: Analytics in Learning and Education.” *EDUCAUSE Review* 46 (5): 31–40.
- Massey, Douglas S., und R. Tourangeau. n.d. “Introduction: New Challenges to Social Measurement.” *The ANNALS of the American Academy of Political and Social Science* 645 (1): 6–22.
- Ohly, S., Sonnentag, S., Niessen, C., und D. Zapf. 2010. „Diary studies in organizational research.” *Journal of Personnel Psychology*. 9(2):79–93
- R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Samoiloa, E., Keusch, F., und F. Kreuter. (2018). Integrating survey and learning analytics data for a better understanding of engagement in MOOCs. In *Data Analytics and Psychometrics: Informing Assessment Practices*, Hrsg. H. Jiao, R.W. Lissitz, und A. Van Wie. Charlotte, NC: Information Age Publishing, 247-261.
- Samoiloa, Evgenia, Florian Keusch, und T. Wolbring. 2017. “Learning Analytics and Survey Data Integration in Workload Research.” *Zeitschrift Für Hochschulentwicklung* 12 (1): 65–78.
- Stier, S., Breuer, J., Siegers, P., und K. Thorson, Hrsg. “Integrating Survey Data and Digital Trace Data.” Special issue, *Social Science Computer Review* (2019).
- Tourangeau, R. 1984. Cognitive science and survey methods. In *Cognitive aspects of survey methodology: Building a bridge between disciplines*, Hrsg. T. Jabine, M. Straf, J. Tanur, und R. Tourangeau, Washington, DC: National Academy Press, 73 - 100.
- Tourangeau, R., L. J. Rips, und K. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge: Cambridge University Press.
- Tourangeau, R., und T. Yan. 2007. “Sensitive Questions in Surveys.” *Psychological Bulletin* 133: 859–83.
- Wang, R., G.M. Harari, P. Hao, X. Zhou, und A. Campbell. 2015. “Smart GPA: How Smartphones Can Assess and Predict Academic Performance of College Students.” In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Hrsg. K. Mase, M. Langheinrich, D. Gatica-Perez, H. Gellersen, T. Choudhury, und K. Yatani, 295–306. New York: Association for Computing Machinery.
- Webb, E., Campbell, D., Schwartz, R., und L. Sechrest. 2000. *Unobtrusive measures*, revised edition. Thousand Oaks, CA: SAGE.

- Willis, G. B. 1999. *Cognitive interviewing: A "How To" guide*. Research Triangle Institute. Meeting of the American Statistical Association. Research Triangle Park, NC: Research Triangle Institute.
- Willis, G. B. 2005. *Cognitive Interviewing. A Tool for Improving Questionnaire Design*. Thousand Oaks: SAGE.
- Willis, G. B. 2015. *Analysis of the Cognitive Interview in Questionnaire Design*. Oxford: Oxford University Press

## Appendix



**Appendix 1:** Boxplots des Workloads in Minuten über Wochen in den drei Kursen

**Appendix 2.** Lineare Regressionen zum studentischen Workload (gemessen via LA):  
Beobachtungen mit LA weniger als fünf Minuten sind entfernt

Modell	Random-Effects-Modell		Fixed-Effects-Modell	
	Koeffizient	95% CI	Koeffizient	95% CI
Umfragedaten	0.10***	0.04, 0.16	0.15***	0.08, 0.21
Konstante	71.02***	53.59, 88.23	-	-
R <sup>2</sup> ( <i>marginal</i> für Random- Effects- Modell)	0.09		0.9	
Log Likelihood für Random- Effects- Modell/	-1530.97		21.11***	

F für Fixed-Effects-Modell		
n	307	307

Signifikanzniveaus: \*\*\*  $p \leq 0.01$  \*\*  $p \leq 0.05$  \*  $p \leq 0.1$