Literale Kompetenzen in empirischen Disziplinen

Erprobung einer Methode zur Messung des Lernzuwachses in schreibintensiven Praktika der Biologie

Nils Cordes, Anne-Kathrin Warzecha

Kurzfassung

Eine Herausforderung in der Lehre besteht darin, herauszufinden, ob Studierende durch den Besuch einer Veranstaltung im gewünschten Ausmaß etwas lernen. Evaluationen können nur zum Teil Antworten darauf liefern, da sie oft Momentaufnahmen darstellen, die von vielen Faktoren beeinflusst werden. Daher wird außerdem eine Methode benötigt, um den Lernzuwachs zu messen. Dies gilt vor allem für Kompetenzen, die üblicherweise über mehrere Semester hinweg erworben werden – wie die Fertigkeit, wissenschaftlich zu schreiben. Im Rahmen dieser Studie wird die Veränderung von Selbsteinschätzungen genutzt, um den Lernzuwachs (im Sinne von „normalized gain" nach Hake) bei der Entwicklung literaler Kompetenzen im Biologiestudium zu messen. Dafür gaben Studierende über einen Zeitraum von einem Jahr zu drei verschiedenen Zeitpunkten Selbsteinschätzungen zu sechs verschiedenen Kompetenzen ab, die als zentral für das Verfassen wissenschaftlicher Arbeiten in empirischen Disziplinen erachtet werden. Dieselben Kompetenzen wurden an zwei Zeitpunkten anhand der von den Studierenden angefertigten Texte bewertet. Wir verglichen die Veränderungen in der Qualität der Texte anschließend mit den Veränderungen in den Selbsteinschätzungen der Studierenden, um herauszufinden, ob aus der Veränderung von Selbsteinschätzungen auf tatsächliche Entwicklungen geschlossen werden kann. Die Veränderungen in den Selbsteinschätzungen und die Veränderungen in den Bewertungen der Texte korrelieren stark. Die große Varianz in den Daten suggeriert jedoch, dass die Methode zur Einschätzung individueller Lernprozesse ungeeignet ist. Auf Kursebene hingegen stellt der so gemessene Lernzuwachs unserer Meinung nach eine Bereicherung für die Lehrevaluation dar und kann konkrete Ansatzpunkte für die Weiterentwicklung von (schreibintensiven) Lehrveranstaltungen liefern.

Schlagworte: Evaluation; Selbsteinschätzung; Lernerfolg; Lernziel; Curriculumsentwicklung

Abstract

A challenge of teaching is to know if students actually learn in a course what they are supposed to learn. Evaluations can only partly answer this question as they are but snapshots in time, influenced by many different factors. Therefore, a method is needed to measure how much students learn, the so-called learning gain. This is especially true for competencies that are acquired over several semesters – like the ability to write scientifically. In this study, we use the change over time in self-assessments to measure learning gain (following Hake's definition of „normalized gain") in writing competencies in a Biology curriculum. Over the course of one year, students filled out a questionnaire at three separate points in time to self-assess how good they are in six competencies needed for writing scientific texts. We also measured these competencies based on the scientific texts written by students at two different points in time during our study. We then compared learning gain based on self-assessment with learning gain based on the texts to find out if change in self-assessments can be used to learn about actual writing developments within courses. Both correlated significantly. The large variance, however, suggests that the method is unfit to illustrate individual learning processes and should only be used to measure course-level developments, where we think it is a valuable addition to standard evaluations. Estimated learning gains can provide specific starting points for further conceptual advancements of (writing-intensive) courses.

1 Einleitung

Das Rückgrat eines jeden naturwissenschaftlichen Studiums muss die Fähigkeit sein, wissenschaftlich zu denken, zu analysieren und zu kommunizieren. Der Bachelor of Science sollte diese Kompetenzen vermitteln, weshalb oft bereits in den ersten Semestern Schreibaufträge feste Bestandteile des Studiums sind. Das Formulieren und Begründen von Fragestellungen, das Auswerten und Interpretieren von Daten, das Erstellen von Diagrammen und der Umgang mit wissenschaftlicher Literatur sind wesentliche Bestandteile von Forschung, ohne die es kaum möglich ist, theoretische Fachinhalte zu bewerten und anzuwenden (Bao, Cai, Koenig et al. 2009; Coil, Wenderoth, Cunningham & Dirks 2010).

Evaluationen und Klausuren sollen in der Regel überprüfen, inwieweit Studierende Fachwissen und Kompetenzen zu einem bestimmten Zeitpunkt besitzen. Weil literale Kompetenzen aber auf dem Weg zur Abschlussarbeit über mehrere Semester hinweg erworben und gefestigt werden, ist es hilfreich, den *Lernzuwachs* (learning gain) beurteilen zu können (Caspersen, Smeby & Aamodt 2017) – ein Maß, das bislang weder von Klausuren noch von Evaluationen adäquat aufgenommen werden kann. Basierend auf Cohens Maß der Effektgröße (*effect size*, Cohen 1977) entwickelte Hake (1998) eine Methode, die das Verständnis für konkrete Konzepte der Physik mit Hilfe von Prä-Post-Tests bestimmt. Der sogenannte *normalized gain* teilt den tatsächlichen Lernzuwachs durch den maximal möglichen Lernzuwachs. Diese Standardisierung

soll helfen, auch unterschiedliche Lernziele miteinander zu vergleichen. Darauf aufbauend entwickelten sich ähnliche Konzepte wie *normalized change* (Marx & Cummings 2007), *individual gain* (Raupach, Münscher, Beißbarth et al. 2011; Schiekirka, Reinhardt, Bei et al. 2013) und *averaged individual gain* (Nissen, Talbot, Thompson & Dusen 2018; Von Korff, Archibeque, Gomez et al. 2016).

In der Physik lassen sich Fachwissen und Fertigkeiten gut mit Konzeptinventaren (Savinainen & Scott 2002) messen. Komplexe, disziplinübergreifende Kompetenzen wie das wissenschaftliche Schreiben, Denken und Argumentieren waren bislang aber nur schwer zu quantifizieren (Caspersen, Smeby & Aamodt 2017; Lauer & Hendrix 2009; Timmerman, Strickland, Johnson et al. 2011). Ohne ein möglichst objektives Maß für die Kenntnisse von Studierenden zu einem bestimmten Zeitpunkt scheint es unrealistisch, den Lernzuwachs bestimmen zu können. *Rubrics* (Anson & Dannels 2002; Timmerman, Strickland, Johnson et al. 2011) sind eine Methode, um Kompetenzen auf spezifische, messbare Kriterien herunterzubrechen. Doch aufgrund ihres großen Zeitaufwandes wäre es wünschenswert, Lernzuwachs nicht nur über die Bewertung von Texten, sondern auch im Rahmen von regelmäßigen Kursevaluationen messen zu können (Mager 1984).

Für die Messung des Lernzuwachses bei einfach zu bewertenden Kernkompetenzen des Medizinstudiums übertrugen Raupach, Münscher, Beißbarth et al. (2011) Hakes Formel von Prüfungen auf Selbsteinschätzungen und zeigten, dass der Lernzuwachs zuverlässig sowohl an konkreten Fragestellungen in einer Umfrage gemessen werden kann als auch an der Leistung bei praktischen Prüfungen. Wir fragten uns, ob diese Form der Lernzuwachsbestimmung auch auf das wissenschaftliche Schreiben und Auswerten übertragen werden kann. Das Ziel dieser Studie ist es daher, am Beispiel von Studiengängen in der Biologie zu überprüfen, ob der Lernzuwachs basierend auf Selbsteinschätzungen zuverlässig den tatsächlichen Lernzuwachs beim Anfertigen von Versuchsprotokollen/Lab Reports (Cargill & O'Connor 2013; Cordes 2016; Lerner 2007; Rosenthal 1987) wiedergibt und ob diese Methode in den empirischen Wissenschaften genutzt werden könnte, um den Erfolg von Kursen zu evaluieren.

2 Methoden

Die Studie lief über die ersten zwei Fachsemester von Studienanfänger*innen der Biologie an der Universität Bielefeld. Das Wintersemester begann im Oktober 2018, das Sommersemester endete im Juli 2019. Alle an der Studie teilnehmenden Studierenden nahmen an den verpflichtenden Praktika des Basismoduls Praxis I und II teil. In diesen Praktika ließen wir die Studierenden an drei verschiedenen Zeitpunkten (T1, T2, T3) eine Umfrage ausfüllen, in der sie ihre Fähigkeiten zum Anfertigen wissenschaftlicher Texte einschätzen sollten. Außerdem verfassten die Studierenden für die Praktika wissenschaftliche Texte in Form von Versuchsprotokollen. Die Studierenden erhielten für das Winter- und das Sommersemester jeweils eines von vier bzw. fünf Themen, zu denen sie im Laufe des Semesters das Versuchsprotokoll schrieben.

Die Verteilung der Themen verlief zufällig und zielte darauf ab, dass alle Themen gleichmäßig auf die Studierenden verteilt waren. Die Studierenden wurden gebeten, die Versuchsprotokolle der einzelnen Semester (P1 und P2) anonymisiert auf einen Universitätsserver zu laden, damit diese für eine Textanalyse von drei Gutachter*innen bewertet werden konnten. So ergaben sich fünf definierte Zeitpunkte, an denen die Studierenden für die Studie unterschiedliche Leistungen erbringen mussten (Abb. 1).

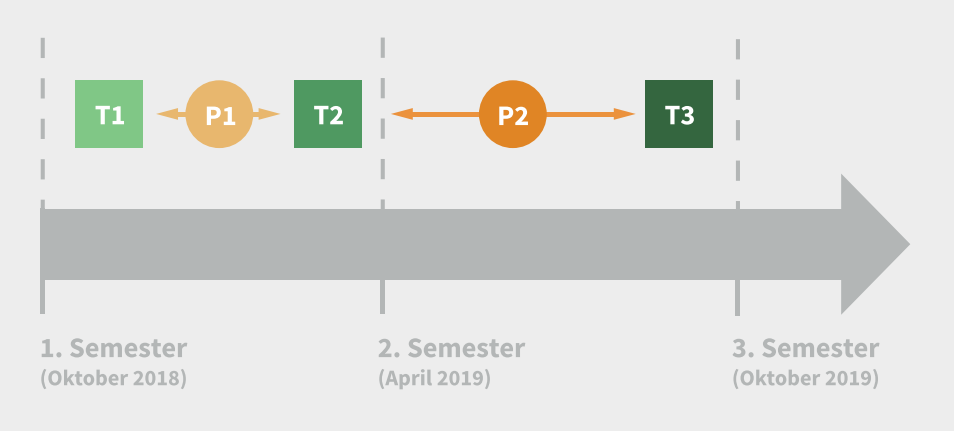

Abbildung 1: Ablauf der Studie. Umfragen wurden zu drei verschiedenen Zeitpunkten im Studium durchgeführt, T1-T3. Dazwischen wurden von den Studierenden Versuchsprotokolle geschrieben, und zwar in den Zeiträumen P1 und P2.

Umfragen

Die Umfragen wurden zu Beginn (T1) und am Ende (T2) des ersten Semesters sowie am Ende des zweiten Semesters (T3) jeweils in den Praktika am Smartphone durchgeführt (Abb. 1). Die Online-Fragebögen dafür wurden mit der Software EvaSys® (Electric Paper, Lüneburg, Deutschland) generiert und die Ergebnisse nach Ablauf der Umfrage in Tabellenform gespeichert. Der Zugang zur Umfrage war nicht passwortgeschützt, wurde aber nach Ablauf der Erhebungsphase geschlossen. Einmal abgeschickte Fragebögen konnten von den Studierenden nicht überarbeitet werden.

Die einzelnen Fragebögen wurden durch einen Selbstgenerierten Identifikationscode (SGIC) anonym immer denselben Studierenden zugeordnet. Jede Umfrage bestand aus Selbsteinschätzungen zu sechs Fähigkeiten des wissenschaftlichen Schreibens plus einer Positivkontrolle, Lernziel 0, von dem wir erwarteten, dass alle Studierenden es erreichten. Die Lernziele wurden als operationalisierte Lernziele (Doran 1981) formuliert (Tab. 1). Die Studierenden gaben ihre Selbsteinschätzung auf einer sechsstufigen Skala von 1 (trifft vollständig zu) bis 6 (trifft gar nicht zu) an.

Tabelle 1: Zusammenstellung der operationalisierten Lernziele, dem SMART-Prinzip (Doran 1981) folgend so formuliert, dass die Studierenden sie eindeutig einschätzen können. Lernziel 0 diente als Positivkontrolle. Gutachter*innen bewerteten anhand der Versuchsprotokolle, inwieweit ein Lernziel erreicht wurde, indem sie einzelne Fragen dazu beantworteten.

	Operationalisiertes Lernziel	**Abkürzung im Rahmen dieser Studie**	**Fragen zur Hilfestellung für Gutachter*innen bei der Bewertung der Versuchsprotokolle**
0	*Ich kann alle Abschnitte eines Versuchsprotokolls benennen.*	Abschnitte	• Sind alle 4 Hauptabschnitte vorhanden? • Sind alle Abschnitte korrekt benannt? • Ist ein Literaturverzeichnis vorhanden? • Ist ein Abstract vorhanden?
1	*Ich kann die Funktionen der einzelnen Abschnitte eines Versuchsprotokolls im Detail erklären.*	Funktionen	• Ist das Versuchsziel klar und verständlich? • Sind alle Abbildungen/Tabellen im Text in Worten beschrieben? • Greift die Diskussion das Versuchsziel auf? Wird die Forschungsfrage beantwortet?
2	*Ich kann wissenschaftliche Versuchsziele (z. B. Forschungsfragen oder Hypothesen) sicher formulieren.*	Versuchsziele	• Wird ein Versuchsziel erwähnt? • Ist das Versuchsziel klar und verständlich? • Werden Ergebnisse visuell dargestellt?
3	*Ich kann aus experimentell gewonnenen Daten ein aussagekräftiges Diagramm erstellen.*	Diagramme	• Greift die Diskussion das Versuchsziel auf? Wird die Forschungsfrage beantwortet? • Werden die Daten der Abbildungen/Tabellen in der Diskussion aufgegriffen und interpretiert?
4	*Ich kann im Detail erklären, wie eine Abbildung (z. B. ein Diagramm) zur Beantwortung einer Forschungsfrage beiträgt.*	Beantwortung der Frage	• Sind alle Abbildungen/Tabellen im Text in Worten beschrieben? • Werden die Daten der Abbildungen/Tabellen in der Diskussion aufgegriffen und interpretiert?
5	*Ich kann abschätzen, wie sich Fehler, die beim Experimentieren passiert sind, auf die Messergebnisse auswirken.*	Fehlerdiskussion	• Wurden Fehler diskutiert?
6	*Ich kann wissenschaftliche Texte als Referenzen in mein Versuchsprotokoll einbinden.*	Referenzen	• Werden Referenzen einheitlich und korrekt genutzt? • Wie sind die Referenzen in den Text integriert? • Welche Arten von Referenzen werden bereits genutzt?

Bewertung von Versuchsprotokollen

Alle Versuchsprotokolle wurden von drei Gutachter*innen gelesen: zwei studentischen Hilfskräften aus dem 6. Semester, die sich über ihr Studium intensiv mit dem Schreiben wissenschaftlicher Texte auseinandergesetzt hatten, und einem wissenschaftlichen Mitarbeiter, der als Schreibberater der Fakultät tätig ist. Deren Aufgabe war es, möglichst gut einzuschätzen, inwieweit die Lernziele erreicht wurden.

Zur Bewertung der Versuchsprotokolle erhielten die Gutachter*innen einzelne, den Lernzielen zugeordnete Fragen (Tab. 1), die mit einer einfachen Rubric (Anson & Dannels 2002; Timmerman, Strickland, Johnson et al. 2011) zu beantworten waren. Vor Beginn der Bewertung wurden die Fragen und die Rubric an sechs Übungsversuchsprotokollen getestet und darauf optimiert, dass alle Gutachter*innen eine möglichst gleiche Vorstellung der Fragen hatten. Die separaten Bewertungen der drei Gutachter*innen wurden für die weitere Analyse gemittelt, für jedes Lernziel summiert und auf die Skala der Umfrage übertragen, von 1 (trifft vollständig zu) bis 6 (trifft gar nicht zu).

Auswertung

In die Berechnung des Lernzuwachses gingen ausschließlich Studierende ein, die mindestens zwei Fragebögen vollständig ausgefüllt hatten und deren SGIC eindeutig zugeordnet werden konnte (vgl. Tab. 2 im Ergebnisteil).

Die statistische Auswertung erfolgte mit der Software R (R Core Team 2016). Um die Zuverlässigkeit der Bewertungen der Versuchsprotokolle zu bestimmen, berechneten wir den Intraclass Correlation Coefficient (ICC) (Gamer, Lemon, Fellows & Singh 2019). Wir analysierten die Varianz mit Hilfe eines linearen Modells und des lme4-Pakets, Befehl „lmer“ (Bates, Maechler, Bolker & Walker 2015).

Die Berechnung des Lernzuwachses für die Lernziele folgte der Formel für „individual gain (%)“ aus Raupach, Münscher, Beißbarth et al. (2011). Wir berechneten für alle Studierenden individuell den Lernzuwachs, und zwar sowohl für die Selbsteinschätzung als auch für die Bewertung von Versuchsprotokollen. In beiden Fällen galt dieselbe Formel, jeweils bestehend aus einem zeitlich früheren Wert (prä) und einem späteren Wert (post), um die zeitliche Veränderung über den Semesterverlauf als Prozentsatz zwischen –100 % und + 100 % angeben zu können. Je nachdem, ob die Veränderung über diesen Zeitraum positiv oder negativ war, wurden verschiedene Formeln benutzt. Wenn $prä - post < 0$, galt

$$Lernzuwachs\left(\%\right) = \frac{prä - post}{prä - 1} \times 100$$

Wenn $prä - post > 0$, galt

$$Lernzuwachs\left(\%\right) = \frac{prä - post}{6 - prä} \times 100$$

Wenn sich die Selbsteinschätzung bzw. Bewertung nicht veränderte, war der Lernzuwachs 0.

Für die Darstellung der Selbsteinschätzung (Abb. 2) und der Bewertung der Versuchsprotokolle (Abb. 3) im Ergebnisteil rechneten wir alle aufgenommenen Daten zur besseren Veranschaulichung in die Skala 0 (sehr schlecht) bis 1 (sehr gut) um. Damit entspricht die Skala besser der des berechneten Lernzuwachses (Abb. 4, Abb. 5).

3 Ergebnisse und Diskussion

Im Wintersemester 2018/19 waren für die Praktika, in denen die Befragung stattfand, 333 Studierende eingeschrieben. Nach Entfernen der unvollständigen Datensätze und solcher, die nicht über den SGIC zugeordnet werden konnten, blieben die in Tabelle 2 aufgeführten Stichprobengrößen übrig.

Tabelle 2: Stichprobengrößen der in die Studie eingeflossenen Datensätze. Nur ein Teil der Studierenden, die sich an der Umfrage beteiligt haben, reichte auch Versuchsprotokolle zur Bewertung ein. 22 Studierende nahmen an allen Umfragen teil und gaben jeweils zwei Versuchsprotokolle ab; diese stellen die Grundlage für die Auswertung von Abbildung 5 dar.

Stichprobe Umfrage	**N**
Erster Fragebogen (T1)	176
Zweiter Fragebogen (T2)	91
Dritter Fragebogen (T3)	176
Vergleich T1 und T2	*90*
Vergleich T2 und T3	*63*
Vergleich T1 und T3	*81*
Stichprobe Versuchsprotokolle	**N**
Versuchsprotokoll 1 (P1)	40
Versuchsprotokoll 2 (P2)	22
Vergleich P1 und P2	*22*

Selbsteinschätzungen der Studierenden

Die Studierenden bewerteten sich grundsätzlich zum Zeitpunkt T1 (Beginn erstes Semester) schlechter als zu den späteren Zeitpunkten T2 und T3 (Ende von Semester 1 bzw. 2) (Abb. 2). Zu Beginn (T1) bewerteten sie sich am besten dabei, aus ihren Daten Diagramme erstellen zu können (Lernziel „Diagramme“). Nach zwei Semestern (Zeitpunkt T3) pendelte sich die Selbsteinschätzung bei allen Lernzielen bis auf „Abschnitte“ im Mittel über alle Studierenden auf Werte zwischen 0,6 und 0,7 ein (Abb. 2). Bei der Kontrolle „Abschnitte“ erwarteten wir, dass annähernd alle Studierenden zwischen den Zeitpunkten T1 und T2 die Fähigkeit, alle vier Abschnitte eines Versuchs-

protokolls zu benennen, erwerben würden. Überraschenderweise schätzten sich die Studierenden hier im Mittel auch zu den Zeitpunkten T2 und T3 noch schlechter ein. 66 von 83 Studierenden gaben sich allerdings zum Zeitpunkt T3 die Höchstbewertung von 1.

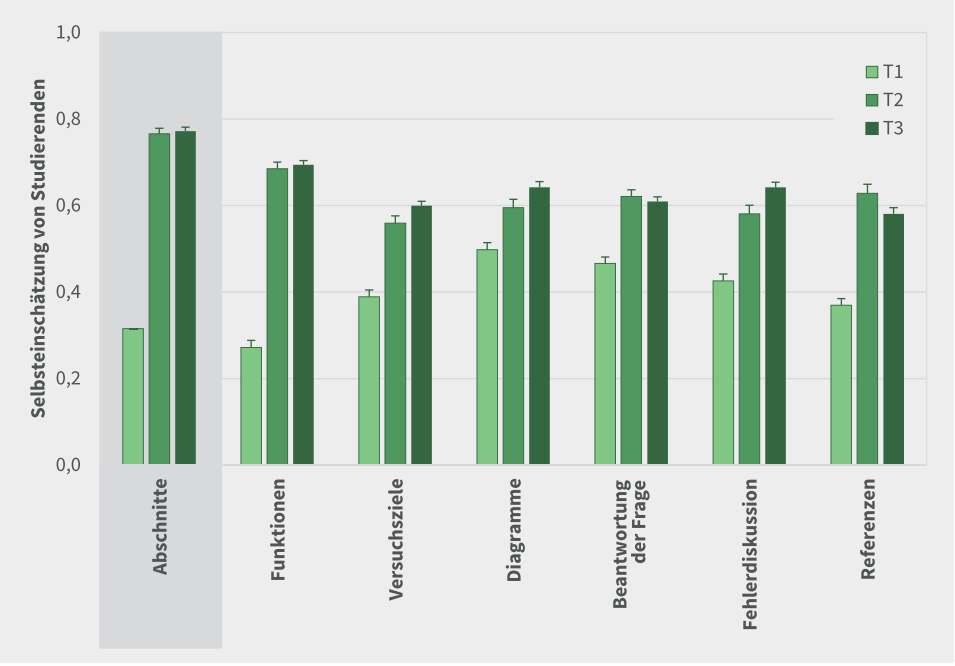

Abbildung 2: Mittlere Selbsteinschätzung der Studierenden zu den sechs Lernzielen und der Positivkontrolle (siehe Methoden), auf 1 normiert (1 = sehr gut, 0 = sehr schlecht) für die Zeitpunkte T1 bis T3. Die Fehlerbalken kennzeichnen den Standardfehler des Mittelwertes, $N_{T1} = 176$, $N_{T2} = 91$, $N_{T3} = 176$.

Bewertungen der Versuchsprotokolle

Zur Bewertung der Übereinstimmung zwischen den einzelnen Gutachter*innen nutzten wir den Intraclass Correlation Coefficient (ICC). Dieser gab eine Übereinstimmung zwischen den Gutachter*innen von 0,678 an (95 %-Konfidenzintervall: 0,649–0,707), was als moderat zuverlässig interpretiert werden sollte (Koo & Li 2016).

Im Gegensatz zu den deutlichen Unterschieden in den Selbsteinschätzungen zwischen Zeitpunkt T1 und T2 wurden bei den Bewertungen der Versuchsprotokolle im Mittel nur wenige Unterschiede zwischen erstem und zweitem Fachsemester gefunden (Abb. 3). Es gab minimale Veränderungen, jedoch sowohl in positiver als auch in negativer Richtung. Die Kontrolle „Abschnitte“ zeigte, dass beim Schreiben der Versuchsprotokolle fast alle Studierenden die richtigen Begriffe benutzten (Abb. 3), ein leichter Widerspruch zu den relativ vorsichtigen Selbsteinschätzungen dazu. Gerade mal 5 der 62 bewerteten Versuchsprotokolle aus beiden Fachsemestern wiesen kleine Fehler in der Benennung der Abschnitte auf. Nur 21 der 62 Versuchsprotokolle enthielten einen Abstract, was aber nicht überrascht, da ein Abstract zwar wünschens-

wert war, aber in den Richtlinien zum Schreiben im ersten Semester noch nicht als notwendig kommuniziert wurde.

Überraschend war, dass Abbildungen größtenteils korrekt und ausführlich beschriftet wurden („Diagramme“), sogar schon im ersten Semester, und dass auf diese Abbildungen in der Diskussion Bezug genommen wurde („Beantwortung der Frage“). Die größten Schwierigkeiten hatten Studierende beim Abschätzen, welche Auswirkungen Fehler auf die Ergebnisse hatten („Fehlerdiskussion“). Das Lernziel „Referenzen“ wurde aus unserer Sicht überraschend positiv bewertet, was aber in erster Linie darauf zurückzuführen war, dass wir die Anzahl der benutzten Quellen nicht berücksichtigten. In der Regel gaben die Studierenden einige wenige Quellen an, konnten diese aber in der Mehrzahl richtig in den Text integrieren (Abb. 3). In einem Drittel der Versuchsprotokolle tauchten bei den Referenzen auch schon einzelne gute wissenschaftliche Primärquellen auf.

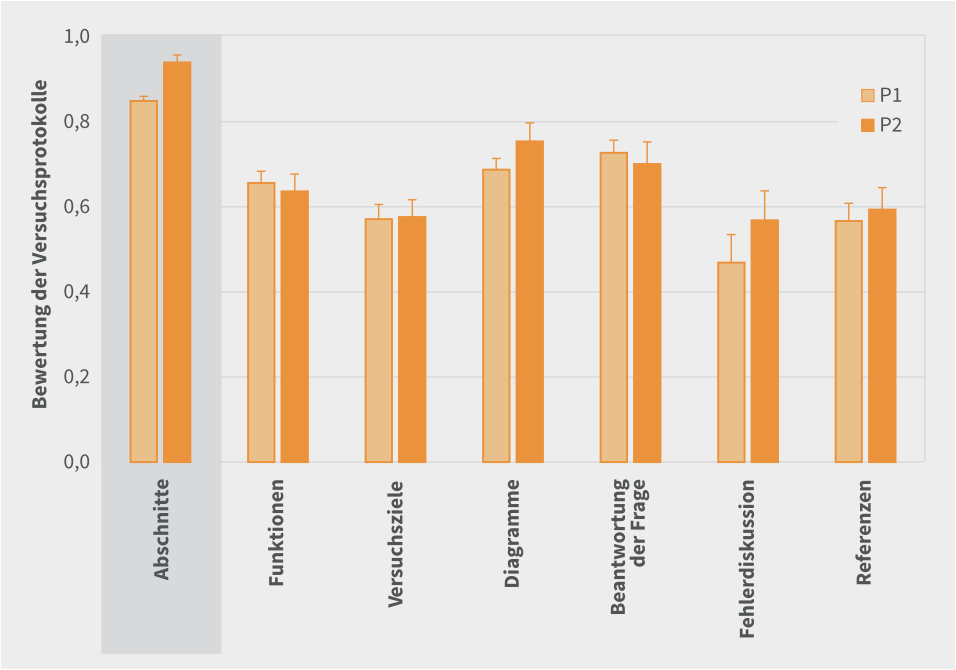

Abbildung 3: Mittlere Bewertung der Versuchsprotokolle zu den Lernzielen (siehe Methoden), auf 1 normiert (1 = sehr gut, 0 = sehr schlecht) für Zeitraum P1 (1. Semester, helle Säule, N = 40) und Zeitraum P2 (2. Semester, dunkle Säule, N = 22). Die Fehlerbalken kennzeichnen den Standardfehler des Mittelwertes.

Lernzuwachs

Sowohl Selbsteinschätzungen als auch Bewertungen sind davon geprägt, dass sie subjektiv und abhängig von einer Vielzahl von Faktoren sind (Wachtel 1998), die als unerklärte Varianz in die Messwerte einfließen. Nimmt man bei den Bewertungen der Versuchsprotokolle dieser Studie typische Faktoren wie die unterschiedlichen Themen (4,3 %), die Zeitpunkte (0,1 %) oder die Studierenden (20,1 %) heraus, bleibt auch

in unseren Daten immer noch ein Anteil von 76 % an unerklärter Varianz übrig (gemischtes lineares Modell).

Das Maß des Lernzuwachses sollte einen Teil dieser Varianz auffangen können (Raupach & Schiekirka 2017). Der Lernzuwachs bei den Selbsteinschätzungen war, wie Abbildung 2 bereits suggeriert, zwischen den Zeitpunkten T1 und T2 (1. Fachsemester) deutlich größer als zwischen T2 und T3 (2. Fachsemester). Basierend auf den Selbsteinschätzungen lernten die Studierenden demnach in allen sechs Kriterien über die ersten zwei Fachsemester dazu, allerdings in unterschiedlich starkem Maße (Abb. 4).

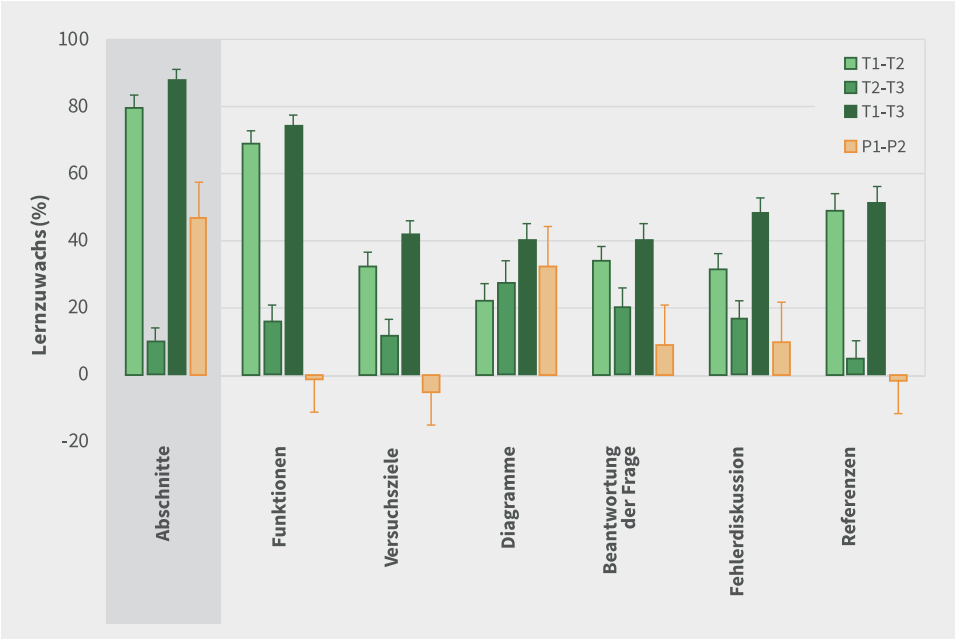

Abbildung 4: Mittlerer Lernzuwachs (in %) für alle Lernziele basierend auf Selbsteinschätzungen (grüne Säulen) und den Bewertungen von Versuchsprotokollen (orange Säulen, siehe Methoden). Untersuchte Zeiträume für die Selbsteinschätzungen waren T1-T2, T2-T3 sowie der gesamte Zeitraum der Studie (die ersten zwei Fachsemester, T1-T3); $N_{T1\text{-}T2} = 90$, $N_{T2\text{-}T3} = 63$, $N_{T1\text{-}T3} = 81$; Lernzuwachs bei den Versuchsprotokollen fand über den Zeitraum P1-P2 statt, N = 22. Die Fehlerbalken kennzeichnen den Standardfehler des Mittelwertes. Siehe Methoden für eine genauere Beschreibung der Zeiträume.

Im zweiten Fachsemester war dieser Effekt erheblich geringer als im ersten. Dies ist auf eine Schwierigkeit bei der Interpretation zurückzuführen, die in Abbildung 4 gut deutlich wird: Das wie erwartet stärkste Lernziel „Abschnitte" zeigt, dass die Verbesserung der befragten Studierenden (N = 90) im Mittel 80 % betrug. Das bedeutet, dass von der noch möglichen Verbesserung nach den Selbsteinschätzungen zum Zeitpunkt T1 80 % zum Zeitpunkt T2 erreicht wurden. Hätten sich alle Studierenden zum Zeitpunkt T2 als perfekt eingeschätzt, läge dieser Anstieg direkt bei 100 %, und zwar unabhängig davon, ob sich jemand zum Zeitpunkt T1 als sehr schlecht (0) oder gut

(80) einschätzte. Was hier verloren geht, ist die Information, dass 62 Studierende (von 90) sich zum Zeitpunkt T2 als perfekt einschätzten. Die geringe Verbesserung zwischen T2 und T3 kommt deshalb daher, dass nur noch 28 Studierende überhaupt eine Möglichkeit zur Verbesserung basierend auf den vorherigen Selbsteinschätzungen hatten. Dies zieht den Mittelwert zwangsläufig nach unten.

Ein weiterer Grund, warum der Lernzuwachs in der Regel zwischen T1 und T2 höher war als zwischen T2 und T3, ist, dass viele dieser Lernziele Fähigkeiten repräsentieren, die Studierende kaum aus der Schule kennen. Die Möglichkeit, in den ersten Wochen des Studiums viel zu lernen, ist deutlich größer als zu späteren Zeitpunkten, vor allem bezogen auf die von uns festgelegten Kriterien, welche die Lernziele der ersten zwei Semester darstellen.

Um den Lernzuwachs der beiden Datenaufnahmen zu vergleichen, müssen wir vergleichbare Zeiträume wählen. Bei einem Schreibauftrag kann man nicht wie bei Klausuren davon ausgehen, dass man einen Wissensstand zu einem definierten Zeitpunkt überprüft, denn der Schreibprozess selbst trägt erheblich dazu bei, dass Studierende lernen (Anderson, Gonyea, Anson et al. 2015). Dies trifft vor allem für unser Kontroll-Lernziel zu, da die Fähigkeit, alle Abschnitte zu benennen und im Text richtig zu nutzen, erst beim Schreiben des Textes (spätestens bei der letzten Kontrolle vor der Abgabe) erworben wird. Die überprüften Fähigkeiten zum Zeitpunkt P1 sollten demnach nicht das geringe Fachwissen von Studierenden zu Studienbeginn (T1) widerspiegeln, sondern viel näher an den abgefragten Fähigkeiten zum Ende des ersten Semesters (T2) sein. Entsprechend ist der Lernzuwachs von P1 zu P2 für die Kontrolle geringer, als man erwarten würde, weil das Schreiben des Versuchsprotokolls P1 bereits dazu geführt hat, dass fast alle Studierenden das Ziel erreichten (vgl. Abb. 3).

Um den Lernzuwachs der beiden Berechnungen zu vergleichen, müsste man demnach die Zeiträume T2-T3 und P1-P2 heranziehen. Schaut man sich alle Studierenden an, die sich sowohl an den Umfragen T2-T3 beteiligt haben als auch alle Versuchsprotokolle verfasst und eingereicht haben, korreliert der Lernzuwachs von T2-T3 und P1-P2 im Kursmittel tatsächlich sehr stark ($r = 0{,}86$, Abb. 5). Die Verschiebung der Regressionsgerade entlang der x-Achse besagt jedoch, dass der Lernzuwachs basierend auf den Selbsteinschätzungen grundsätzlich größer ist als der aus den Bewertungen. Wir interpretieren dies so, dass wir bei den Bewertungen einiger Kriterien kritischer waren als die Studierenden bei ihren Selbsteinschätzungen.

Aus Abbildung 5 geht außerdem hervor, dass die Studierenden im Anfertigen von Diagrammen in den Basismodulen relativ starke Fortschritte gemacht haben, während das Formulieren von Versuchszielen in der kleinen Stichprobe eher schlechter wurde. Tatsächlich waren beide Aspekte Lernziele des Basismoduls 2 im Sommersemester. Man könnte daraus schließen, dass die Studierenden eines der beiden Lernziele besser erreichten als das andere. Was verursacht diese Unterschiede? Wie stehen diese beiden Lernziele im Zusammenhang mit dem, was im ersten Semester thematisiert wurde? Welche anderen Aspekte haben geholfen oder waren hinderlich dabei, das eine Lernziel stärker zu erreichen als das andere? Dies sind die Fragen, die man nach einer Kursevaluation stellen sollte, die jedoch ohne Kenntnis des Lernzuwachses

kaum zu beantworten sind. Das Maß des Lernzuwachses hilft also, Diskussionen vom Kursablauf, den Rahmenbedingungen und den Dozierenden wegzulenken und Fragen zu Lernzielen und Lernerfolg zu stellen.

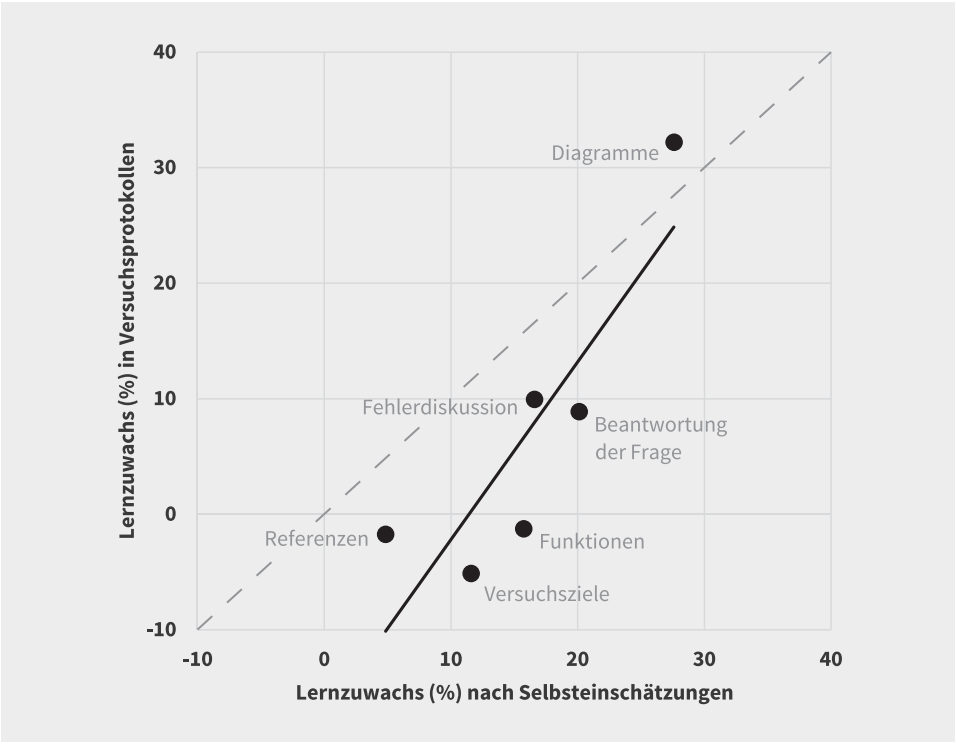

Abbildung 5: Vergleich des mittleren Lernzuwachses (in %) für die sechs untersuchten Lernziele basierend auf Selbsteinschätzungen zu den Zeitpunkten T2-T3 (x-Achse) und Bewertung der Versuchsprotokolle (y-Achse) (N = 22). Die Mittelwerte korrelieren mit r = 0,86 (p = 0,027). Die gestrichelte Linie stellt die zu erwartende Korrelation dar, wenn der Lernzuwachs nach Selbsteinschätzungen und Versuchsprotokollen identisch wäre.

Unsere Daten zeigen, dass die Methode des Lernzuwachses grundsätzlich einen guten Hinweis darauf gibt, ob Studierende über einen definierten Zeitraum etwas dazulernen oder nicht. Auf Kursebene korrelieren Selbsteinschätzungen und Bewertungen der Versuchsprotokolle, sodass der Lernzuwachs als Maß für den Erfolg eines Kurses ein wertvolles Werkzeug der Evaluation sein kann. Dies deckt sich mit Erfahrungen aus dem Medizinstudium, nach denen objektive Maße und Selbsteinschätzungen auf Kursebene ähnlich stark korrelierten (Schiekirka, Reinhardt, Bei et al. 2013).

Was die Methode nicht leisten kann, ist, den Lernerfolg von Individuen realistisch abzubilden (Lam 2009; Schiekirka, Reinhardt, Bei et al. 2013). Es spielen zu viele Variablen in die Ergebnisse, als dass ein individuelles Ergebnis ein realistisches Bild der Person wiedergeben könnte. Vor allem weichen bei Einzelnen die Selbsteinschät-

zungen zum Teil stark von der Realität ab (Kruger & Dunning 1999; Deslauriers, McCarty, Miller et al. 2019), und wie wir zeigen konnten, sind selbst möglichst objektive Bewertungen bei manchen Lernzielen immer noch von starker Varianz geprägt.

Es ist außerdem wichtig, sich bei der Interpretation der Ergebnisse klarzumachen, dass die Veränderung in der Selbsteinschätzung keinen Hinweis darauf gibt, wie gut die Studierenden tatsächlich bei einem Lernziel sind (Caspersen, Smeby & Aamodt 2017). Es kann schnell passieren, dass man einen Lernzuwachs von 10 % so interpretiert, dass ein Kurslernziel nicht erreicht wurde. In manchen Fällen mag das auch stimmen, doch im Fall dieser Studie (vgl. Abb. 4) ist der erwartet geringe Lernzuwachs für „Abschnitte" in T2-T3 lediglich darauf zurückzuführen, dass die Studierenden im ersten Semester bereits sehr viel dazugelernt hatten (80 %). Man muss den Lernzuwachs also immer im Kontext des Curriculums betrachten.

Der Lernzuwachs, sei er durch Selbsteinschätzungen oder durch Bewertungen gewonnen, ist daher in erster Linie als eine Evaluation des Lernerfolgs einer Kohorte über einen definierten Zeitraum zu sehen. Dieser kann indirekt auf die Effektivität einer Veranstaltung hinweisen und hilfreiche Ansätze dafür liefern, in welche Richtung zukünftige Entwicklungen von Veranstaltungen oder Curricula gehen sollten (Evans, Howson & Forsythe 2018). Inwieweit nun Selbsteinschätzungen dazu beitragen können, den Lernzuwachs auch beim wissenschaftlichen Schreiben zu bestimmen, bleibt mit bislang zu wenigen Untersuchungen eine offene Frage. Wie jede Form von Evaluation kann auch der Lernzuwachs nur Ausschnitte liefern. Doch mit seinem lernziel- und fortschrittsorientierten Blick sind diese Ausschnitte aus unserer Sicht eine Bereicherung für lernzielorientierte Evaluationen, die nötige Diskussionen lostreten und konkrete Ansatzpunkte für die Weiterentwicklung nicht nur von schreibintensiven Lehrveranstaltungen liefern können.

Literatur

Anderson, P., Gonyea, R. M., Anson, C. M., Paine, C. & Students, F. (2015). The Contributions of Writing to Learning and Development: Results from a Large-Scale Multi-Institutional Study. *Research in the Teaching of English*, 50(2), 199–235.

Anson, C. M. & Dannels, D. P. (2002). Developing Rubrics for Instruction and Evaluation. In D. Roen, V. Pantoja, L. Yena, S. K. Miller & E. Wagonner (Hg.), *Strategies for Teaching First-Year Composition*, 387–401. Urbana, Ill.: National Council of Teachers of English.

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J. et al. (2009). Physics: Learning and Scientific Reasoning. *Science*, 323(5914), 586–587.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.

Cargill, M. & O'Connor, P. (2013). *Writing Scientific Research Articles*. Oxford: Wiley-Blackwell.

Caspersen, J., Smeby, J. C. & Aamodt, P. O. (2017). Measuring Learning Outcomes. *European Journal of Education*, 52(1), 20–30.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale: Lawrence Erlbaum Associates, Inc.

Coil, D., Wenderoth, M. P., Cunningham, M. & Dirks, C. (2010). Teaching the Process of Science: Faculty Perceptions and an Effective Methodology. *CBE–Life Sciences Education*, 9, 524–535.

Cordes, N. (2016). *Schreiben im Biologiestudium*. Opladen/Toronto: Verlag Barbara Budrich/utb.

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K. & Kestin, G. (2019). Measuring Actual Learning versus Feeling of Learning in Response to Being Actively Engaged in the Classroom. *Proceedings of the National Academy of Sciences U. S. A.*, 116, 19251–19257.

Doran, G. T. (1981). There's a S. M. A.R.T Way to Write Management's Goals and Objectives. *Management Review*, 70, 35–36.

Evans, C., Howson, C. K. & Forsythe, A. (2018). Making Sense of Learning Gain in Higher Education. *Higher Education Pedagogies*, 3(1), 1–45.

Gamer, M., Lemon, J., Fellows, I. & Singh, P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement*. R Package v. 0.84.1. Verfügbar unter http://www.r-project.org (Zugriff am: 09.07.2020).

Hake, R. R. (1998). Interactive-Engagement versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses. *American Journal of Physics*, 66(1), 64–74.

Koo, T. K. & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163.

Kruger, J. & Dunning, D. (1999). Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of Personality and Social Psychology*, 77, 1121–1134.

Lam, T. C. M. (2009). Do Self-Assessments Work to Detect Workshop Success? *American Journal of Evaluation*, 30(1), 93–105.

Lauer, T. & Hendrix, J. (2009). A Model for Quantifying Student Learning via Repeated Writing Assignments and Discussions. *International Journal of Teaching and Learning in Higher Education*, 20(3), 425–433.

Lerner, N. (2007). Laboratory Lessons for Writing and Science. *Written Communication*, 24(3), 191–222.

Mager, R. F. (1984). *Developing Attitude Toward Learning*. Atlanta: The Center for Effective Performance.

Marx, J. D. & Cummings, K. (2007). Normalized Change. *American Journal of Physics*, 75(1), 87–91.

Nissen, J. M., Talbot, R. M., Thompson, A. N. & Dusen, B. Van (2018). Comparison of Normalized Gain and Cohen's D for Analyzing Gains on Concept Inventories. *Physical Review Special Topics – Physics Education Research*, 14, 1–12.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Verfügbar unter http://www.R-project.org (Zugriff am: 09.07.2020).

Raupach, T., Münscher, C., Beißbarth, T., Burckhardt, G. & Pukrop, T. (2011). Towards Outcome-Based Programme Evaluation: Using Student Comparative Self-Assessments to Determine Teaching Effectiveness. *Medical Teacher*, 33(8), e446–e453.

Raupach, T. & Schiekirka, S. (2017). *Handreichung zur Lernerfolgsevaluation*. Lüneburg: Electric Paper Evaluationssysteme GmbH.

Rosenthal, L. C. (1987). Writing Across the Curriculum: Chemistry Lab Reports. *Journal of Chemical Education*, 64(12), 996–998.

Savinainen, A. & Scott, P. (2002). The Force Concept Inventory: A Tool for Monitoring Student Learning. *Physics Education*, 37(1), 45–52.

Schiekirka, S., Reinhardt, D., Bei, T., Anders, S., Pukrop, T. & Raupach, T. (2013). Estimating Learning Outcomes from Pre- and Posttest Student Self-Assessments: A Longitudinal Study. *Academic Medicine*, 88(3), 369–375.

Timmerman, B. E., Strickland, D. C., Johnson, R. L. & Paynec, J. R. (2011). Development of a „Universal" Rubric for Assessing Undergraduates' Scientific Reasoning Skills Using Scientific Writing. *Assessment and Evaluation in Higher Education*, 36(5), 509–547.

Von Korff, J., Archibeque, B., Gomez, K. A., Heckendorf, T., McKagan, S. B., Sayre, E. C. et al. (2016). Secondary Analysis of Teaching Methods in Introductory Physics: A 50 K-Student Study. *American Journal of Physics*, 84(12), 969–974.

Wachtel, H. K. (1998). Student Evaluation of College Teaching Effectiveness: A Brief Review. *Assessment and Evaluation in Higher Education*, 23(2), 191–212.

Autor und Autorin

Nils Cordes und Anne-Kathrin Warzecha lehren und forschen an der Fakultät für Biologie der Universität Bielefeld. Kontakt: ncordes@uni-bielefeld.de; ak.warzecha@uni-bielefeld.de