

# Lernaufgaben mit automatisiertem Feedback in einem digitalen Hochschultutorium zur Vorlesung der Allgemeinen Chemie als Angebot zur Binnendifferenzierung

*Florian Trauten, Carolin Eitemüller, Maik Walpuski*

**Zusammenfassung:** Die hohen Abbruchquoten in naturwissenschaftlichen Studiengängen im Allgemeinen und speziell in der Chemie wurden zum Anlass genommen, ein digitales, binnendifferenzierendes Förderangebot für Chemiestudierende im ersten Semester zur Erhöhung des Studienerfolgs zu entwickeln. Da die Überwindung von Vorwissensdefiziten entscheidend für den Verbleib im Studium ist (Heublein et al., 2017), wurden Online-Lernaufgaben entwickelt, die basierend auf der individuellen Performance in einem multiple-try Feedback Algorithmus automatisiert adaptives Feedback bereitstellen. In einem Prä-post-Vergleichsgruppendesign wurde zudem die Rolle des Vorwissens auf die Wirksamkeit der zwei adaptiven Feedback-Algorithmen (elaboriert vs. korrektiv) untersucht, da die aktuelle Befundlage diesbezüglich keine eindeutigen Schlüsse zulässt. Für Studierende, die zuvor keinen Chemie Leistungskurs belegt haben, konnte dabei ein signifikanter positiver Effekt für adaptives elaboriertes Feedback gefunden werden.

**Schlüsselwörter:** Adaptives Feedback, Digitalisierung, Online-Lehre, Chemie, Studieneingangsphase, Hochschuldidaktik

## Learning tasks with automatic adaptive feedback in a digital university tutorial for a lecture in general chemistry as a tool for internal differentiation

**Summary:** High dropout rates in natural science and specifically in chemistry studies led us to provide online learning tasks for internal differentiation to foster freshmen's study success in chemistry study programs. Deficits in content knowledge have to be compensated during the study entry phase, otherwise the likelihood of student dropout is fuelled (Heublein et al., 2017). Thus, based on a multiple-try feedback algorithm various online learning tasks had been developed to provide adaptive feedback based on the learners' performance while processing the learning task. As preceding studies were not able to clarify the exact impact of prior content knowledge on the efficacy of guiding informative components on learning process, two types of feedback (elaborative vs. corrective) have been compared within a quantitative study with a pre-post comparison-group design. Here, the adaptive elaborated feedback was found to be significantly better for low-achieving students than corrective feedback alone.

**Keywords:** adaptiv feedback, digitalization, online teaching, chemistry, study entry phase, university didactics

## 1 Einleitung

Der Übergang von der Schule zur Hochschule stellt Studierende vor eine Vielzahl von Herausforderungen (Trautwein & Bosse, 2017). Die hohe Abbruchquote in Chemiestudiengängen deutscher Universitäten von aktuell 47% (Heublein et al., 2020) lässt darauf schließen, dass es nur einem Teil der Studierenden gelingt, diese Anforderungen zu meistern. Da Leistungsprobleme für 84% der Abbrechenden mindestens als wichtiger Grund aufgeführt werden (Heublein et al., 2017), kann man an den Universitäten grundsätzlich von einem *Performance-Problem* in Chemiestudiengängen sprechen.

Neben der Überwindung von Leistungsdefiziten müssen sich die Studierenden zu Beginn des Studiums in neue Sozialgefüge integrieren (Tinto, 1988), sodass das Verhältnis zu den Lehrpersonen und/oder Mitstudierenden gerade in der frühen Studieneingangsphase von eher distanzierterer Natur ist. Feedback muss zudem nun – anders als an der Schule – aktiv eingefordert werden (Heublein et al., 2017; Rost, 2018). Der Einfluss dieses *Feedback-Problems* auf das *Performance-Problem* erscheint fast als logische Folge, wenn man beachtet, dass Feedback an Hochschulen besonders selten von Studierenden aus der Gruppe der Studienabbrechenden in Anspruch genommen wird (Heublein et al., 2017), Feedback aber eine der wichtigsten Einflussgrößen auf Lernprozesse darstellt (Hattie & Timperley, 2007). Die Kontaktaufnahme zu anderen Studierenden und Dozierenden, um Feedback zu erhalten, kann so zu einem entscheidenden Faktor für den Verbleib an der Hochschule und für den Studienerfolg werden (Heublein et al., 2017; Sarcletti & Müller, 2011), es sei denn, es wird auf anderen Wegen Feedback bereitgestellt.

Für das Fach Chemie konnte gezeigt werden, dass Lücken im Vorwissen im Verlauf des ersten Semesters nicht aufgeholt werden (Averbeck, 2020). Über informatives tutorielles Feedback (ITF), das z.B. in einer Online-Lernumgebung Informationen zur Bewältigung einer Aufgabe zur Verfügung stellt, ohne unmittelbar die Lösung zu präsentieren, kann personenunabhängig effektiv Hilfestellung bereitgestellt werden (Narciss & Huth, 2004). Fraglich ist allerdings, ob elaboriertes Feedback das Vorwissen aller Studierenden konzeptionell berücksichtigen kann. Denn während Novizen tendenziell von erklärendem Feedback profitieren (Feedback Principle; Johnson & Priest, 2014; Moreno & Mayer, 2007), ist es naheliegend, dass dieser Vorteil bei zunehmender Expertise verschwindet (Expertise Reversal Effekt; Chen et al., 2017; Kalyuga, 2014). Ergebnisse aus Metastudien zur Wirksamkeit von Feedback im Allgemeinen (Bangert-Drowns et al., 1991; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Wisniewski et al., 2019) und Studien, die spezifisch den Vorwissenseinfluss untersuchen (Fyfe et al., 2012; Narciss, 2006; Smits et al., 2008), zeigen ein uneindeutiges und zum Teil kontroverses Bild von Effektstärken.

Bei den vorangestellten Studien variieren Stichprobe, Feedbackmaßnahme und/oder Aufgabenbereich, sodass aktuell keine gesicherten Erkenntnisse vorliegen. Ziel des hier vorgestellten Projektes war es daher zu ermitteln, ob mit einem binnendifferenzierenden Förderangebot, das die individuellen Voraussetzungen der Erstsemesterstudierenden in Bezug auf deren fachspezifische Vorkenntnisse durch adaptives Feedback stärker berücksichtigt, der Studienerfolg am Ende des ersten Semesters erhöht werden kann. Die Schlüsselrolle, die der Disziplin der Allgemeinen Chemie für das gesamte Chemiestudium zukommt, bietet daher berechtigte Gründe, genau dort eine Intervention anzusetzen.

## 2 Stand der Forschung

Eine Interventionsmaßnahme für eine vom Wissensstand her heterogene Lerngruppe, die das Vorwissen konzeptionell berücksichtigen soll, steht einer Reihe an Herausforderungen gegenüber, insbesondere, wenn Feedback automatisiert und fehlerspezifisch vom Computer ausgegeben werden soll. Für ein besseres Verständnis, wie Feedback wirkt, modelliert Narciss (2018) ausgewählte Lernprozesse in ihrem *Interactive two feedback-loops* (ITFL) Modell. Dieses weist deutliche Parallelen zu dem Modell des selbstregulierten Lernens von Butler und Winne (1995) auf. Es wird hier allerdings verstärkt Augenmerk auf die modellierten Subprozesse von internem und externem Feedback sowie die Einflussnahme von externem Feedback auf interne Informationsverarbeitungsprozesse gelegt. Das Modell in Abbildung 1 verdeutlicht auf anschauliche Weise die ablaufenden Prozessschritte der Informationsverarbeitung im Individuum bei der Bearbeitung einer Aufgabe und wie performancebasiertes Feedback entstehen sollte und wie Feedback auf diese wirken kann.

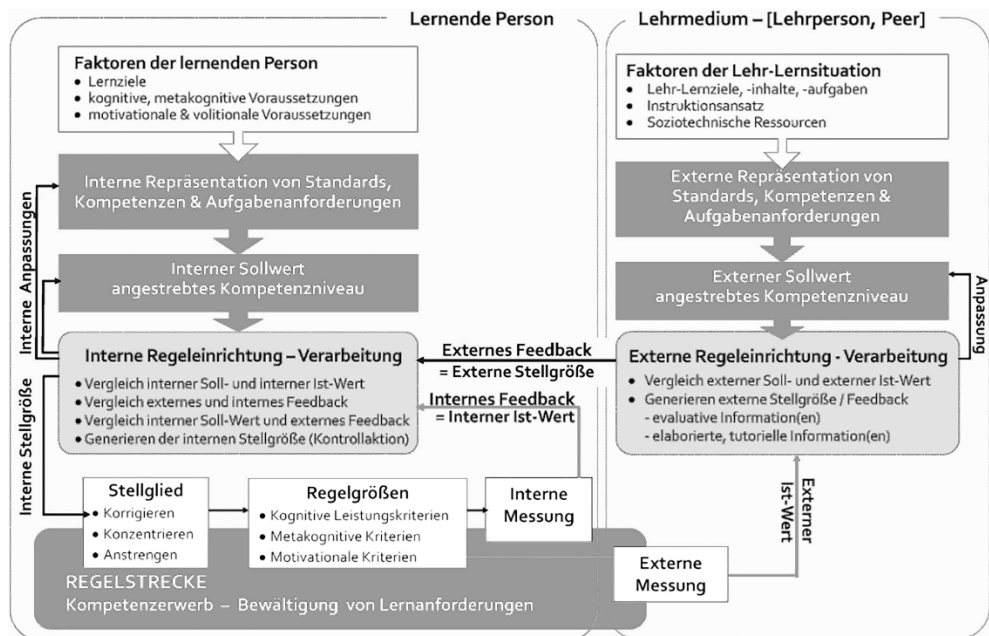


Abbildung 1: Interactive two feedback-loops model – ITFL-Modell (Narciss, 2018)

Daraus lassen sich weitere Anforderungen an die angestrebten digitalen Lernaufgaben ableiten. Wird mit der Bearbeitung einer Aufgabe begonnen, so bedingen die Faktoren der lernenden Person maßgeblich, wie die Aufgabeninformation verarbeitet werden. Das Vorwissen bedingt, ob z.B. Heuristiken für die spezifische Aufgabe herangezogen werden können oder nicht. Damit resultiert auf Personenebene – auf Basis einer individuellen Repräsentation der Aufgabe – eine von individuellen Faktoren bedingte Performance in einer Aufgabe. Diese Summe an individualitätsstiftenden Faktoren muss bei der Konzeption von Feedback bestmöglich berücksichtigt werden. Dieser systemtheoretische Ansatz geht davon aus, dass in dem jeweiligen Regelkreis (lernende Person oder Lehrmedium) bestimmte Regelgrößen fort-

laufend erfasst werden, sodass sich entsprechend des regelkreisspezifischen Blickwinkels ein *Ist-Wert* ergibt, der innerhalb des Regelkreises an die informationsverarbeitende Instanz rückgemeldet wird. Diese generiert dann entsprechend internes bzw. externes Feedback.

Der externe Regelkreis muss dabei nicht zwingend in einer Person stattfinden, sondern kann auch von einer automatisierten digitalen Lernaufgabe als Lernmedium vollzogen werden. Liegt eine Diskrepanz zwischen *Ist-* und *Soll-Wert* vor, werden dort, wie auch bei Butler und Winne (1995), Korrekturgrößen oder -maßnahmen spezifiziert (sogenannte *Stellgrößen*), um eine Veränderung in Richtung des Soll-Wertes herbeizuführen. Nimmt man also das beobachtbare Verhalten in Form der Performance in einer Aufgabe der Lernenden, so sind erfahrenere Lernende in der Lage, ihre Lösung eigenständig zu bewerten und gegebenenfalls auch zu korrigieren (internes Feedback). Wenn das interne Feedback kompetenzbedingt aber nicht ausreicht, um Selbstregulationsprozesse in Gang zu setzen, ist externe Unterstützung geboten. Auf Basis dieser externen Ist-Wert-Soll-Wert-Diskrepanz kann dann ein entsprechendes Feedback formuliert werden. Externes Feedback, das auf Basis von Fehlern Defizite aufzeigt, muss darauf aufbauend adaptive Hilfestellungen geben, die durch aktive Wissenskonstruktion Hilfe zur Selbsthilfe leisten. Denn das Ziel ist es, dass die Lernenden langfristig in die Lage versetzt werden, mittels internen Feedbacks ihre Performance selbst zu überwachen. Diese Korrekturen werden dann von einer korrigierenden Instanz (dem sogenannten *Stellglied*) in der lernenden Person bewertet und bestenfalls umgesetzt. Dabei wird ein neuer Ist-Zustand generiert, und der Loop beginnt erneut. Soll überprüft werden, ob das vorherige Feedback zur Verbesserung geführt hat, muss ein neuer Ist-Zustand erhoben werden. Folglich bedarf es eines multiple-try Feedbacks (Narciss, 2018). Für die Passung von Feedback zum Individuum spielt es daher eine große Rolle, wie detailliert und wie oft der Ist-Zustand erhoben wird. Feedback Preavailability darf nicht gegeben sein, denn die Performance eines Individuums zu adressieren, setzt mindestens eine vom Feedback unbeeinflusste Performance in einer Situation voraus, z.B. einen ersten Antwortversuch in einer Lernaufgabe. Soll das Feedback konkrete Hilfestellung geben, bieten ausschließlich ergebnisbezogene Feedbackarten (*knowledge of result*, *knowledge of performance*; vgl. Tabelle 1) nur wenig Unterstützung für die Wissenskonstruktion an, da sie relativ wenige Anhaltspunkte für Handlungs- bzw. Wissensoptimierung geben.

Feedback sollte folglich nicht nur evaluative Funktionen erfüllen, sondern vielmehr wie ein Tutor beziehungsweise eine Tutorin strategisch hilfreiche Informationen zur Überwindung von Hürden im Lernprozess liefern. Ein so beschaffenes Feedback wird als *informatives tutorielles Feedback* (ITF) bezeichnet. Dieses Feedback macht es sich zum Ziel, regulierend auf Lernprozesse zu wirken und Hilfestellung zu geben, sodass Lernende Wissens (strukturen) beziehungsweise Kompetenzen erwerben, um die Anforderungen der Lernsituation zu meistern, ohne dass ihnen dabei zuvor die Aufgabenlösung preisgegeben wurde (Narciss, 2006, 2014, 2018). Dazu bedient es sich je nach Zielsetzung verschiedener elaborierter Feedback-Typen (s. Tabelle 1).

Entlang verschiedener Metastudien (Bangert-Drowns et al., 1991; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Wisniewski et al., 2019) zeigen sich allerdings große Schwankungen in den Effektstärken hinsichtlich der Effektivität von Feedback-Typen im Allgemeinen. Aufgrund großer Varianz hinsichtlich der Effektstärken in Verbindung mit unterschiedlichen Forschungsdesigns kann die Feedbackforschung zurzeit auch keine gesicherten Erkenntnisse hinsichtlich des Vorwissenseinflusses auf die Wirksamkeit von Feedback unterschiedlichen Umfangs liefern (Fyfe et al., 2012; Narciss, 2006; Smits et al., 2008). Fyfe et al.

(2012) konnten dabei hinsichtlich des Erlernens von Äquivalenzrechnungen feststellen, dass bei steigendem Vorwissen der Vorteil elaborierter Hilfestellung zurückgeht und das eigenständige Finden von Fehlern zu höherem Lernerfolg führt. Bei Narciss (2006) wurde ebenfalls im mathematischen Kontext ein Vergleich zwischen rein korrekktivem und elaboriertem Feedback, hier explizit ITF, angestrebt. Hier konnten keine Treatment-Unterschiede unter Berücksichtigung des Vorwissens gefunden werden. Smits et al. (2008) konnten feststellen, dass rein evaluatives Feedback für Vorwissensstarke im Rahmen von Lerneinheiten zur Genetik zu besseren Leistungen führte. Bezüglich Studien, die den Zusammenhang von Vorwissen und Feedback explizit untersucht haben, zeigt sich also eine unklare Ergebnislage. Da für chemienahe Kontexte ebenfalls nur Indizien für einen solchen Zusammenhang gefunden werden konnten (Albacete & VanLehn, 2000; Moreno, 2004; Narciss & Huth, 2004), zeigt sich bereits für die domänenspezifische Feedbackforschung eine explizite Forschungslücke.

*Tabelle 1:* Klassifikation unterschiedlicher Feedback-Typen nach inhaltlichen Gesichtspunkten (Narciss, 2006, S. 23)

| Bezeichnung <sup>a</sup>   |                                                  | Beispiele für Feedback-Inhalte                                                                                                                                                                                                     |
|----------------------------|--------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Elaborierte Feedback-Typen | Knowledge of Performance [KP]                    | <ul style="list-style-type: none"> <li>• 15 von 20 Aufgaben richtig</li> <li>• 85% der Aufgaben gelöst</li> </ul>                                                                                                                  |
|                            | Knowledge of result/response [KR]                | <ul style="list-style-type: none"> <li>• Falsch/richtig</li> <li>• Nicht richtig/stimmt</li> </ul>                                                                                                                                 |
|                            | Knowledge of correct response [KCR]              | <ul style="list-style-type: none"> <li>• Angabe der korrekten Antwort</li> <li>• Markierung der korrekten Antwort</li> </ul>                                                                                                       |
|                            | Knowledge on task constraints [KTC]              | <ul style="list-style-type: none"> <li>• Hinweise auf Art der Aufgabe</li> <li>• Hinweise auf Bearbeitungsregeln</li> <li>• Hinweise auf Teilaufgaben</li> <li>• Hinweise auf Aufgabenanforderungen</li> </ul>                     |
|                            | Knowledge about mistakes [KM]                    | <ul style="list-style-type: none"> <li>• Anzahl der Fehler</li> <li>• Ort der Fehler/ des Fehlers</li> <li>• Art der Fehler/ des Fehlers</li> <li>• Ursache/n des/r Fehler(s)</li> </ul>                                           |
|                            | Knowledge on how to proceed [KH]<br>(„know how“) | <ul style="list-style-type: none"> <li>• Fehlerspezifische Korrekturhinweise</li> <li>• Aufgabenspezifische Lösungshinweise</li> <li>• Hinweise auf Lösungsstrategien</li> <li>• Leitfragen</li> <li>• Lösungsbeispiele</li> </ul> |
|                            | Knowledge on meta-cognition [KMC]                | <ul style="list-style-type: none"> <li>• Hinweise auf meta-kognitive Strategien</li> <li>• Metakognitive Leitfragen</li> </ul>                                                                                                     |

<sup>a</sup> Bei der Wahl der Bezeichnungen wurde einerseits darauf geachtet, die in der Feedback-Forschung üblichen Bezeichnungen zu verwenden beziehungsweise weiter zu führen, andererseits sollte die inhaltliche Feedback-Facetten in der Bezeichnung deutlich werden.

Spezifisch für automatisiertes Feedback in Online-Lernumgebungen erstreckt sich hingegen über den gesamten Bereich der Naturwissenschaften eine eklatante Forschungslücke: Cavalcanti et al. (2021) betrachten beispielsweise in ihrem Literatur-Review 3,178 Online-Artikel, die über ein automatisiertes Suchverfahren anhand von Schlagwortgruppen erhalten wurden (u.a. von SpringerLink, Scopus und ScienceDirect). Durch ein anschließendes dreischrittiges Review-Verfahren konnten daraus 63 Artikel klassifiziert werden, die sich hinreichend mit dem Thema automatisierten Feedbacks in Online-Lernumgebungen befassen. Mathematik war dabei mit lediglich drei Artikeln der einzige naturwissenschaftsnähe Vertreter. Über alle

Artikel konnte allerdings unabhängig von der Fachrichtung bei 32 Artikeln ein empirischer Nachweis für positive Effekte von automatisiertem Feedback auf die Performance von Nutzerinnen und Nutzern gefunden werden, wodurch sich tendenziell ein Trend ergibt, den es in chemischen Kontexten zu replizieren gilt. Auf den Einfluss von Vorwissen auf die Wirksamkeit von Feedback wurde in dieser Studie nicht eingegangen. Jenseits des Reviews liegen den technischen Stand betreffend auch nur wenige Anwendungsbeispiele für automatisiertes Feedback im Bereich chemiespezifischer Aufgaben vor. Diese beschränken sich auf kostenpflichtige Programme (z.B. Mastering Chemistry, ALEKS, Achieve), die überwiegend in den Vereinigten Staaten und Kanada genutzt werden. Weitere individuelle Online-Lösungen weisen hingegen funktionale Defizite auf, da sie z.B. Reaktionsgleichungen in der Chemie als offenen Aufgabentyp nicht berücksichtigen (für einen Überblick vgl. Eitemüller et al., 2023).

### 3 Fragestellung

Über einen systematischen Vergleich von zwei vom Informationsgehalt her unterschiedlichen Feedback-Typen, eingebettet in einen multiple-try Feedback Algorithmus, sollen weitere Erkenntnisse über den Einfluss des Vorwissens der Lernenden auf die lernförderliche Wirkung von Feedback-Typen gewonnen werden.

FF: Wie wirken sich zwei vom Feedbackumfang unterschiedliche Feedback Algorithmen (Error-Specific-Tutoring Feedback vs. korrekatives Feedback) in Abhängigkeit vom fachspezifischen Vorwissen auf den Studienerfolg von Erstsemesterstudierenden in der Allgemeinen Chemie aus?

Als Maß für den Studienerfolg wird in dieser Arbeit das Fachwissen (Post) – geprüft durch einen unabhängigen Fachwissenstest – sowie der Klausurerfolg am Ende des Semesters verstanden. Der Error-Specific-Tutoring Feedback Algorithmus (EST-Feedback) zeichnet sich durch verschiedene ITF-Typen aus. Diese scheinen besonders vielversprechend für Studierende mit Vorwissensdefiziten zu sein. Denn besonders diese Studierenden benötigen beim Auftreten von Fehlern elaboriertes Feedback (Krause et al., 2009; Mayer & Johnson, 2010; Moreno, 2004). So erfordert das selbstständige Finden von Rechenfehlern und/oder fehlerhaften Ansätzen im Lösungsprozess bereits ein gewisses Grundmaß an Vorwissen (Schuhmacher, 2008). Daraus ergibt sich folgende erste Hypothese:

H1: Studierende mit geringem fachspezifischem Vorwissen profitieren in ihrem Studienerfolg stärker durch ein Training mit einem EST-Feedback als Studierende mit hohem Vorwissen.

Bei zunehmender Expertise ist allerdings zu erwarten, dass der Nutzen von elaboriertem Feedback sinkt, sodass sich für Fortgeschrittene und Experten kein Vorteil mehr ergibt (Expertise Reversal Effect; Kalyuga, 2014) oder sogar Nachteile durch zu umfangreiches Feedback entstehen können (z.B. Smits et al., 2008). Ein korrektiver Feedback Algorithmus (KOR-Feedback), der nur auf eine fehlerhafte Lösung hinweist, scheint damit für Fortgeschrittene die bessere Wahl zu sein.

- H2: Studierende mit hohem fachspezifischem Vorwissen profitieren in ihrem Studienerfolg stärker durch ein Training mit einem KOR-Feedback als Studierende mit geringem Vorwissen.

## 4 Material

### 4.1 Der Feedback Algorithmus

Wie bereits in Abschnitt 2 beschrieben, wird im Rahmen des ITFL-Modells eine vorherige Studierendenleistung benötigt, um fehlerspezifisches Feedback geben zu können. Das EST-Feedback soll aber zusätzlich zu einer fehlerspezifischen Rückmeldung ITF bereitstellen. Dieses soll passend zu bestimmten Fehlerkonstellationen automatisiert eine Hilfestellung enthalten, die den Lernenden helfen soll, die Aufgabe selbst zu vollenden, aber ohne die Lösung der Aufgabe zu präsentieren. Somit bedarf es einer Aufgabe, die einem multiple-try Schema mit mindestens zwei möglichen Antwortversuchen folgt. In Anlehnung an den Bug-Related-Tutoring Feedback Algorithmus (Narciss & Huth, 2004) wurde ein multiple-try Algorithmus mit drei möglichen Antwortversuchen entwickelt, der bei Lernaufgaben im ersten und zweiten Feedback neben korrektivem Feedback (KR) noch elaborierte Feedback-Typen in Form von ITF präsentiert (KM, KH; s. Tabelle 1). Das letzte Feedback enthielt neben dem Indizieren von spezifischen Fehlern (KM) zusätzlich die richtige Lösung (KCR). Die richtige Lösung wurde dabei feingliedrig mit Teilschritten (KH) präsentiert. Parallel zu diesem EST-Feedback Algorithmus wurde ein dreistufiger korrektiver Feedback Algorithmus (KOR-Feedback) entwickelt, der in der ersten und zweiten Rückmeldung nur KR und in der letzten Feedback Nachricht KCR präsentierte.

Durch zwei korrekte Antworten im ersten Lösungsversuch oder durch sechs Aufgaben mit beliebiger Performance gelangten die Studierenden zu einer Testaufgabe, mit der anteilig Bonuspunkte für die Klausur gesammelt werden konnten. Daraus resultierte nicht nur adaptives Feedback, sondern auch eine adaptive Aufgabenmenge. Zur Beantwortung der Forschungsfrage wurden zu zwanzig Lernzielen basierend auf diesem Schema insgesamt 120 Online-Aufgaben mit adaptivem fehlerspezifischem Feedback (EST-Feedback) beziehungsweise korrektivem Feedback (KOR-Feedback) in der Webapplikation JACK® (Pobel & Striwe, 2019; Striwe, 2016) realisiert (sechs je Lernziel). Die Testaufgabe am Ende diente zur Lernzielüberprüfung. Da sich diese interaktiven Aufgaben als externes Tool in die Lernplattform Moodle einbetten lassen, eigneten sie sich sehr gut für einen webbasierten Übungsbetrieb.

### 4.2 Lernzielbasierte Digitalisierung von Chemieaufgaben

Die Auswahl spezifischer Lernziele des gegebenen domänenspezifischen Wissens als anzustrebende Kompetenz sowie typischer Aufgaben, die diese kompetenzbildenden Lernziele bedienen, sind laut Narciss und Huth (2004) grundlegend für die Spezifizierung von Feedbackelementen. Chemie ist für viele ohne mathematische Formeln, Reaktionsgleichungen, Diagramme usw. überhaupt nicht denkbar.

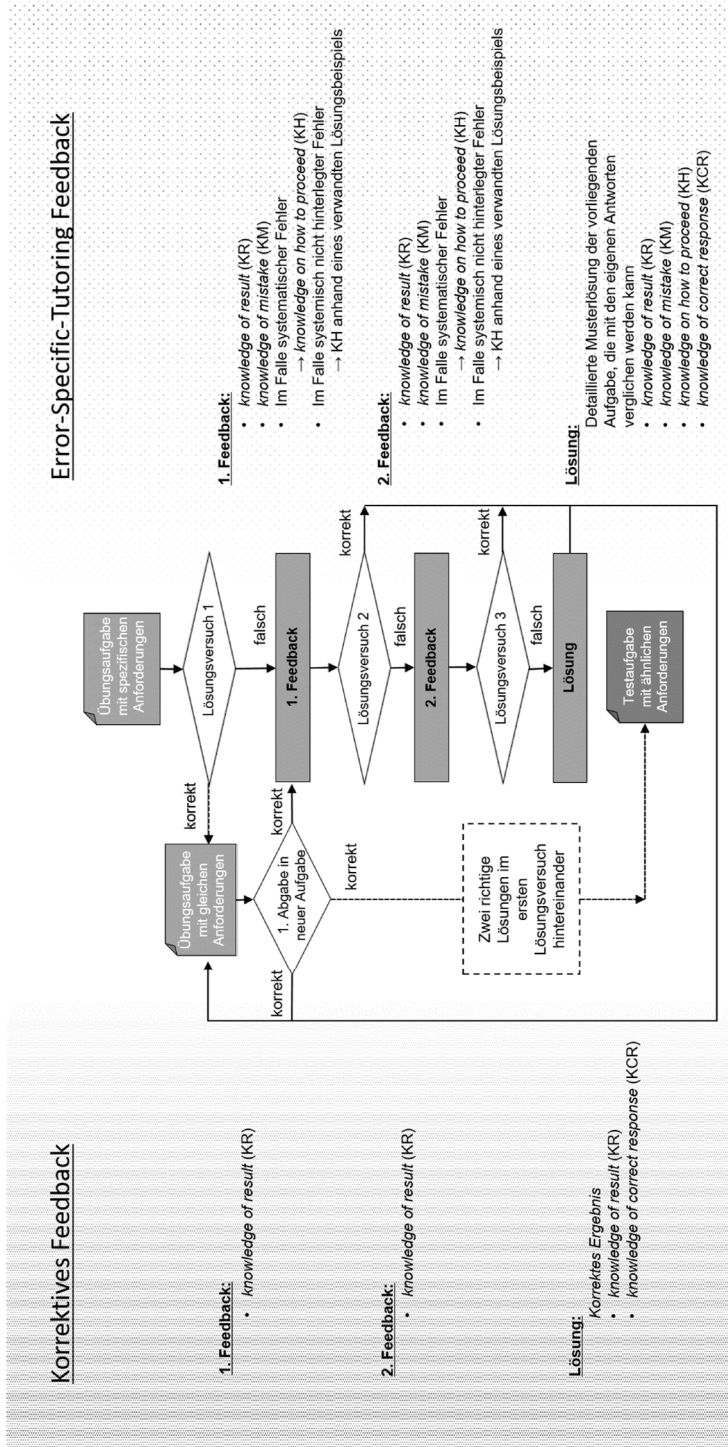


Abbildung 2: Gegenüberstellung des korrekativen Feedback Algorithmus (KOR-Feedback) mit dem Error-Specific-Tutoring-Feedback Algorithmus (EST-Feedback)

Eine Analyse von Lehrbüchern, die Studierende in der Eingangsphase des Chemiestudiums häufig zum Lernen nutzen, konnte zeigen, dass wenigstens 87% aller Lehrbuchseiten mindestens eine Visualisierung enthalten. Mit wenigstens 60% Verteilungsanteil stellen symbolische Repräsentationsformen dabei den Hauptanteil. Im Mittel sind 34% davon Reaktionsgleichungen (vgl. Dickmann, 2019). All diese Repräsentationsformen in domänenspezifischer Schreibweise müssen durch die Software adäquat abgebildet werden. Klassische Übungs- und Klausuraufgaben repräsentieren in diesem Zusammenhang die intendierten Lernziele. Eine Analyse von Übungsaufgaben, Klausuraufgaben und weiteren Modulinhalten nach Lernzielen lieferte daher im Rahmen dieser Arbeit entsprechende Operatoren und damit die nötigen Anforderungen an das Software-Tool. Parallel dazu wurden typische Fehlerquellen ermittelt und bei der Konzeption des Feedbacks berücksichtigt, wodurch sich hinsichtlich der automatisierten Fehlererkennung weitere Anforderungen an das Software-Tool ergaben. Dieser Prozess ist in Abbildung 3 verdeutlicht.

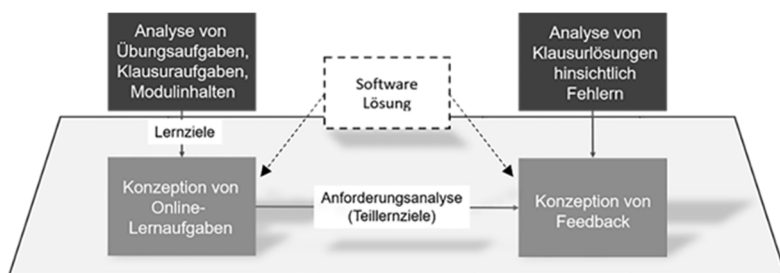


Abbildung 3: Überführung der Papier-Bleistift-Aufgaben in digitale Lernaufgaben unter Berücksichtigung der Einschränkungen durch die Software Lösung

Bei der Überführung von klassischen Paper-Pencil Aufgaben in ein Online-Format ist zudem zu vermuten, dass sich durch diesen Transfer die Anforderungen der Aufgaben verändern würden. Für chemische Kontexte konnte ein Effekt des Mediums allerdings widerlegt werden, solange die Lernziele zwischen analogen und digitalen Lernaufgaben konstant gehalten wurden. Einschränkungen konnten eindeutig nicht in der Aufgabenstellung, sondern bei der Eingabe der Lösung verortet werden (Ashton et al., 2005), z.B. wenn gewohnte Schreibweisen nicht möglich waren. Gesucht wurde daher eine Softwarelösung, die zugleich die Möglichkeit für eine chemiespezifische Schreibweise bot und zugleich eine automatisierte synchrone Evaluation von Antworten, unter anderem in dieser spezifischen Schreibweise, ermöglichte. In Zusammenarbeit mit dem Moodle-Kompetenzzentrum der Universität Duisburg-Essen sowie im interdisziplinären Austausch mit Doktorierenden und Professor/innen der Informatik, rückte das vom Ruhr Institute for Software Technology entwickelte Java-basierte E-Assessment System „JACK® – Lernen. Testen. Prüfen.“ in den Fokus.<sup>1</sup>

Während Freitext und Zeichenaufgaben aufgrund von hochkomplexen Evaluationsstrukturen für fehlerspezifisches Feedback nicht erschlossen werden konnten, ist regelbasiertes Feedback zu verschiedensten mathematischen Aufgabentypen, die einen hohen Stellenwert in der Allgemeinen Chemie haben (Goldhausen, 2015), einfach zu realisieren. Zudem wurde

<sup>1</sup> Sowohl formative als auch summative Assessments waren hiermit zu realisieren. Jede Einreichung einer Aufgabe wird dabei vollständig unabhängig von einer anderen getestet. Bis 2015 wurde JACK® im Informatikkontext an 4 Universitäten in Deutschland und England eingesetzt und erhielt dabei sehr viel positiven Zuspruch seitens der Studierendenschaft (Striewe, 2016).

ein Editor für chemische Summenformeln und Reaktionsgleichungen (Eitemüller et al., 2023; Pobel & Striewe, 2019), sowie ein Eingabetool für Atomorbitaldiagramme (Striewe et al., 2020) fertiggestellt. So ergab sich für die domänenexklusiven Schreibweisen der Chemie mit ihrer spezifischen Semantik die nötige Fehlersensitivität für passgenaues Feedback.

## 5 Studiendesign, Stichprobe und Testinstrumente

### 5.1 Studiendesign und Stichprobe

Aufgrund der oben genannten Schlüsselrolle des Moduls der Allgemeinen Chemie wurden die Aufgaben in einem Online-Tutorium inhaltlich parallel zu den Vorlesungsinhalten arrangiert und hinsichtlich ihrer Lernwirksamkeit getestet. Als Teilstichprobe der Grundgesamtheit der Chemiestudierenden an deutschen Universitäten wurden die Erstsemesterstudierenden der Universität Duisburg-Essen mit der Fachwahl Bachelor Chemie gewählt. Aufgrund der inhaltlichen Parallelität und des gemeinsamen Besuchs der Vorlesungen in den ersten Semestern wurden Studierende des Studiengangs Bachelor Water Science mit hinzugezogen.

Die Einteilung der Studierenden in die Übungsgruppen hatte bei dieser Feldstudie nach der ersten Veranstaltung schnell innerhalb der ersten Semesterwoche zu erfolgen. In der Kürze der Zeit ist eine Auswertung von Fachwissensdaten nicht möglich gewesen. Daher wurden die Studierenden über die Kurswahl in der Oberstufe (Chemie Leistungskurs: ja/nein), die als guter Prädiktor für das Vorwissen gilt (Averbeck, 2020), gleichmäßig und randomisiert auf die Interventionsgruppen aufgeteilt ( $2 \times 2$ -Design, Kurswahl  $\times$  Feedback-Typ, vgl. Tabelle 2). Eine Poweranalyse mit G\*Power (Faul et al., 2007) ermittelt zum Nachweis mittlerer Effektstärken ( $f^2 = 0.15$ ,  $\alpha = 0.05$ ,  $Power = 0.80$ ) inklusive Korrektur für hierarchische Datenstrukturen eine notwendige Stichprobe von 42 Probanden pro Kategorie (LK/kein LK), also eine minimale Gesamtstichprobengröße von  $N = 84$  Probanden. Da erfahrungsgemäß aufgrund des Dropouts über die Semester und des damit verbundenen Schwundes an Probanden damit zu rechnen war, dass die Gesamtstichprobengröße über ein Semester nicht ausreichend für die anschließende Auswertung sein könnte, wurden insgesamt drei Interventionsstudien über die Wintersemester (WiSe) 2018–2020 durchgeführt.

Tabelle 2: Interventionsgruppen im  $2 \times 2$ -Design

| Kurswahl Chemie    | Feedback-Typ             |                          |
|--------------------|--------------------------|--------------------------|
|                    | EST                      | KOR                      |
| Kein Leistungskurs | EST <sub>(kein LK)</sub> | KOR <sub>(kein LK)</sub> |
| Leistungskurs      | EST <sub>(LK)</sub>      | KOR <sub>(LK)</sub>      |

Für die drei Interventionsjahrgänge konnten in der ersten Semesterwoche auf Basis der Kurswahl folgende Teilstichproben erhalten werden (s. Tabelle 3, linke Seite). Die Bearbeitung der Übungsaufgaben ist mit Bonuspunkten für die Klausur und des Weiteren mit einer Probandenvergütung von 50 € pro Person belohnt worden. Dazu mussten die Studierenden mindestens zwei Aufgaben zu einem Lernziel in Folge richtig lösen oder alle sechs Aufgaben eines Aufgabenpakets sowie eine abschließende Testaufgabe bearbeiten (s. Abbildung 2). Da zudem nur sehr wenige Studierende alle verfügbaren Aufgabensets im Sinne des Algorithmus

vollständig bearbeitet haben, wurde bereits im Rahmen der Pilotierungsstudie, in der sich dieses Verhalten bereits abzeichnete, die aktive Arbeit mit den Übungsaufgaben an die Regel geknüpft, dass die Hälfte der zehn in der Pilotierung verfügbaren Aufgabensets mit je zwei Lernaufgaben bearbeitet worden sein musste. Die verfügbaren Aufgabensets wurden dann sukzessiv auf 20 Lernziele erweitert. Um am Ende der Studie einen Effekt auf den Studien-erfolg der Studierenden untersuchen zu können, wurde diese Regel für die aktive Übungs-teilnahme auch für die Folgejahre beibehalten. So ergibt sich folgende Verteilung der „akti-ven“ Gesamtkohorte für die Evaluation des Fachwissenszuwachs (s. Tabelle 3, rechte Seite). Um tiefergehende Informationen über die Wirkweise der Lernaufgaben in Abhängigkeit von dem unterstützten Feedback-Typ gewinnen zu können, konnte zusätzlich auf eine Kontroll-kohorte aus dem WiSe 2016 zurückgegriffen werden. Diese besuchte bei identischen Vorle-sungsinhalten *nur* eine reine Präsenzübung ohne interaktive Lernaufgaben.

Tabelle 3: Verteilung der „aktiv“ Studierenden pro Interventionsgruppe

| DEMO               | Interventionsgruppe |     | $\Sigma$ | „aktiv“ | Interventionsgruppe |     | $\Sigma$ |
|--------------------|---------------------|-----|----------|---------|---------------------|-----|----------|
|                    | KOR                 | EST |          |         | KOR                 | EST |          |
| Kein Leistungskurs | 90                  | 96  | 186      | →       | 28                  | 33  | 61       |
| Leistungskurs      | 50                  | 55  | 105      |         | 30                  | 31  | 61       |
| $\Sigma$           | 140                 | 151 | 291      |         | 58                  | 64  | 122      |

Mithilfe eines Pre-Post-Vergleichsgruppendesigns sollten die Aufgaben hinsichtlich ihrer Lernwirksamkeit im Rahmen des folgenden Designs geprüft werden (s. Tabelle 4). Averbek (2020) konnte neben dem Chemie-Vorwissen noch weitere Einflussfaktoren (kognitive Fähigkeiten, mathematische Fähigkeiten, Studien- und Lern- und Leistungsmotivation) auf den Studienerfolg identifizieren.

Tabelle 4: Studiendesign für die Interventionsstudie

| Semesterverlauf                  | EST <sub>(kein LK)</sub> + EST <sub>(LK)</sub>      | KOR <sub>(kein LK)</sub> + KOR <sub>(LK)</sub> |
|----------------------------------|-----------------------------------------------------|------------------------------------------------|
| Erste Semesterwoche              | Demografie-Fragebogen                               |                                                |
|                                  | Allgemeine Chemie (Pre)                             |                                                |
|                                  | Kognitiver Fähigkeitentest                          |                                                |
|                                  | Test für mathematische Fähigkeiten                  |                                                |
|                                  | Tutorial für den Umgang mit JACK                    |                                                |
|                                  | Studienmotivation (Erwartung, Wert, Kosten)         |                                                |
|                                  | Allgemeine Lern- und Leistungsmotivation im Studium |                                                |
| Semesterbegleitende Intervention | Arbeiten mit den Lernaufgaben                       |                                                |
|                                  | empfundene Aufgabenschwierigkeit/                   |                                                |
|                                  | Denkanstrengung/Nützlichkeit des Feedbacks/         |                                                |
| Letzte Semesterwoche             | Nutzerfreundlichkeit der Aufgaben, time-on-task     |                                                |
|                                  | Allgemeine Chemie (Post)                            |                                                |
| Vorlesungsfreie Zeit             | Klausur/Nachschreibklausur                          |                                                |

Diese wurden als Kontrollvariablen vor der Intervention miterhoben. Um den Arbeitsaufwand, der sich durch die Studie für die Testpersonen ergab, angemessen darstellen zu können, werden diese hier mit dargestellt, im Umfang dieses Papers aber nicht näher behandelt.

Die Pilotierung der interaktiven, durch Feedback gestützten Lernaufgaben wurde im selben Design wie die Hauptstudie durchgeführt und unterscheidet sich lediglich durch die Menge an verfügbaren JACK-Aufgaben von der Hauptstudie. Bedingt durch die COVID-19 Pandemie ergaben sich nur für die Kohorte des WiSe 2020/2021 (HS 2) Einschränkungen. Da dieses Semester an der Universität als reines Online-Semesters durchgeführt und zudem bedingt durch die Maßnahmen gegen das Virus verspätet in das Semester gestartet wurde, musste auf einige Testinstrumente verzichtet werden (mathematische Fähigkeiten, Studienmotivation, Lern- und Leistungsmotivation, Selbstkonzept und Lernstrategien).

Tabelle 5: Stichprobenbeschreibung

|     | Demografie |                  |                   | + | Allg. Chemie (Pre) |                   |                   | + | Allg. Chemie (Post) |                   |                   | + | Aktive<br>Übungsteilnahme |                   |                   |
|-----|------------|------------------|-------------------|---|--------------------|-------------------|-------------------|---|---------------------|-------------------|-------------------|---|---------------------------|-------------------|-------------------|
|     | N          | %                | %(♀) <sup>1</sup> |   | N                  | %                 | %(♀) <sup>1</sup> |   | N                   | %                 | %(♀) <sup>1</sup> |   | N                         | %                 | %(♀) <sup>1</sup> |
| Ges | 291        | 100 <sup>2</sup> | 42.6              |   | 234                | 80.4 <sup>2</sup> | 42.7              |   | 149                 | 51.2 <sup>2</sup> | 36.9              |   | 122                       | 41.9 <sup>2</sup> | 40.2              |
| Pil | 84         | 100 <sup>3</sup> | 39.3              |   | 77                 | 91.7 <sup>3</sup> | 39.0              |   | 54                  | 64.3 <sup>3</sup> | 35.2              |   | 39                        | 46.4 <sup>3</sup> | 43.6              |
| HS1 | 92         | 100 <sup>3</sup> | 41.3              |   | 83                 | 90.2 <sup>3</sup> | 42.2              |   | 58                  | 63.0 <sup>3</sup> | 25.7              |   | 53                        | 60.4 <sup>3</sup> | 24.5              |
| HS2 | 115        | 100 <sup>3</sup> | 44.3              |   | 74                 | 64.3 <sup>3</sup> | 47.3              |   | 37                  | 32.2 <sup>3</sup> | 56.8              |   | 30                        | 26.1 <sup>3</sup> | 63.3              |

Anmerkungen: <sup>1</sup> Anteil weiblicher Studierender bezogen auf den jeweiligen MZP; <sup>2</sup> bezogen auf die Gesamtstichprobe (Studierende, die den Demografie-Fragebogen bearbeitet haben); <sup>3</sup> bezogen auf die Teilstichprobe je WiSe (Studierende, die den Demografie-Fragebogen bearbeitet haben)

Über die drei Interventionsstudien wurde eine Gesamtkohorte von  $N = 291$  Studierenden<sup>2</sup> ( $M_{Alter} = 20.72$  Jahre, 57.4% männlich, 42.6 % weiblich) aus Erstsemesterstudierenden der Studiengänge Bachelor Chemie und Bachelor Water Science erhalten (s. Tabelle 5).

## 5.2 Testinstrumente

Damit ein Vergleich zwischen den Jahrgängen hinsichtlich der Leistungsdaten möglich wurde (allgemeine Chemie Fachwissen, mathematische Fähigkeiten, kognitive Fähigkeiten), wurden die Leistungsdaten mithilfe der Item Response Theorie (Rasch Modelle) ausgewertet. Durch Fixierung der Itemschwierigkeiten lassen sich auch sukzessiv entstandene Stichproben auf einer gemeinsamen Skala abbilden. Denn anders als die klassische Testtheorie, die von homogenen Itemschwierigkeiten ausgeht, ermittelt man in der IRT sowohl die Personenfähigkeit ( $\theta$ ) als auch die Itemschwierigkeit ( $\sigma$ ) und setzt sie für jedes Item in Bezug zueinander, sodass sich letztlich beides auf einer eindimensionalen Logit-Skala einordnen lässt. Wenn der Personenparameter ( $\theta$ ) einem Itemparameter ( $\sigma$ ) entspricht, beträgt dabei die Lösungswahrscheinlichkeit eines Items 50%. Da sich die Itemschwierigkeiten zwischen unterschiedlichen Stichproben fixieren (also gleichsetzen) lassen, liegen die in diesem Modell geschätzten Personenfähigkeiten ( $\theta$ ) auf derselben Skala und lassen sich so gleichwertig verwenden und vergleichen. Der WLE-Schätzer gilt dabei als geeignetster Schätzer für Personenfähigkeiten im Rahmen von IRT-Modellen (Walter, 2005). Im Folgenden wird daher nicht mit z-standardisierten Personenfähigkeiten, sondern mit den WLEs (Weighted Maxi-

2 Die Stichprobe hat einen Anteil von 17.6% Probanden, die nicht in Deutschland geboren wurden. Dabei haben 40.2% einen Migrationshintergrund, d.h., dass mindestens ein Elternteil nicht in Deutschland geboren wurde. Bei 29.1% der Probanden ist es bei keinem Elternteil der Fall. Insgesamt gaben 69.8% der Probanden die deutsche Sprache als Muttersprache an.

mum Likelihood Estimates; Warm, 1989), also Personenfähigkeitsparametern auf einer Logit-Skala gerechnet.

Die Itemschwierigkeiten wurden im Rahmen des interdisziplinären Projektes ALSTER (Akademisches Lernen und Studienerfolg in der Eingangsphase von naturwissenschaftlich-technischen Studiengängen) auf den Pre-Messzeitpunkt der Kontrollkohorte fixiert, da hier die größte Stichprobe für eine valide Schätzung gegeben war (Averbeck, 2020). Die Erhebung des Fachwissens in der Allgemeinen Chemie bildet im Rahmen dieses Prä-Post-Vergleichsgruppendesigns das zentrale Testinstrument zur Erfassung des Studienerfolgs und liegt somit im Fokus dieses Papers. Hierzu wurde auf den Fachwissenstest von Freyer (2013), modifiziert nach Averbeck (2020), zurückgegriffen. Er liefert eine akzeptable bis gute EAP/PV-Reliabilität von .765 zum ersten und von .802 zum zweiten Messzeitpunkt. Der Item-Fit aller 35 Items (Range: 0.76–1.275) liegt zudem innerhalb der von Bond et al. (2020) empfohlenen Spanne für gute Infit- und Outfit-Parameter für Item Mittelwertquadrate (Range: 0.7–1.3; S. 244). Für das allgemeine Chemie-Fachwissen liegt die WLE Reliabilität bei .752 für den ersten beziehungsweise .789 für den Post-Messzeitpunkt am Ende des ersten Semesters. Für die mathematischen Fähigkeiten zu Beginn des ersten Semesters zeigt sich mit .842 ebenfalls eine gute WLE Reliabilität über KG bis hin zur HS1.

Tabelle 6: Testinstrumente zur Erfassung von Kontrollvariablen

| Skala                     | Quelle                  | N <sub>Items</sub> | EAP/PV |
|---------------------------|-------------------------|--------------------|--------|
| Kognitive Fähigkeiten     | Heller & Perleth (2000) | 25                 | .77    |
| Mathematische Fähigkeiten | Müller et al. (2018)    | 23                 | .86    |

## 6 Ergebnisse

### 6.1 Leistungsunterschiede im Fachwissen zwischen Interventionsgruppen und Kontrollgruppe

Über einen Vergleich mit der Kontrollkohorte wurde zunächst untersucht, ob ein Training mit EST- bzw. KOR-Feedback (in Abhängigkeit vom Vorwissen) zu signifikant besseren Leistungen im Fachwissen am Ende des ersten Semesters führt. Hierzu wurden die Daten des unabhängigen Fachwissenstests analysiert, die sowohl für alle drei Interventionsjahrgänge als auch für die Kontrollkohorte zur Verfügung standen. Der Einfluss der Interventionsbedingungen, bei sonst gleicher Lehre und gleichen Dozierenden sowie vergleichbarer Abiturnote ( $F(1,1216) = 1,243, p = .265$ ), auf den Fachwissensstand am Ende des ersten Semesters wurde mithilfe einer multiplen linearen Regression untersucht, hierbei zeigen sich signifikante Effekte ( $F(4, 224) = 106.8, p \leq .001$ ). Neben dem Fachwissen Chemie zu Beginn des Semesters und der Kurswahl in der Oberstufe wurden in einem ersten Modell zwei weitere Variablen hinzugezogen. Durch diese zwei Dummy-Variablen EST-Feedback (0 = nein, 1 = ja) und KOR-Feedback (0 = nein, 1 = ja) wird der Vergleich beider Feedback-Typen mit der Kontrollkohorte ermöglicht. Das Fachwissen zu Beginn des Semesters ( $\beta = .740, p \leq .001$ ) und die Kurswahl ( $\beta = .103, p \leq .05$ ) bieten erwartungskonform gute prädiktive Kraft. Zusätzlich trägt der Prädiktor, ob EST-Feedback erhalten wurde ( $\beta = .107, p \leq .05$ ), signifikant zur Varianzaufklärung bei (vgl. Modell 1 in Tabelle 7). Der Erhalt von KOR-

Feedback ist hingegen unerheblich für den Fachwissensstand am Ende des Semesters (s. Modell 1, Tabelle 7). Ob die Studierenden unabhängig von ihrer Kurswahl vom EST-Feedback profitieren, wurde in einem weiteren Modell überprüft.

Tabelle 7: Hierarchische lineare Regressionen zur Vorhersage des Fachwissens (Post)

| Modell                  | Modell 1 |          |      | Modell 2 |          |      | Modell 3 |          |      |
|-------------------------|----------|----------|------|----------|----------|------|----------|----------|------|
| Variable                | B        | $\beta$  | SE   | B        | $\beta$  | SE   | B        | $\beta$  | SE   |
| Konstante               | .453 *** |          | .076 | .410 *** |          | .083 | .408 *** |          | .078 |
| Allg. Ch.               | .930 *** | .734 *** | .055 | .920 *** | .725 *** | .055 | .920 *** | .730 *** | .055 |
| Fachw. (pre)            |          |          |      |          |          |      |          |          |      |
| Kurswahl                | .261 *   | .107 *   | .105 | .408 *   | .167 *   | .158 | .414 **  | .169 **  | .125 |
| EST-Feedback            | .285 *   | .107 *   | .116 | .489 **  | .184 **  | .151 | .491 **  | .184 **  | .148 |
| KOR-Feedback            | .165     | .060     | .120 | .129     | .047     | .159 | .135     | .049     | .120 |
| Kurswahl × EST-Feedback |          |          |      | -.469 *  | -.134 *  | .235 | -.474 *  | -.136 *  | .214 |
| Kurswahl × KOR-Feedback |          |          |      | .015     | .004     | .242 |          |          |      |
| korrr. R <sup>2</sup>   | .646     |          |      | .650     |          |      | .656     |          |      |
| $\Delta R^2$            |          |          |      | .004     |          |      | .01 *    |          |      |

Anmerkungen: \*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

Ein zweites Modell (vgl. Tabelle 7), welches die Interaktion aus Feedback und Kurswahl in der Oberstufe berücksichtigt, zeigt allerdings einen signifikanten negativen Effekt für die Interaktion aus EST-Feedback und Kurswahl ( $\beta = -.134$ ,  $p \leq .05$ ). Da dieses Modell nicht signifikant besser misst als Modell 1 ( $\Delta R^2 = .004$ ,  $p = .089$ ) und der Interaktionseffekt von KOR-Feedback und Kurswahl in der Oberstufe mit  $\beta = .004$  ( $p = .170$ ) vernachlässigbar gering ausfällt, wurde ein weiteres Modell mit einem Interaktionsterm aus EST-Feedback und der Chemie-Kurswahl berechnet (vgl. Modell 3 in Tabelle 7). Es zeigt sich eine signifikante Verbesserung der Varianzaufklärung im Vergleich zu Modell 1 bei sonst gleichbleibenden Wirkzusammenhängen ( $\Delta R^2 = .01$ ,  $p \leq .001$ ). Der Vorteil von EST-Feedback ( $\beta = .184$ ,  $p \leq .01$ ) wird durch den Interaktionsterm nahezu aufgehoben ( $\beta = -.136$ ,  $p \leq .05$ ).

Dass sich nur ein Vorteil für diejenigen Studierenden durch EST-Feedback ergibt, die keinen Chemie-Leistungskurs in der Oberstufe besucht haben, zeigt sich auch rein deskriptiv in Abbildung 4. Studierende ohne Leistungskurs, die aber EST-Feedback bekommen haben ( $EST_{(\text{kein LK})}$ ), haben ein vergleichbares Fachwissen ( $M = 1.18$ ,  $SD = 1.16$ ) wie Studierende der Kontrollgruppe ( $KG_{(\text{LK})}$ ), die einen Chemie-Leistungskurs besucht haben ( $M = 1.28$ ,  $SD = 1.23$ ).

Zu Semesterbeginn lässt sich kein Unterschied zwischen den Subgruppen mit gleicher Kurswahl hinsichtlich des Vorwissens in der allgemeinen Chemie feststellen ( $ANOVA_{(\text{kein LK})}$ :  $F(2, 112) = 1.152$ ,  $p = .320$ ;  $ANOVA_{(\text{LK})}$ :  $F(2, 73) = 0.127$ ,  $p = .881$ ). Die entsprechende ANOVA, die den Fachwissensstand zum Ende des Semesters zwischen den Interventionsgruppen vergleicht, zeigt im post-hoc Test (Games-Howell) keinen signifikanten Unterschied zwischen der Interventionsgruppe ohne Leistungskurs mit EST-Feedback zu allen anderen Subgruppen, die einen Leistungskurs in der Oberstufe belegt hatten ( $EST_{(\text{LK})}$ ,  $KOR_{(\text{LK})}$  und  $KG_{(\text{LK})}$ ) ( $F(5, 223) = 11.26$ ,  $p \leq .001$ ). Die Gruppe  $EST_{(\text{kein LK})}$  ist zudem signifikant besser ( $p \leq .01$ ) als die Kontrollgruppe ohne Leistungskurs ( $M = .30$ ,  $SD = 1.23$ ) (vgl. Abbildung 4). Im Vergleich zur Gruppe  $KOR_{(\text{kein LK})}$  ( $M = 0.45$ ,  $SD = 0.92$ ) ist zumindest ein möglicher Trend zugunsten der Gruppe  $EST_{(\text{kein LK})}$  zu erkennen ( $p = .082$ ).

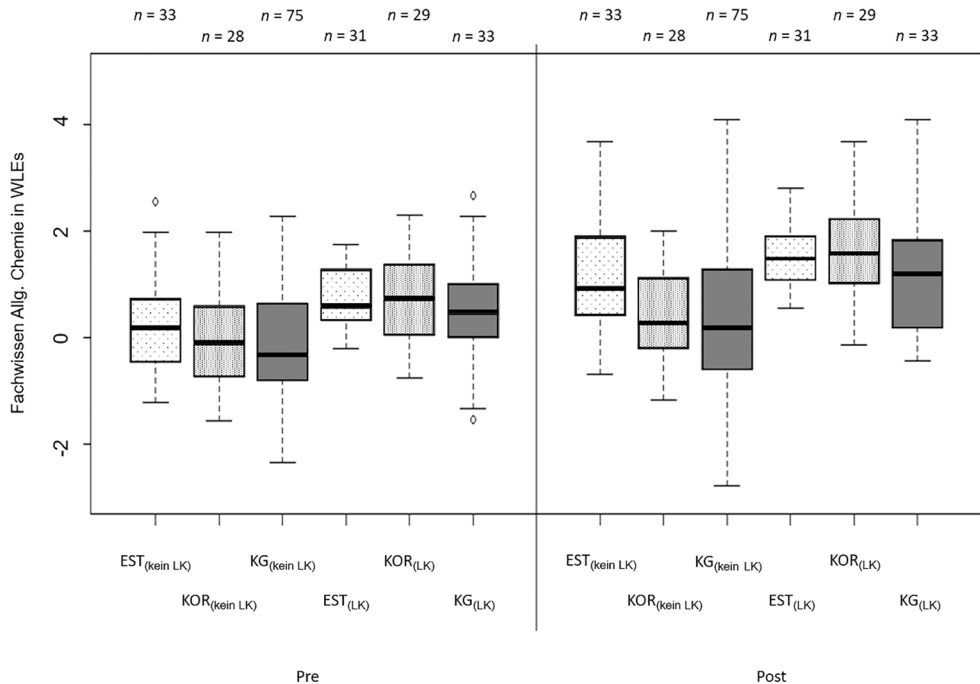


Abbildung 4: Entwicklung des Chemie Fachwissens nach Interventionsgruppe (Kohortenvergleich)

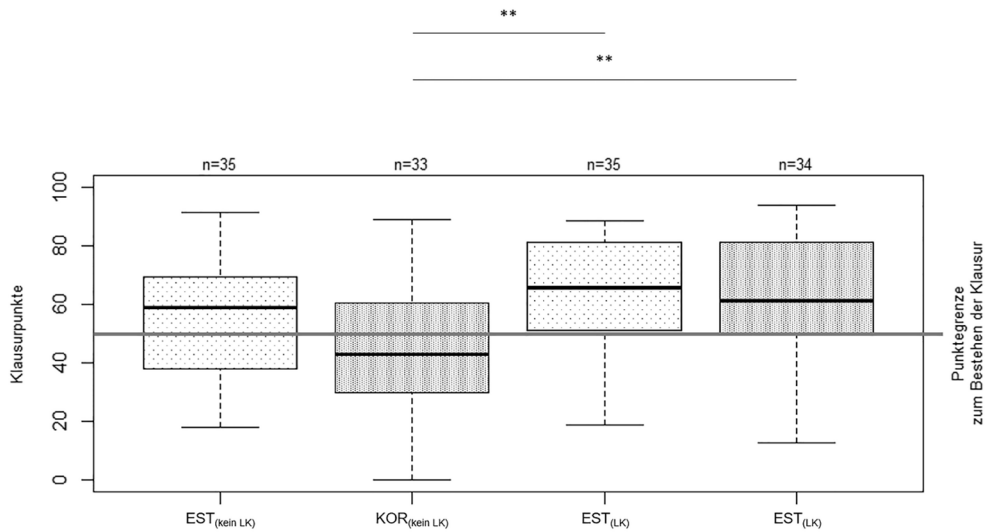
## 6.2 Unterschiede in den Klausurleistungen nach Interventionsgruppe

Während die Klausuren über die Pilotierung und Hauptstudie 1 in Absprache mit dem Dozenten am ersten und zweiten Versuchstermin identisch gehalten werden konnten, musste aufgrund der Pandemie im WiSe 2020 das Klausurformat kurzfristig geändert werden. Daher sind die Ergebnisse zum Klausurerfolg beschränkt auf die WiSe 2018 und 2019 ( $N_{Post} = 93$ ). Aufgrund einer universitätsinternen Freiversuchsregelung wird der Klausurerfolg als die beste Punktzahl angesehen, die über die ersten beiden Klausur-Termine je Proband/in erreicht wurde. Durch mindestens 50 von 100 erzielten Punkten bei einer Teilnahme an mindestens einem der zwei Klausur-Termine, die in der vorlesungsfreien Zeit nach Semesterende innerhalb von vier Wochen zur Verfügung standen, konnte die Klausur bestanden werden. Über eine einfaktorielle ANOVA ( $F(3, 89) = 5.384, p \leq .01, \eta^2 = .154$ ) konnten für den Klausurerfolg im post-hoc Test (Tukey) signifikante Unterschiede zwischen den Interventionsgruppen ermittelt werden (s. Abbildung 5).

Dabei blieben Studierende der Gruppe mit KOR-Feedback ohne Leistungskurs ( $M = 39.36, SD = 20.20$ ) signifikant hinter den Studierenden mit Leistungskurswahl zurück, unabhängig davon, welches Feedback sie bekommen haben. Zwischen EST- ( $M = 60.52, SD = 17.44$ ) und KOR-Feedback ( $M = 61.00, SD = 22.23$ ) findet sich kein Unterschied, wenn vormals ein Leistungskurs besucht wurde. Da sich zwischen der Gruppe EST<sub>(kein LK)</sub> und KOR<sub>(kein LK)</sub> keine Unterschiede hinsichtlich des Fachwissens (Pre) und mathematischen und kognitiven Fähigkeiten finden ließen, profitieren Studierende der Gruppe ohne Leistungskurs

am meisten vom EST-Feedback ( $M = 55.61$ ,  $SD = 21.98$ ), da sie hinsichtlich des Klausurerfolgs nicht von beiden Gruppen mit Leistungskurs hintergrund unterscheidbar sind und sich zudem ein Trend hinsichtlich eines Vorsprungs gegenüber der gleich beschulten Gruppe, die mit KOR-Feedback gelernt hat, abzeichnet.

Abbildung 5: Klausurpunkte in Abhängigkeit von der Interventionsgruppe WiSe 2018 und 2019 (\*\*  $p \leq .01$ )



## 7 Diskussion und Fazit

Die Ergebnisse deuten grundsätzlich auf einen Vorteil des EST-Feedbacks hin, besonders scheinen aber Studierende mit geringerem Vorwissen vom EST-Feedback zu profitieren. Betrachtet man die Chemie-Kurswahl in der Oberstufe als zusätzliches Maß für das Vorwissen der Studierenden, so zeigt sich hinsichtlich des Klausurerfolgs für Studierende der Gruppe EST<sub>(kein LK)</sub> eine Performance, die nicht mehr von der der Studierenden mit früherer Leistungskurswahl (EST<sub>(LK)</sub> und KOR<sub>(LK)</sub>) zu unterscheiden ist, während Studierende der Gruppe KOR<sub>(kein LK)</sub> signifikant zurückbleiben und im Mittel mit knapp 39.5 Punkten nicht einmal die Klausur bestehen. Hier deutet sich also bereits konform zur Hypothese 1 ein Vorteil von EST-Feedback für Studierende mit geringem Vorwissen an.

Zwischen EST<sub>(LK)</sub> und KOR<sub>(LK)</sub> gibt es allerdings keinen Unterschied in den Klausurleistungen, weshalb es für Hypothese 2, also einen Nachteil für Vorwissensstarke resultierend aus EST-Feedback, zunächst einmal keine bestätigenden Indizien gibt. Betrachtet man mit Modell 2 nun die Regression zur Vorhersage des Fachwissens zum Post-Messzeitpunkt unter Kontrolle des Fachwissens (Pre), dann zeigt sich der Vorteil für EST-Feedback nur für Studierende ohne Leistungskurs (EST<sub>(kein LK)</sub>). Wurde hingegen ein Leistungskurs Chemie in der Oberstufe belegt und mit EST-Feedback gelernt (EST<sub>(LK)</sub>), so äußert sich in der Interaktion ein schwacher Expertise Reversal Effekt, der den Vorteil des EST-Feedbacks nahezu aufhebt. Auf Basis dessen war für Studierende der Gruppe KOR<sub>(LK)</sub> ein Vorteil für das Lernen mit

KOR-Feedback erwartet worden. Der Interaktionsterm aus Kurswahl und KOR-Feedback weist zwar konform zu Hypothese 2 ein positives Vorzeichen auf, doch ist er weder signifikant noch von nennenswerter Effektstärke ( $\beta = .004, p = .170$ ). Das Modell 3, das auf Basis dessen zur Überprüfung nur eine Interaktion für EST-Feedback annimmt, führt zudem zu einer signifikant besseren Varianzaufklärung.

Eine direkte Interaktion zwischen Feedback-Typ und den harten Vorwissensdaten aus dem Fachwissenstest ließ sich hingegen in keiner Analyse zeigen. Der Fachwissenstest von Averbeck (2020) besteht allerdings zu einem Drittel aus Items, die Faktenwissen, und zu zwei Dritteln aus Items, die Zusammenhänge abfragen, und von der inhaltlichen Komponente scheinen die 35 Items auch überwiegend Aspekte zu prüfen, die grundlegend für die Allgemeine Chemie sind und sich in den Lehrplänen von Grundkurs und Leistungskurs wiederfinden. Der Test weist gute Kennwerte hinsichtlich der Gütekriterien auf. Fraglich ist aber, ob er den gesamten Vorteil abbilden kann, der sich durch einen Leistungskurs ermöglicht.

Auf der Internetseite des Landesinstituts für Schule (QUA-LiS NRW) findet sich folgender Ausschnitt:

„[Als Bruttounterrichtszeit ergeben sich] ... für die Einführungsphase 90 Unterrichtsstunden, für den Grundkurs in der Q1 ebenfalls 90 und in der Q2 60 Stunden und für den Leistungskurs in der Q1 150 und für Q2 90 Unterrichtsstunden“ (QUA-LiS NRW, 2022).

Demnach ergibt sich für einen Chemie-Leistungskurs ein Zugewinn an Lernzeit, welcher der gesamten Lernzeit eines Grundkurses in der Q1 oder Einführungsphase entspricht. Im Rahmen eines Leistungskurses besteht folglich viel mehr Zeit, durch kumulatives Lernen ein Wissensnetz aufzubauen. Die geringere Lernzeit bietet im Grundkurs deutlich weniger Möglichkeiten, Zusammenhänge im gleichen Umfang wie in einem Leistungskurs zu vermitteln. Das EST-Feedback präsentiert aber gerade diese Zusammenhänge anwendungsbezogen u.a. in Form von Lösungsbeispielen. Da sich diesbezüglich aber nur für Studierende ein signifikanter Vorteil zeigt, die früher einen Grundkurs Chemie belegt hatten, ist es möglich, dass die Aufgaben mit EST-Feedback dabei helfen, bei gleicher Lernzeit fehlendes Zusammenhangswissen aufzuholen beziehungsweise vorhandenes (Insel-)Wissen miteinander zu verknüpfen, was sich in einer höheren Test- und Klausurperformance bei Studierenden der Gruppe  $EST_{(\text{kein LK})}$  – verglichen mit anderen Studierenden mit einem Grundkurs in der Oberstufe – zeigt.

Einigt man sich darauf, das Konstrukt Vorwissen als Kombination aus dem Fachwissenstest (Pre) und der Chemie-Kurswahl in der Oberstufe zu beschreiben, dann lässt sich Hypothese 1 bestätigen, denn es scheinen laut Datenlage besonders Studierende ohne frühere Leistungskurswahl von den Aufgaben mit EST-Feedback zu profitieren. Wenngleich sich ein schwindender Vorteil der Unterstützung in der Interaktion aus EST-Feedback und Leistungskurswahl für  $EST_{(\text{LK})}$  zeigt, lassen sich keine Indizien für Hypothese 2 finden, weshalb sie verworfen werden muss.

In einem klassischen Übungsbetrieb (vgl.  $KG_{(\text{kein LK})}$  und  $KG_{(\text{LK})}$ ) bietet es sich also an, diese Lernaufgaben mit EST-Feedback als zusätzliches Lernangebot und damit als binnendifferenzierendes Förderangebot einzusetzen, da aktives Lernen mit den Aufgaben zu einem signifikant besseren Fachwissensstand und Klausurerfolg am Ende des ersten Semesters führt. Damit liefert diese Studie hinsichtlich elaborierten Feedbacks vergleichbare Ergebnisse wie Fyfe et al. (2012). Die sehr komplexen Aufgaben, die die Allgemeine Chemie bereithält, können allerdings der Grund sein, warum Studierende der Gruppe  $EST_{(\text{LK})}$  keine

Nachteile durch Lernen mit dem EST-Feedback erfahren. Daraus ergibt sich für die Studieneingangsphase eine spezifische Handlungsempfehlung für das Fach Chemie, etwaige digitale Lernmodule, die elaboriertes Feedback bereitstellen, einzurichten beziehungsweise (falls vorhanden) systematisch auszubauen. Ob sich durch eine systematische Intervention mit solchen digitalen Lernmodulen langfristig Abbruchquoten reduzieren lassen, gilt es allerdings noch zu prüfen.

Im Rahmen der Feedback-Forschung liefert diese Arbeit einen (vergleichsweise selten anzutreffenden) systematischen Vergleich zweier Feedback-Typen. Naturwissenschaftliche Fächer und insbesondere die Chemie als Forschungsbereich sind allerdings in der Feedback-Forschung unterrepräsentiert beziehungsweise hinsichtlich der Beforschung der Wirksamkeit von automatisiertem Feedback nach Cavalcanti et al. (2021) nicht vorhanden. Allerdings konnte auch diese Arbeit zeigen, dass automatisiertes (elaboriertes) Feedback grundsätzlich einen positiven Effekt auf die Performance hat. Angelehnt an den positiven Trend, der sich bei 32 der von Cavalcanti et al. (2021) ermittelten Artikel abzeichnet, liefert diese Arbeit damit nun auch einen empirischen Nachweis für positive Effekte automatisierten Feedbacks auf die Performance von Nutzerinnen und Nutzern für die Chemie und damit ebenfalls einen guten Ansatzpunkt für etwaige Replikationsstudien.

## Anmerkung

Dieses Projekt wurde gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 397641476.

## Literatur

- Albacete, P. L. & VanLehn, K. A. (2000). Evaluating the effectiveness of a cognitive tutor for fundamental physics concepts. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 22. Verfügbar unter <https://escholarship.org/uc/item/0166b7p0>
- Ashton, H. S., Beevers, C. E., Korabinski, A. A. & Youngson, M. A. (2005). Investigating the medium effect in computer-aided assessment of school chemistry and college computing national examinations. *British Journal of Educational Technology*, 36(5), 771–787. <https://doi.org/10.1111/j.1467-8535.2005.00501.x>
- Averbeck, D. (2020). *Zum Studienerfolg in der Studieneingangsphase des Chemiestudiums* (Studien zum Physik- und Chemielernen, Bd. 308). Logos.
- Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A. & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, 61(2), 213–238. <https://doi.org/10.3102/00346543061002213>
- Bond, T. G., Yan, Z. & Heene, M. (2020). *Applying the Rasch model. Fundamental measurement in the human sciences* (Fourth edition). Routledge. <https://doi.org/10.4324/9780429030499>
- Butler, D. L. & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3), 245–281.
- Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D. et al. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2, 100027. <https://doi.org/10.1016/j.caeai.2021.100027>

- Chen, O., Kalyuga, S. & Sweller, J. (2017). The Expertise Reversal Effect is a variant of the more general Element Interactivity Effect. *Educational Psychology Review*, 29(2), 393–405. <https://doi.org/10.1007/s10648-016-9359-1>
- Dickmann, T. (2019). *Visuelles Modellverständnis und Studienerfolg in der Chemie. Zwei Seiten einer Medaille* (Studien zum Physik- und Chemielernen). Logos.
- Eitemüller, C., Trauten, F. Striewe, M. & Walpuski, M. (2023). Digitalization of multistep chemistry exercises with automated formative feedback. *Journal of Science Education and Technology*, 32, 453–467. <https://doi.org/10.1007/s10956-023-10043-2>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.
- Freyer, K. (2013). *Zum Einfluss von Studieneingangsvoraussetzungen auf den Studienerfolg Erstsemestrierender im Fach Chemie* (Studien zum Physik- und Chemielernen, Bd. 156). Logos.
- Fyfe, E. R., Rittle-Johnson, B. & DeCaro, M. S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *Journal of Educational Psychology*, 104(4), 1094–1108. <https://doi.org/10.1037/a0028389>
- Goldhausen, I. (2015). *Mathematische Modelle im Chemieunterricht*. Dissertation, Universität Kassel. Uni-Edition GmbH.
- Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Heller, K. A. & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4-12+R)*. Göttingen: Beltz Test.
- Heublein, U., Ebert, J., Hutzsch, C., Isleib, S., König, R., Richter, J. et al. (2017). *Zwischen Studienerwartungen und Studienwirklichkeit. Ursachen des Studienabbruchs, beruflicher Verbleib der Studienabbrecherinnen und Studienabbrecher und Entwicklung der Studienabbruchquote an deutschen Hochschulen* (Forum Hochschule). Deutsches Zentrum für Hochschul- und Wissenschaftsforschung.
- Heublein, U., Richter, J. & Schmelzer, R. (2020). *Die Entwicklung der Studienabbruchquoten in Deutschland*. [https://doi.org/10.34878/2020.03.DZHW\\_BRIEF](https://doi.org/10.34878/2020.03.DZHW_BRIEF)
- Johnson, C. I. & Priest, H. A. (2014). The feedback principle in multimedia learning. In R. E. Mayer (Hrsg.), *The Cambridge Handbook of Multimedia Learning* (S. 449–463). Cambridge University Press.
- Kalyuga, S. (2014). The expertise reversal principle in multimedia learning. In R. E. Mayer (Hrsg.), *The Cambridge Handbook of Multimedia Learning* (S. 576–597). Cambridge University Press. <https://doi.org/10.1017/CBO9781139547369.028>
- Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>
- Krause, U.-M., Stark, R. & Mandl, H. (2009). The effects of cooperative learning and feedback on e-learning in statistics. *Learning and Instruction*, 19(2), 158–170. <https://doi.org/10.1016/j.learninstruc.2008.03.003>
- Mayer, R. E. & Johnson, C. I. (2010). Adding instructional features that promote learning in a game-like environment. *Journal of Educational Computing Research*, 42(3), 241–265. <https://doi.org/10.2190/EC.42.3.a>
- Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science*, 32(1/2), 99–113. <https://doi.org/10.1023/B:TRUC.0000021811.66966.1d>
- Moreno, R. & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review*, 19(3), 309–326. <https://doi.org/10.1007/s10648-007-9047-2>

- Müller, J., Stender, A., Fleischer, J., Borowski, A., Dammann, E., Lang, M. & Fischer, H. E. (2018). Mathematisches Wissen von Studienanfängern und Studienerfolg. *Zeitschrift für Didaktik der Naturwissenschaften*, 24, 183–199.
- Narciss, S. (2006). *Informatives tutorielles Feedback. Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 56). Waxmann.
- Narciss, S. (2014). Modelle zu den Bedingungen und Wirkungen von Feedback in Lehr-Lernsituationen. In H. Ditton (Hrsg.), *Feedback und Rückmeldungen. Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (S. 43–82). Waxmann.
- Narciss, S. (2018). Feedbackstrategien für interaktive Lernaufgaben. In H. M. Niegemann & A. Weinberger (Hrsg.), *Handbuch Bildungstechnologie. Konzeption und Einsatz digitaler Lernumgebungen* (Springer reference, S. 1–24). Springer. [https://doi.org/10.1007/978-3-662-54373-3\\_35-1](https://doi.org/10.1007/978-3-662-54373-3_35-1)
- Narciss, S. & Huth, K. (2004). Fostering achievement and motivation with bug-related tutoring feedback in a computer-based training for written subtraction. *Learning and Instruction*, 16(4), 310–322. <https://doi.org/10.1016/j.learninstruc.2006.07.003>
- Pobel, S. & Striwe, M. (2019). Domain-specific extensions for an e-assessment system. In M. A. Herzog, Z. Kubincová, P. Han & M. Temperini (Hrsg.), *Advances in Web-Based Learning – ICWL 2019* (Bd. 11841, S. 327–331). Springer International Publishing.
- QUA-LiS NRW (Rüdiger Käuser, Hrsg.). (2022). *Schulentwicklung NRW – Lehrplannavigator S II – Gymnasiale Oberstufe – Chemie – Hinweise und Beispiele – Beispiel eines schulinternen Lehrplans für die gymnasiale Oberstufe im Fach Chemie*. Schulentwicklung NRW. <https://www.schulentwicklung.nrw.de/lehrplaene/lehrplannavigator-s-ii/gymnasiale-oberstufe/chemie/hinweise-und-beispiele/schulinterner-lehrplan/schulinterner-lehrplan.html>
- Rost, F. (2018). *Lern- und Arbeitstechniken für das Studium* (Lehrbuch, 8., vollständig überarbeitete und aktualisierte Auflage). Springer VS.
- Sarcelletti, A. & Müller, S. (2011). Zum Stand der Studienabbruchforschung. Theoretische Perspektiven, zentrale Ergebnisse und methodische Anforderungen an künftige Studien. *Zeitschrift für Bildungsforschung*, 1(3), 235–248. <https://doi.org/10.1007/s35834-011-0020-2>
- Schuhmacher, R. (2008). Der produktive Umgang mit Fehlern. In R. Caspary (Hrsg.), *Nur wer Fehler macht, kommt weiter* (S. 49–72). Herder.
- Smits, M. H., Boon, J., Sluijsmans, D. M. & van Gog, T. (2008). Content and timing of feedback in a web-based learning environment: effects on learning as a function of prior knowledge. *Interactive Learning Environments*, 16(2), 183–193. <https://doi.org/10.1080/10494820701365952>
- Striwe, M. (2016). An architecture for modular grading and feedback generation for complex exercises. *Science of Computer Programming*, 129, 35–47. <https://doi.org/10.1016/j.scico.2016.02.009>
- Striwe, M., Trauten, F. & Eitemüller, C. (2020). Aufgaben mit automatischem Feedback zu chemischen Atom-Orbitalmodellen. In R. Zender, D. Ifenthaler, T. Leonhardt & C. Schumacher (Hrsg.), *DELFI 2020. Die 18. Fachtagung Bildungstechnologien der Gesellschaft für Informatik e.V., Fachtagung vom 14.-18. September 2020 Online* (GI-Edition. Proceedings, Bd. 308, S. 109–119). Gesellschaft für Informatik.
- Tinto, V. (1988). Stages of student departure: Reflections on the longitudinal character of student leaving. *Journal of Higher Education*, 59(4), 438. <https://doi.org/10.2307/1981920>
- Trautwein, C. & Bosse, E. (2017). The first year in higher education – critical requirements from the student perspective. *Higher Education*, 73(3), 371–387. <https://doi.org/10.1007/s10734-016-0098-5>
- Walter, O. (2005). *Kompetenzmessung in den PISA-Studien. Simulationen zur Schätzung von Verteilungsparametern und Reliabilitäten*. Pabst.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427–450. <https://doi.org/10.1007/BF02294627>

Wisniewski, B., Zierer, K. & Hattie, J. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 1–14. <https://doi.org/10.3389/fpsyg.2019.03087>

### *Kontakt*

Florian Trauten · Dr. Carolin Eitemüller · Prof. Dr. Maik Walpuski  
Universität Duisburg-Essen  
Didaktik der Chemie  
Schützenbahn 70  
45127 Essen  
E-Mail: [florian.trauten@uni-due.de](mailto:florian.trauten@uni-due.de)  
E-Mail: [carolin.eitemueller@uni-due.de](mailto:carolin.eitemueller@uni-due.de)  
E-Mail: [maik.walpuski@uni-due.de](mailto:maik.walpuski@uni-due.de)