Z Erziehungswiss (2017) 20:728–747
DOI 10.1007/s11618-017-0776-9

ZfE CrossMark

ALLGEMEINER TEIL

Formative Leistungsdiagnostik und *Learning Analytics*: Entwicklung, Nutzung und Optimierung eines onlinebasierten Kurses für die Diagnostik und Förderung von Grundwissen im Kompetenzbereich Sprachbetrachtung

Uwe Maier · Carolin Ramsteck · Kathrin Hoffmann

Online publiziert: 30. August 2017

Zusammenfassung *Learning analytics* ist ein neues, im deutschsprachigen Raum kaum rezipiertes Forschungsfeld, in dem automatisch gespeicherte Daten einer Lernplattform analysiert werden, um das Lernangebot zu optimieren sowie Rückmeldungen für Lehrende und Lernende zu generieren. Bisher liegen hierzu lediglich Untersuchungen aus dem universitären Bereich vor. Ebenso gibt es nur wenige Studien, in denen eine theoriebasierte Klassifikation der Daten erfolgte. Ziel dieser Studie ist die Entwicklung und Anwendung eines Klassifikationssystems für Log-Daten und Kursbewertungen aus einem Moodle-Kurs für die formative Diagnostik von grammatikalischem Grundwissen (fünf Klassen, Sekundarstufe I, 129 Schülerinnen und Schüler). Mit der Analyse der kodierten Log-Daten und deren Häufigkeiten konnte ein differenziertes Bild der Kursnutzung gezeichnet werden. Ebenso ergaben sich vielfältige Hinweise für die Optimierung des Lernangebots. Allerdings erklärt nur ein Teil der Interaktionskategorien die im Kurs erfassten Lernzuwächse. Abschließend werden Implikationen für die Weiterentwicklung von *learning analytics*-Methoden für die Erforschung computergestützter, formativer Diagnostik diskutiert.

Schlüsselwörter Computerunterstützter Unterricht · Deutschunterricht · Leistungsdiagnostik · Formative Evaluation

Prof. Dr. U. Maier (✉) · C. Ramsteck · K. Hoffmann
Institut für Erziehungswissenschaft, Pädagogische Hochschule Schwäbisch Gmünd, Oberbettringerstraße 200, 73525 Schwäbisch Gmünd, Deutschland
E-Mail: uwe.maier@ph-gmuend.de

C. Ramsteck
E-Mail: carolin.ramsteck@ph-gmuend.de

K. Hoffmann
E-Mail: kathrin.hoffmann@ph-gmuend.de

Formative assessment and learning analytics: development usage and optimization of an online course for the diagnostics and promotion of basic knowledge in the area of language

Abstract *Learning analytics* is a young field of educational research and aims to analyze data from learning management systems in order to improve digital learning scenarios or to generate feedback to learners and instructors. To date, studies in this field of research are limited to higher education and there are only few examples of theory-based classifications of learner-computer interactions. This paper describes the development and application of a theory-based classification of log-files and course statistics for a computer-assisted formative assessment of basic German language knowledge (five secondary classrooms, lower secondary level, 129 students). The analysis of interaction frequencies yields several possibilities for improving the course design. Only some interaction categories explain learning progressions within the formative assessment course. Finally, implications for developing learning *analytics* methods for the analysis and improvement of computer-assisted formative assessment are discussed.

Keywords Blended-learning · Computer-assisted testing · Curriculum-based assessment · Formative assessment · Individualized instruction

1 Einleitung

Heterogene Lerngruppen sind eine zunehmend größere Herausforderung für Lehrkräfte in der Sekundarstufe. Um einen an den Lernvoraussetzungen der Schülerinnen und Schüler orientierten Unterricht realisieren zu können, benötigen Lehrkräfte Verfahren der formativen Diagnostik mit daran anschließenden Förderoptionen. Die Entwicklung digitaler Unterrichtstechnologien eröffnet hierfür neue Möglichkeiten. Dieser Artikel nimmt den schulischen Einsatz der Lernplattform Moodle für die formative Diagnostik in den Blick. Die Nutzung eines digitalen Lernangebots wird anhand von Kursbewertungen und Log-Daten rekonstruiert. Die Studie knüpft damit an zwei Forschungsstränge an. Einmal wird auf Konzepte der formativen Lernverlaufsdiagnostik zurückgegriffen, um eine computergestützte, formative Diagnostik für Grundwissen im Bereich Sprachbetrachtung zu entwickeln. Ein zweiter Anknüpfungspunkt sind Studien aus dem Bereich *learning analytics*, in denen Methoden zur Analyse von Lernplattformnutzerdaten für die Optimierung von Lernangeboten erprobt werden.

2 Theorie und Forschungsstand

2.1 Formative Lernverlaufsdiagnostik

Unter formativer Lernverlaufsdiagnostik versteht man die wiederholte Messung von Schülerleistungen sowie die Rückmeldung der individuellen Lernverläufe an Leh-

rende und Lernende (Strathmann und Klauer 2010; Maier 2014). Ziel einer Lernverlaufsdiagnostik ist die Adaption von Unterricht oder die Förderung einzelner Schülerinnen und Schüler.

Deutschsprachige Studien zu formativer Lernverlaufsdiagnostik beziehen sich auf die US-amerikanisch geprägte Forschungslinie zu *curriculum-based measurement* (z. B. Topping und Fisher 2003; Fuchs 2004; Nunnery et al. 2006; Yeh 2010). Im Review von Stecker et al. (2005) werden Effekte von *curriculum-based measurement* sowohl für *special education* als auch für *general education* als bedeutsam eingeschätzt. Allerdings spielen jeweils unterschiedliche Implementationsfaktoren eine Rolle. Beispielsweise sind Verfahren formativer Diagnostik dann besonders effektiv, wenn sich Lehrkräfte mit dem Leistungsprofil der ganzen Klasse auseinandersetzen und im Anschluss daran kooperative Lernstrategien etablieren. Deutschsprachige Beispiele für Lernverlaufsdiagnosen sind die Lernfortschrittsdiagnostik Lesen (Walter 2011), die Verlaufsdiagnostik für das sinnerfassende Lesen (Walter 2013), die Lernverlaufsdiagnostik Grundrechenarten (Strathmann und Klauer 2010) oder die internetbasierte Lernverlaufsdiagnostik „quop“ für Leseverständnis und Grundrechenarten (Souvignier et al. 2014). Kernstück all dieser Verfahren sind gleich schwierige Paralleltests, die eine Messung von Lernzuwächsen bei überwiegend prozeduralisierten Grundfertigkeiten ermöglichen.

Das Konzept *mastery assessment* (Alternative Begriffe: *mastery tests, criterion-referenced tests)* hat große Ähnlichkeit mit der Lernverlaufsdiagnostik, gilt ebenfalls als effizient und geht zurück auf Studien zum *mastery learning* (Bloom 1974; Kulik et al. 1990; Guskey 2015). Beide Konzepte unterscheiden sich vor allem im Hinblick auf die Konstruktion und Validierung der Tests (Zimmerman und Dibenedetto 2008). Beim *curriculum-based measurement* wird ein Konstrukt mit möglichst gleich schweren Paralleltests über mehrere Messzeitpunkte hinweg erfasst. Deshalb werden basale Prozeduren wie Lesen oder Rechenfertigkeiten, die über einen längeren Zeitraum geschult werden müssen, untersucht.

Der Einsatz von kurzen Paralleltests ist allerdings nur in bestimmten Lerndomänen oder bei bestimmten Kompetenzfacetten möglich. Die Leseflüssigkeit ist hierfür ein Paradebeispiel. Sie lässt sich sehr gut und isoliert mit kurzen und dennoch validen Paralleltests prüfen und spielt über einen langen Zeitraum für die Lesekompetenzentwicklung eine wichtige Rolle. In vielen Lerndomänen, vor allem ab der Sekundarstufe 1, wird diese Vorgehensweise zunehmend erschwert. Sowohl in den sprachlichen als auch in den mathematisch-naturwissenschaftlichen Fächern zeichnen sich Lerndomänen durch ein vielfältig vernetztes deklaratives und prozedurales Wissen aus. Bei der Abbildung des Lernfortschritts durch formative Tests muss dabei die inhaltliche Progression berücksichtigt werden. In *mastery learning*-Programmen wird diesem Umstand Rechnung getragen. Es findet zunächst eine curriculare Analyse der erwünschten Lernprogression statt. Danach werden *mastery tests* entwickelt, die prüfen, ob eine Schülerin oder ein Schüler eine bestimmte Fähigkeit oder einen bestimmten Wissensbereich auf der jeweiligen Stufe der Lernprogression beherrscht oder nicht. Als *passing score* hat sich der Wert von 80 % etabliert (Zimmerman und Dibenedetto 2008).

2.2 *Learning analytics*: Auswertung von Log-Daten in Lernplattformen

In dem relativ neuen Forschungsbereich *learning analytics* geht es um die Auswertung von Nutzerdaten und Log-files aus Lernplattformen, um Online-Kurse oder Blended-Learning-Angebote zu optimieren. Auf der Homepage der ersten Learning Analytics Tagung (2011) findet sich folgende Definition: „Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs."

Die überwiegende Mehrheit der Studien bezieht sich allerdings auf universitäre Lehrangebote. Am gängigsten ist die Auswertung von Nutzerdaten und Log-files in Standard-Lernplattformen wie Moodle oder Blackboard. Die Forschungsdesigns sind in der Regel nicht experimentell angelegt, sondern *post hoc*-Analysen, um Prädiktorvariablen für den Lernerfolg im Kurs zu ermitteln und letztendlich Schlussfolgerungen für eine Kursoptimierung abzuleiten. Dabei werden unterschiedliche Analyse- und Datenauswertungsmethoden verwendet. Xing et al. (2015) unterscheiden zwischen *black box*-Modellen und *white box*-Modellen. In *black box*-Modellen lernen neuronale Netzwerke anhand von großen Mengen an Nutzeraktivitätsprofilen, welches Aktivitätsmuster mit hohem Output (Testleistung in einem Kurs) zusammenhängt. Das neuronale Netzwerk kann zwar ein erfolgreiches von einem nicht erfolgreichen Nutzeraktivitätsprofil mit großer Wahrscheinlichkeit unterscheiden. Allerdings können mit diesem Verfahren keine Prädiktoren bestimmt werden, die zu einer Optimierung von digitalen Lernangeboten führen.

In sog. *white box*-Modellen werden einzelne Regeln oder Prädiktoren gezielt im Hinblick auf ihre Vorhersagekraft getestet (z. B. Iglesias-Pradas et al. 2015). In vielen Studien werden multiple Regressionsanalysen oder multiple Korrelationen für einzelne Aktivitäten in der Lernplattform berechnet (z. B. Macfadyen und Dawson 2010; Tempelaar et al. 2015). Sehr komplexe Interaktionszusammenhänge können eventuell nicht aufgedeckt werden, allerdings lassen sich einzelne Aktivitäten ganz klar im Hinblick auf ihre prädiktive Kraft differenzieren. Beispielsweise spielen rein quantifizierende Indikatoren (Zeit, die Studierende mit einer Lernplattform verbringen, Anzahl der Klicks, Downloadraten von Lernmaterialien) eine geringe Rolle, um die abschließende Kursleistung vorherzusagen. Sehr hohe Zusammenhänge gibt es dagegen zwischen der abschließenden Kursleistung und Indikatoren für eine aktive Mitarbeit im Kurs (z. B. formative Tests, Anzahl von Diskussionsbeiträgen im Forum, der Anzahl erledigter Aufgabenstellungen oder der Anzahl von Mails an Mitstudierende bzw. Lehrende).

Learning-analytics-Studien werden von einigen Autoren als atheoretisch kritisiert, weil die Ansätze induktiv sind und die Auswahl der untersuchten Variablen sich nicht an instruktionalen Theorien orientiert (z. B. Clow 2013). Um diesem Theoriedefizit entgegenzuwirken und eine gewisse Vergleichbarkeit zwischen verschiedenen Studien zu ermöglichen, schlagen Agudo-Peregrina et al. (2014) eine Klassifikation für LMS (*Learning Management System*)-Interaktionsdaten im Hinblick auf Interaktionsagenten (Lehrer-Schüler, Schüler-Schüler, Schüler-Lerninhalt), Häufigkeit und Partizipation (aktiv vs. passiv) vor. Agudo-Peregrina et al. (2014) klassifizieren die Interaktionsdaten aus verschiedenen universitären Kursen und prüfen die prädiktive

Kraft der einzelnen Dimensionen für die Erklärung der abschließenden Kursleistung. Bei der ersten Dimension sind vor allem Interaktionen mit anderen Studierenden und den Lehrenden relevant. Ebenso Interaktionen, die entlang der dritten Dimension als aktive Interaktionen klassifiziert wurden (z. B. Aufgabenstellungen und formative Tests). Allerdings konnten Agudo-Peregrina et al. (2014) diese Zusammenhänge nur bei Online-Kursen finden, jedoch nicht in Lehrveranstaltungen mit Lernplattformen als Ergänzung (Blended-Learning-Szenarien).

In einer Studie von You (2016) wird verstärkt die Bedeutung des selbstregulierten Lernens in einem universitären Online-Kurs in den Blick genommen und es wird ein über die bisher praktizierten Verfahren hinausgehendes Klassifikationssystem für LMS-Log-Daten erprobt. Anhand von Log-Daten bewerteten *human raters* die Regelmäßigkeit und die Vollständigkeit der Nutzung von Lehrvideos. Indikatoren für selbstgesteuertes Lernen, wie die rechtzeitige Abgabe von Aufgaben oder die Bestätigung des Lesens von Kursinformationen, waren relevante Prädiktoren zur Vorhersage der Kursleistung. Rein quantifizierende Indikatoren, wie die Gesamtdauer der Betrachtung von Lehrvideos oder die Anzahl geschriebener Nachrichten, hatten dagegen keine Relevanz für die Erklärung der Kursleistung.

Eine ähnlich differenzierte, jedoch automatisierte Auswertung von Kursinteraktionen publizierten Ruipérez-Valiente et al. (2015). Sie entwickelten Lernindikatoren aus den Log-Daten der Lernplattform *Khan Academy* (Ereignisse, Zeitmarken usw.), u. a. für die Gesamtnutzung der Lernplattform (Anzahl von Videos und Übungen usw.), den Lernfortschritt (Zunahme bei korrekten Übungen, Effizienz beim Bearbeiten der Übungen, Fortschritt und Effizienz beim Betrachten der Videos), die zeitliche Verteilung der Plattformnutzung (tägliche Nutzungsintervalle) oder die Übungsgewohnheit (welche Übungen zu absolvieren sind; Nutzung von Lernhinweisen; doppelte Bearbeitung von Übungen). In dieser Studie werden allerdings keine Zusammenhänge zwischen den klassifizierten Indikatoren und den Kursergebnissen berechnet. Es wird lediglich dargestellt, welches Potenzial die Indikatoren für die Analyse der Lernstrategien und die Gestaltung von Rückmeldungen haben.

3 Zielsetzung des Projektes und Forschungsfragen

In dieser Studie soll erkundet werden, inwiefern sich Methoden aus *learning analytics*-Studien für die Evaluation und Optimierung einer formativen Diagnostik mit der Lernplattform Moodle heranziehen lassen. Ein Forschungsdefizit bisheriger *learning analytics*-Studien ist, dass relativ wahllos die verfügbaren Kurs- und Log-Daten im Hinblick auf ihre prädiktive Kraft zur Vorhersage der Kursleistung herangezogen werden. Ebenso sind die bisherigen Befunde aus dem universitären E-Learning-Bereich wenig aussagekräftig für die speziellen Anforderungen an formative Lernverlaufsdiagnosen in der Sekundarstufe I. Anknüpfungspunkte für diese Studie bieten die Klassifikationssysteme von You (2016) sowie Ruipérez-Valiente et al. (2015), in denen Interaktionssequenzen kategorisiert wurden, die konzeptionellen Überlegungen zum Kursdesign folgen.

Eine erste Zielsetzung der Studie ist deshalb die Entwicklung einer Klassifikation von systematischen und nicht systematischen Mustern der Nutzung einer formativen

Lernverlaufsdiagnostik. Von systematischer Nutzung soll dann gesprochen werden, wenn eine Schülerin oder ein Schüler sowohl die formativen Tests als auch die weiterführenden Lernmaterialien im Sinne der Prinzipien formativer Diagnostik nutzt. Ein Prinzip ist die zeitliche Nähe zwischen formativem Test, Rückmeldung und auf die Rückmeldung bezogener Weiterarbeit. Ein weiteres Prinzip, v. a. des *mastery assessment*, ist die Wiederholung des formativen Tests oder der Übungen bis zur sicheren Beherrschung des Wissens. Die in der Lernplattform möglichen Interaktionen und Interaktionsfolgen sollen vor dem Hintergrund dieser Prinzipien als systematische oder nicht systematische Nutzung klassifiziert werden.

Eine zweite Zielsetzung ist die Untersuchung von Zusammenhängen zwischen Nutzungsmustern und Lernzuwächsen. Das in einem ersten Schritt zu entwickelnde Klassifikationssystem ist hierfür die methodische Grundlage für die Datenauswertung. Damit sollte sowohl eine Quantifizierung von systematischer oder unsystematischer Nutzung möglich sein als auch eine Quantifizierung des Lernfortschritts innerhalb des Kurses. Auf Basis dieser Auswertungen sollen folgende Fragen beantwortet werden:

1. In welchem Verhältnis stehen systematische und unsystematische Interaktionen der Schülerinnen und Schüler insgesamt?
2. Lassen sich auf Schülerebene die im Kurs erzielten Lernzuwächse mit den klassifizierten Interaktionen erklären? Vermutet wird, dass die als systematisch klassifizierten Interaktionen bei der Diagnostik und der Weiterarbeit mit den Lernzuwächsen zusammenhängen.

4 Methode

4.1 Stichprobe

Am Projekt beteiligten sich im Durchführungszeitraum von Schuljahr 2013/2014 bis 2014/2015 insgesamt 21 Deutschlehrkräfte in neun Schulen (Realschulen und Mittelschulen, Klassenstufe 7/8). Die Rekrutierung der Lehrkräfte erfolgte über Fortbildungen und Rundschreiben, in denen die Projektteilnahme angeboten wurde. Ein großer Teil der Lehrkräfte kann als computeraffin bezeichnet werden. An einigen der Projektschulen standen gut ausgestattete PC-Räume zur Verfügung. Es gab jedoch auch Lehrkräfte ohne Erfahrung mit dem Einsatz digitaler Lernmaterialien im Unterricht oder ohne entsprechende Infrastruktur. In diesen Schulen wurde das Projekt mit einem mobilen Tablet-Klassensatz durchgeführt. Die hier vorgestellten und diskutierten Analysen beziehen sich auf fünf Deutschlehrkräfte sowie 129 Schülerinnen und Schüler, die über einen sehr langen Zeitraum (ein Schulhalbjahr bis 1,5 Schuljahre) mit dem Moodle-Kurs gearbeitet hatten. Damit wird sichergestellt, dass die Schülerinnen und Schüler die Möglichkeiten der formativen Lernverlaufsdiagnostik in allen Bereichen auch ausschöpfen konnten. Durch diese Reduzierung der Stichprobe wird die Übertragbarkeit der Befunde zwar eingeschränkt. Andererseits können dadurch die Zusammenhänge zwischen Nutzungsmustern und Lernerfolg genau beschrieben werden.

4.2 Entwicklung des Moodle-Kurses

Sprachbetrachtung spielt sowohl in der Primarstufe als auch in der Sekundarstufe I eine wichtige Rolle. Da in dieser Lerndomäne sowohl deklaratives als auch darauf aufbauendes prozedurales Wissen über mehrere Schuljahre hinweg erworben wird, bietet sich eine wiederholte Diagnose und Förderung grundlegender deklarativer Lerninhalte an. Die deutschdidaktische und sprachwissenschaftliche Literatur diskutiert unter den Stichworten Sprachbetrachtung, Sprachbewusstsein oder Sprachbewusstheit eine neue Akzentuierung des traditionellen Grammatik- und Rechtschreibunterrichts im Fach Deutsch (z. B. Ossner 2006; Bredel 2007). Im Vordergrund steht das Ziel, dass Schülerinnen und Schüler grammatikalische Phänomene und sprachliche Regelungen in Verwendungszusammenhängen entdecken und dieses Wissen für die eigene Sprachproduktion funktional nutzen können. Eichler und Nold (2007, S. 63) differenzieren dabei zwischen Sprachreflexion/Grammatik und sprachlichem Handeln als zwei miteinander verbundenen Teilbereichen der Sprachbewusstheit. Bei der Sprachbetrachtung steht deklaratives Wissen (Grammatikwissen) im Vordergrund. Beim sprachlichen Handeln wird explizites Wissen (Grammatikwissen) oder implizites Wissen (Sprachgefühl) angewendet, um eigene oder fremde sprachliche Produktionen zu korrigieren.

In einem Moodle-Kurs wurden fünf Test- und Übungsmodule zu den Themen Wortarten, Kommasetzung, Satzglieder, Zeitformen sowie Groß- und Kleinschreibung entwickelt. Diese Themen sind deklarative Facetten des Kompetenzbereichs Sprachbetrachtung. Es handelt sich um Inhalte, die im Grammatikunterricht der Sekundarstufe I eine zentrale Rolle spielen und in verschiedenen Jahrgangsstufen auf unterschiedlichem Niveau behandelt und geübt werden. Die Test- und Übungsmodule wurden von einer Deutschlehrkraft (Realschule) und einer Germanistin entwickelt und in Moodle umgesetzt. Die Test- und Übungsmodule wurden innerhalb des Projektteams auf Funktionalität und Verständlichkeit geprüft. Der an der Testentwicklung beteiligte Deutschlehrer erprobte alle Module in einer Pilotstudie in zwei Klassen.

Die Lernplattform Moodle wurde verwendet, weil sich die Funktionalität „Test“ gut für die flexible Programmierung einer formativen Diagnostik eignet. An der Pädagogischen Hochschule Schwäbisch Gmünd konnte der Moodle-Host für das Projekt genutzt werden. Die am Projekt beteiligten Schulklassen erhielten externe Accounts, die Lehrkräfte wurden als Tutoren angelegt und konnten die Ergebnisse ihrer Klasse über das Internet einsehen.

Die Aufgabenentwicklung erfolgte unter Rückgriff auf aktuelle Lehrwerke (vor allem *Klartext-Sprach-Lesebuch Deutsch*, Westermann Verlag). Insbesondere die Einteilung in drei Niveaustufen wurde dem curricularen Vorgehen in den Lehrwerken angeglichen. Hierfür wurde analysiert, in welchen Lehrwerken welche grammatikalischen Aspekte wann behandelt werden. Niveau 1 entspricht Grundschulwissen, das zu Beginn der Sekundarstufe I beherrscht werden sollte (z. B. die Wortarten Nomen, Verben, Adjektive und Artikel). Auf Niveaustufe 2 wird Wissen im Bereich Sprachbetrachtung geprüft, das im Bildungsplan für die Realschule in den Jahrgangsstufen 6 bis 8 verankert ist. Die Tests auf Niveaustufe 3 prüfen grammatikalisches Wissen, das in Schulbüchern eher selten angesprochen wird. Oder sie enthalten Aufgaben mit

komplexerem Wortmaterial und höheren Transferleistungen, die zum Teil auch als prozedurale Facetten des Kompetenzbereichs Sprachbetrachtung verstanden werden können (z. B. Anwendung von Zeitformen in verschiedenen Textsorten).

Bei der Testentwicklung wurden die spezifischen Testaufgabenformate in Moodle genutzt. Für die Bereiche Wortarten, Zeitformen (Niveau 1 und 2) und Satzglieder wurde das Aufgabenformat Single Choice gewählt. Die Aufgaben zur Kommasetzung wurden als Lückentexte angelegt, in denen die Kommata durch ein Drop-down-Menü vervollständigt werden müssen. Das Format Lückentext wurde auch bei den Zeitformen (Niveau 3) verwendet. Zur Prüfung von Wissen im Bereich der Groß- und Kleinschreibung wurden Audiodateien in die Testitems eingebunden. Die Schülerinnen und Schüler hörten ein Wort und mussten damit einen vorgegebenen Satz mit Lücke ergänzen.

Die Schülerinnen und Schüler konnten die einzelnen Lernbereiche frei wählen, wurden allerdings durch die Kurseinstellungen gezwungen, mit dem formativen Test auf Niveaustufe 1 zu beginnen. Bei Erreichen des Mindestwerts von 80 % richtig gelöster Aufgaben wurde der nächsthöhere Test automatisch freigeschaltet. Falls nicht, konnte ein Test frühestens nach 24 h wiederholt werden. Damit wurde verhindert, dass Lernende durch mehrfaches Ausprobieren den Test bestehen. Im Sinne einer formativen Diagnostik standen bei Nichtbestehen eines Tests spezifische Übungsaufgaben für die jeweilige Niveaustufe zur Verfügung. Die Übungsaufgaben und die Tests unterscheiden sich in der Art der Rückmeldung. Während bei den Tests die Rückmeldung erst nach Bearbeitung aller Aufgaben angezeigt wurde, erhielten die Lernenden unmittelbar nach jeder Übungsaufgabe eine Rückmeldung zur Korrektheit der Lösung. Danach konnte die Übungsaufgabe sofort noch einmal bearbeitet werden. Für alle Bereiche wurden zudem Glossare angelegt, die alphabetisch sortierte Einträge zu den sprachlichen Phänomenen der Bereiche auflisten. Die Einträge geben nach einer Beschreibung Beispiele für das Phänomen an und enthalten auch Merksätze zur Ableitung von Regeln.

Die Lehrkraft konnte mit der Teilnehmerübersicht das Gesamtergebnis der Klasse oder einzelner Schülerinnen und Schüler in den Testmodulen oder für den gesamten Kurs einsehen. Ebenso konnten klassenbezogene Lösungshäufigkeiten einzelner Testaufgaben angezeigt werden.

4.3 Durchführung des Moodle-Kurses und Nutzungsdauer

Die beteiligten Lehrkräfte wurden gebeten, den Moodle-Kurs als formative Lernverlaufsdiagnostik in ihren laufenden Deutschunterricht sinnvoll zu integrieren. Gespräche, Interviews und Beobachtungen in den Klassenzimmern ergaben, dass Grammatikthemen einerseits gesondert behandelt und geübt, andererseits aber auch im Zusammenhang mit Schreibaufgaben oder der Bearbeitung einer Lektüre aufgegriffen werden. Dies entspricht zumindest teilweise dem aktuellen fachdidaktischen Konzept eines integrativen Deutschunterrichts, wie er auch in den meisten Schulbüchern verankert ist. Aus diesem Grund wurde keine festgelegte Lernschrittfolge im Sinne des *mastery learning* vorgegeben. Bereits die große Spannweite der Niveaustufen (von Primarstufe bis Ende Sek 1) sprach gegen eine genaue Vorgabe instruktionaler Schrittfolgen für den Unterricht. Die Lehrkräfte nahmen den Moodle-Kurs entweder

zum Anlass, einen Themenbereich zu wiederholen oder sie setzten den Kurs dann ein, wenn ein bestimmter Aspekt der Sprachbetrachtung zur Behandlung im Unterricht anstand. Diese flexible Möglichkeit der Implementation war für eine hohe Akzeptanz unter den beteiligten Lehrkräften verantwortlich.

In den 21 beteiligten Klassen wurden im Schnitt 7.2 Tests durchgeführt. Die Anzahl der durchgeführten Tests variierte jedoch sehr stark zwischen den Klassen (SD = 5,4). Um tatsächlich auch längere Lernverläufe abbilden zu können, wurden für diese Studie die fünf Klassen mit den höchsten Beteiligungsquoten ausgewählt. Die Spannweite der gemittelten Testdurchführungen pro Schülerin bzw. Schüler liegt zwischen 10,6 und 14,6. Damit liegen in diesen Klassen ausreichend Datenpunkte vor, um eine differenzierte Analyse von Lernverlaufsmustern vornehmen zu können. Die projektbegleitenden Beobachtungen und Lehrerinterviews ergaben zudem, dass in zwei der untersuchten Klassen der Kurs sehr häufig als Hausaufgabe eingesetzt wurde. In den anderen Klassen dagegen eher in Selbstlernzeiten während des Unterrichtsvormittags. Die fünf ausgewählten Klassen repräsentieren damit eine gewisse Bandbreite verschiedener Implementationsmöglichkeiten.

4.4 Testschwierigkeit, interne Konsistenz und Validität der formativen Tests

Tab. 1 zeigt für jeden Lernbereich und jedes Niveau die Testschwierigkeiten für den jeweils ersten Testversuch. Der maximale Punktwert lag bei jedem Test bei 100. Die mittleren Testwerte für die Niveaustufen 1 und 2 liegen überwiegend zwischen 70 und 90 Punkten. Eine Ausnahme in dieser Hinsicht ist der deutlich schwierigere Kommasetzungstest auf Niveau 1. Die Gespräche mit den Lehrkräften lieferten allerdings eine plausible Erklärung hierfür. Der Bereich Kommasetzung (v. a. das optionale Komma) gilt als stark vernachlässigt in vielen Lehrwerken und spielte im Deutschunterricht der meisten Lehrkräfte bisher kaum eine Rolle. Die hohen Niveau 1-Testwerte in den anderen Bereichen sind ein Indiz dafür, dass die Primarstufenlerninhalte zu Beginn der Sekundarstufe I genügend oft wiederholt wurden. Hohe mittlere Testwerte beim Erstversuch auf Niveau 2 sind allerdings bereits ein Selektionseffekt, der auf Schülerinnen und Schüler mit erfolgreichem Testversuch auf Niveau 1 zurückgeht.

Die internen Konsistenzen der formativen Tests sind gut bis sehr gut und zwar auch bei Tests mit wenigen Aufgaben. Die hohe Varianz bei den Aufgabenzahlen ist durch die Testaufgabenformate bedingt. Die Zeitformen wie auch die Kommasetzung wurden beispielsweise durch Lückentextaufgaben geprüft. Somit enthält eine Aufgabe mehrere Teilaufgabenstellungen, prüft also mehrere Phänomene ab. Eine Schwachstelle von Moodle ist, dass diese Teilaufgaben nicht tabellarisch ausgegeben werden können, was im Grunde genommen eine Analyse auf unterster Itemebene für diesen Test verhindert.

Die einzelnen formativen Tests für die fünf Lernbereiche beanspruchen durch die streng an Lehrplänen und Schulbüchern orientierte Entwicklung eine hohe curriculare Validität. Diese wurde durch Rückmeldungen während des Projektes und durch die Abschlussbefragung der Lehrkräfte bestätigt. Das Item „Die Test-und Übungsaufgaben decken sich inhaltlich mit dem, was auch mir im Bereich Sprach-

Tab. 1 Mittlere Testwerte und interne Konsistenzen (Cronbach's Alpha) auf Basis des jeweils ersten Testversuchs pro Modul und Niveau

	Wortarten	Kommasetzung	Zeitformen	Satzglieder	Groß- und Kleinschreibung
Niveau 1	M = 89,0 SD = 11,1 α = 0,83 *n* = 423 40 Aufgaben	M = 58,7 SD = 19,3 α = 0,91 *n* = 384 10 Aufgaben	M = 84,3 SD = 17,4 α = 0,78 *n* = 265 16 Aufgaben	M = 78,1 SD = 26,0 α = 0,95 *n* = 205 27 Aufgaben	M = 82,4 SD = 16,0 α = 0,94 *n* = 238 42 Aufgaben
Niveau 2	M = 71,5 SD = 18,8 α = 0,92 *n* = 307 50 Aufgaben	M = 87,5 SD = 13,3 α = 0,81 *n* = 74 11 Aufgaben	M = 75,4 SD = 18,8 α = 0,80 *n* = 179 20 Aufgaben	M = 88,9 SD = 10,0 α = 0,72 *n* = 110 29 Aufgaben	M = 74,6 SD = 14,1 α = 0,90 *n* = 121 45 Aufgaben
Niveau 3	M = 76,6 SD = 16,9 α = 0,93 *n* = 115 60 Aufgaben	M = 90,4 SD = 7,1 α = 0,72 *n* = 42 10 Aufgaben	M = 66,0 SD = 21,4 α = 0,70 *n* = 79 8 Aufgaben	M = 55,0 SD = 17,7 α = 0,84 *n* = 59 29 Aufgaben	M = 76,8 SD = 10,7 α = 0,86 *n* = 41 42 Aufgaben

betrachtung wichtig ist" erreichte auf einer Skala von 1 (trifft nicht zu) bis 5 (trifft zu) einen Mittelwert von 4,2 (N = 20; SD = 0,54; Range von 3 bis 5).

4.5 Kategorisierung der Lernverlaufsmuster auf Individualebene

Die erste Zielsetzung des Beitrags ist die Entwicklung einer theoriebasierten Kategorisierung von Ereignissen innerhalb des Moodle-Kurses. Das Kategoriensystem wurde in verschiedenen Schritten entwickelt und an Teildatensätzen immer wieder erprobt. Leitend waren dabei folgende Fragen:

- Welche der durch das Lernmanagementsystem aufgezeichneten Interaktionen stehen für bestimmte Teilschritte formativer Diagnostik?
- Welche Interaktionen lassen sich vor dem Hintergrund der Prinzipien formativer Lernverlaufsdiagnostik als effektiv und welche als weniger effektiv bezeichnen?

Die erste Frage wurde wie folgt beantwortet: Jede Interaktion innerhalb des Moodle-Kurses kann zunächst einmal entweder als diagnostische Aktivität oder als Weiterarbeit im Lernbereich bezeichnet werden. Die diagnostischen Aktivitäten lassen sich wiederum in Vorwissensdiagnostik und Diagnostik von Lernfortschritten untergliedern. Die ersten Testversuche werden als Vorwissensdiagnostik für den Zeitraum im Deutschunterricht, in dem der Moodle-Kurs eingesetzt wurde, verstanden. Lernfortschritte liegen dann vor, wenn nach einem gescheiterten Testversuch die Wiederholung gelingt oder weitere Niveaustufen erfolgreich absolviert werden. Wenn eine Lehrkraft das Thema Satzglieder im Unterricht behandelt und dann erst in den Moodle-Kurs einsteigt, werden die Lernzuwächse vor Beginn der Arbeit mit dem Moodle-Kurs nicht abgebildet. Dies ist eine sehr strenge Definition, die allerdings dazu beiträgt, dass die Effekte der Übungen und Glossare im Kurs nicht überschätzt werden.

Tab. 2 Kategoriensystem für die Analyse der in Moodle gespeicherten Interaktionen und Ereignisse

Systematische Vorwissensdiagnostik		
A1	Erstversuch erfolgreich	Der erste Testversuch bei einer Niveaustufe ist erfolgreich (Testwert mind. 80 %)
A2	Erstversuch nicht erfolgreich	Der erste Testversuch bei einer Niveaustufe ist nicht erfolgreich (Testwert unter 80 %)
A3	Versuch nach Weiterarbeit nicht erfolgreich	Erfolgloser Testversuch nach passenden Übungen auf dem jeweiligen Niveau oder passenden Glossaren
Unsystematische Diagnostik		
B1	Weiterer Fehlversuch ohne Weiterarbeit	Erfolglose Testwiederholung ohne vorherige Übung oder Beschäftigung mit dem Glossar
B2	Unnötige Testwiederholung	Wiederholung eines bereits bestandenen Tests
Lernfortschritt		
C	Erfolgreiche Testversuche nach erstmaligem Scheitern	Nach einem nicht erfolgreichen Testversuch gelingt der Testversuch auf dieser Niveaustufe und/oder auf höheren Niveaustufen
Systematische Weiterarbeit		
D1	Passende Übung	Auf der Niveaustufe des gescheiterten Testversuchs oder auf der Niveaustufe des nachfolgenden Testversuchs (vorbereitende Übung) wird eine Übung absolviert. Der Abstand zwischen Test und Übung ist nicht größer als zwei Monate
D2	Passendes Glossar	Zum Bereich der gerade eben erfolgreich oder nicht erfolgreich durchgeführten oder der nachfolgenden Testversuche wird ein Glossar gelesen. Der Abstand zwischen Test und Glossar ist nicht größer als zwei Monate
Unsystematische Weiterarbeit		
E1	Nicht passende Übung	Es wird eine Übung zu einem bereits bestandenen Test gemacht bzw. auf die Übung folgt nicht der entsprechende Test
E2	Nicht passendes Glossar	Testversuche zu diesem Glossar werden nicht unternommen oder liegen über zwei Monate zurück

Um die zweite Frage beantworten zu können, wird zwischen systematischen und weniger systematischen Interaktionsfolgen unterschieden. Systematisch bedeutet, dass die Abfolge von diagnostischen Aktivitäten und Weiterarbeit inhaltlich aufeinander bezogen sind und in einer zeitlichen Nähe zueinander stehen. Um diese Prinzipien operationalisieren zu können, wurden zunächst die individuellen Lernverläufe auf Basis der Testergebnisse, der Datums- und Zeitangaben für die Tests und Übungen sowie der Glossar-Log-Daten rekonstruiert. Hierfür mussten verschiedene Moodle-Datenquellen gesondert aufbereitet und in einer Excel-Tabelle zusammengeführt werden. Auf einer Zeitleiste konnte dann für jede Schülerin und jeden Schüler dargestellt werden, wann welche Aktivitäten stattfanden. Unter Berücksichtigung theoretischer Merkmale formativer Diagnostik und der oben dargestellten Konstruktionsprinzipien für den Moodle-Kurs konnten einzelne Aktionen entweder als zielkonforme, systematische Nutzung oder als nicht zielkonforme, unsystematische Nutzung klassifiziert werden (Tab. 2). Bei der systematischen Nutzung konnte zudem zwischen Diagnostik und Weiterarbeit nach der Diagnostik unterschieden werden.

Das Kategoriensystem kann als niedrig-inferent bezeichnet werden, weil jedes Ereignis exakt zugeordnet werden konnte. Man hätte die Auswertung theoretisch auch programmieren können. Allerdings sind die Zuordnungsregeln nicht so einfach wie bei Ruipérez-Valiente et al. (2015). Auch aufgrund des explorativen Charakters der Studie wurde eine Kategorisierung durch *human raters* durchgeführt (analog zu You 2016). Bei der Entwicklung wurde das Kategoriensystem mehrfach geändert. Die endgültige Kodierung wurde von zwei Autoren dieses Artikels durchgeführt. Eine Klasse wurde gemeinsam kodiert, um das Kategoriensystem zu optimieren und zu erlernen. Anschließend wurden zwei weitere Klassen pro Kodierer kategorisiert und stichprobenartig gegenkontrolliert.

Insgesamt wurden in den fünf ausgewerteten Klassen 3966 Interaktionen in 569 Modulbearbeitungen klassifiziert. Die Modulbearbeitungen teilen sich wie folgt auf: 93 Groß- und Kleinschreibung, 127 Kommasetzung, 108 Satzglieder, 127 Wortarten, 114 Zeitformen. Die Module Wortarten und Kommasetzung sind in der Kursanordnung ganz oben und es wurde den Lehrkräften geraten, damit zu beginnen.

5 Ergebnisse

5.1 Nutzung des Moodle-Kurses und Kursdesign

In welchem Verhältnis stehen systematische und unsystematische Interaktionen (Forschungsfrage 1)? Tab. 3 zeigt die deskriptive Statistik für alle Interaktionskategorien in den 569 Modulbearbeitungen auf der Aggregationsebene einzelner Schülerinnen bzw. Schüler. Für jede teilnehmende Person wurden für alle bearbeiteten Module die Kategorienhäufigkeiten aufsummiert. Damit ergibt sich ein Bild der Arbeitsweise im Moodle-Kurs insgesamt. Hat eine Schülerin bzw. ein Schüler mehr Module bearbeitet, dann sind diese Werte auch insgesamt höher. Es werden somit Mittelwerte für die Gesamtaktivität der Lernenden angezeigt.

Die Werte schwanken zum Teil deutlich und nicht bei jeder Modulbearbeitung konnten alle Interaktionskategorien beobachtet werden. Die prozentualen Häufigkeiten des Wertes 0 für die Kategorien A1 und A2 liegen bei ca. 5 %. Nur sehr wenige Schülerinnen und Schüler haben tatsächlich gar keine Testversuche unternommen, sondern nur Übungen bearbeitet oder Glossare angeklickt. Die prozentualen Häufigkeiten des Wertes 0 liegen für die Kategorien E2 (nicht passendes Glossar) und B2 (Wiederholung eines bestandenen Tests) sehr hoch. Bei einem Drittel der Schülerinnen und Schüler konnte kein Lernfortschritt kategorisiert werden. Alle Verteilungen haben eine positive Schiefe (rechtsschief). Die Höhe der Schiefe in Kategorie B2 geht vor allem auf den Ausreißerwert zurück.

Abb. 1 zeigt die Kategorienmittelwerte nach Modulen. Die Kategorienmittelwerte der Interaktionen aus Tab. 3 werden in die Kategorienmittelwerte der einzelnen Module aufgeschlüsselt. Die Profillinien verdeutlichen einerseits den Zusammenhang innerhalb eines Moduls und andererseits die Differenzen zwischen den Modulen innerhalb jeder Kategorie. Man kann damit erkennen, wie sich die Arbeitsweisen zwischen den Modulen unterscheiden. Auf den ersten Blick sieht man nahezu parallel verlaufende Profillinien der Module Groß-/Kleinschreibung, Satzglieder, Wortarten

Tab. 3 Deskriptive Statistik für die Interaktionskategorien über alle fünf Module

	Min	Max	Mittelwert	Standard-abweichung	Rel. Häufigk. des Wertes 0	Schiefe
A1 Erstversuch erfolgreich	0	11	3,71	2,57	5,4	0,85
A2 Erstversuch nicht erfolgreich	0	8	3,20	1,72	3,8	0,43
A3 Erfolglose Testwiederholung nach Glossar oder Übungen	0	5	0,88	1,28	57,7	1,40
B1 Erfolglose Testwiederholung ohne Übungen oder Glossar	0	9	1,33	2,00	46,9	2,07
B2 Wiederholung eines bereits bestandenen Testversuchs	0	5/24[a]	0,56	2,23	74,6	9,38
C Lernfortschritt	0	8	1,91	2,01	33,1	1,07
D1 Passende Übung	0	29	3,98	4,50	18,5	2,44
D2 Passendes Glossar	0	16	2,81	3,39	33,1	1,67
E1 Nicht passende Übung	0	27	3,74	4,63	22,3	2,44
E2 Nicht passendes Glossar	0	7	0,44	1,00	70,8	3,41

[a] Bei B2 gibt es einen Ausreißerwert von 24

und Zeitformen bei den Diagnostikkategorien. Bei den Kategorien zur Weiterarbeit weichen die Profillinien dagegen deutlich voneinander ab. Die Profillinie des Kommasetzungsmoduls ist insgesamt atypisch.

Die meisten Testaktivitäten konnten den Kategorien A1 und A2 zugeordnet werden. Dies entspricht der vorgegebenen Vorgehensweise, ein Kursmodul zu wählen und zunächst eine möglichst hohe Niveaustufe zu erzielen. Die hohe Standardabweichung bei A1 war zu erwarten und verweist einmal auf die Heterogenität des Vorwissens der beteiligten Schülerinnen und Schüler, aber auch auf das vergleichsweise schwierige Kommasetzungsmodul. Der Mittelwert von B1 liegt höher als der Mittelwert von A3. Eine Testwiederholung ist eher dann erfolglos, wenn keine passenden Übungen oder Glossare bearbeitet wurden. Die Kategorie B2 hat einen kleinen Mittelwert. Über 90 % der Schülerinnen und Schüler haben hier den Wert 0 oder 1. Die hohe Standardabweichung geht u. a. auf den Ausreißerwert von 24 zurück. Bezogen auf die Forschungsfrage gibt es damit ein deutliches Übergewicht der als systematisch kategorisierten Testaktivitäten. Die modulbezogene Darstellung zeigt jedoch, dass das schwierige Kommasetzungsmodul von diesem Befund abweicht. Durch das geringe Vorwissensniveau sind erfolglose Testwiederholungen ohne vorhergehende Bearbeitung von Übungen oder Glossaren wesentlich höher. Eventuell waren viele Schülerinnen und Schüler frustriert, bereits bei Niveau 1 zu scheitern und setzten auf eine *trial and error*-Strategie.

Im Mittel gab es pro Schülerin bzw. pro Schüler zwei erfolgreiche Tests, die als Lernfortschritt gewertet werden konnten (Kategorie C). Die Verteilung ist allerdings rechtsschief. 32 % der Schülerinnen und Schüler erzielten keinen Lernfortschritt und

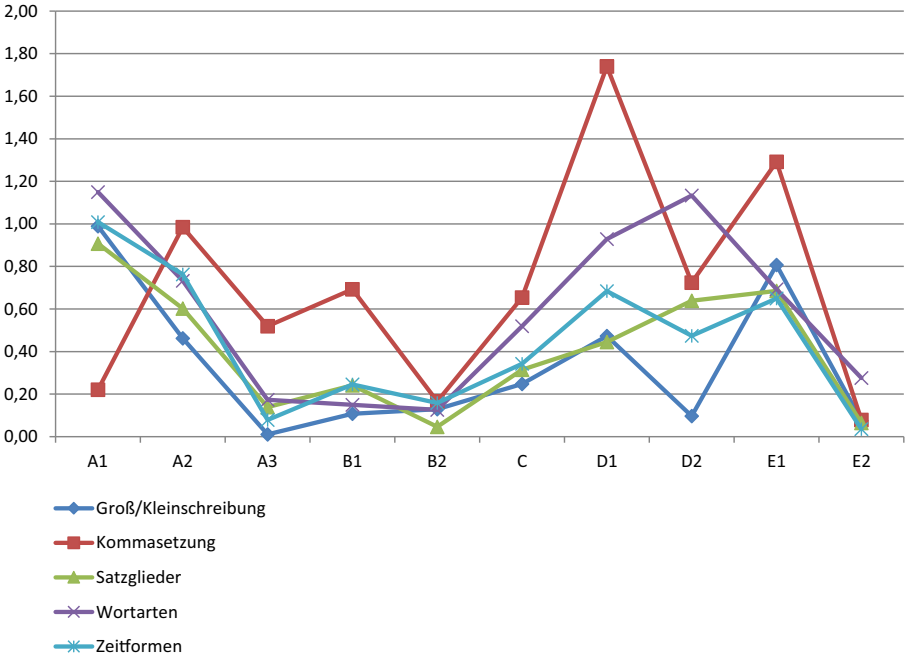

Abb. 1 Mittelwerte der Interaktionskategorien nach Modulen

ca. 50 % einen Lernfortschritt zwischen 1 und 4. Auch diese Verteilung der Lernfortschritte war zu erwarten und spiegelt die Leistungsheterogenität der Schülerinnen und Schüler wider. Die Verteilung der Lernfortschritte auf die Module ist unterschiedlich. Im Kommasetzungsmodul wurden die meisten Lernfortschritte erzielt, weil hier das Vorwissensniveau am geringsten war. Die zweithäufigsten Lernfortschritte gab es im Modul Wortarten, obwohl dort das Vorwissensniveau am höchsten war.

Übungen wurden insgesamt häufiger kategorisiert als Glossare. Wenn mit Glossaren gearbeitet wurde, dann in der Regel mit dem zum gerade bearbeiteten Modul passenden Glossar. Bei den Übungen halten sich die systematische und die unsystematische Nutzung die Waage. Der hohe Wert für nicht passende Übungen hängt unter anderem mit der strengen Kategorisierungsregel zusammen. Nicht nur das Modul, sondern auch die Niveaustufe musste passen. Bei der modulbezogenen Darstellung fallen sofort wieder die hohen Werte für passende und nicht passende Übungen bei der Kommasetzung auf. Die häufig nicht bestandenen Tests in diesem Modul haben eine hohe Übungsfrequenz induziert, allerdings nicht immer systematisch. Die Grafik zeigt auch, dass im Modul Groß- und Kleinschreibung die nicht passend gewählten Übungen überwiegen. Dies könnte folgenden Grund haben: Beim Modul Wortarten wird genau einer Niveaustufe eine Übung zugeordnet, bei den anderen Modulen sind es mehrere Übungen, die jeweils inhaltlich spezifiziert sind. Bei Groß- und Kleinschreibung wurden Übungen mehreren Niveaustufen zugeordnet, sodass eher eine passende Übung gewählt werden konnte.

Auch bei den Glossaren gibt es große modulspezifische Differenzen. In den Glossaren für die Module Kommasetzung, Satzglieder, Wortarten und Zeitformen konnten die Einträge sehr einfach einem Testinhalt zugeordnet werden. Bei den Groß- und Kleinschreibregeln waren die Einträge nach Wortarten sortiert. Diese Systematik war für die Schülerinnen und Schüler anscheinend zu schwierig bzw. nicht kompatibel mit den im Unterricht gelernten Regeln, sodass die Nutzung praktisch unterblieb. Eine zeitlich nicht passende Nutzung findet sich in geringfügigem Umfang lediglich beim Glossar für die Wortarten. Dieses Glossar wurde von den Schülerinnen und Schülern vermutlich als sehr klar strukturiert und hilfreich wahrgenommen und deshalb nach Beendigung ihrer Testaktivitäten in diesem Modul für die Bearbeitung weiterer Übungen herangezogen.

5.2 Interaktionen und Lernzuwächse

Lassen sich auf Schülerebene die im Kurs erzielten Lernzuwächse mit den klassifizierten Interaktionen erklären (Forschungsfrage 2)? Vermutet wird, dass die als systematisch klassifizierten Interaktionen bei der Diagnostik und der Weiterarbeit mit Lernzuwächsen zusammenhängen. Um diese Frage beantworten zu können, wurde für jedes Modul eine lineare multiple Regression mit der Kategorie C als abhängige Variable berechnet (Tab. 4). Für eine lineare Regression gelten bestimmte Voraussetzungen (Gollwitzer et al. 2015). Die Regressoren können bei einer linearen Regression metrisch oder dichotom skaliert sein. Die Verteilung innerhalb der Variablen spielt bei einer Regression keine Rolle, d. h., sie müssen nicht zwingend normalverteilt sein. Damit dürften die überwiegend rechtsschiefen Verteilungen der als Regressoren verwendeten Kategorien keine Beeinträchtigung der Ergebnisse darstellen. Eine weitere Voraussetzung für eine lineare Regression ist allerdings ein metrisches Skalenniveau der abhängigen Variablen. Ebenso sollte die abhängige Variable im Prinzip stetig sein. Die Kategorie C als abhängige Variable ist die Summe der gezählten Lernfortschritte innerhalb eines Moduls. Streng genommen müsste man deshalb eine Ordinalskalierung annehmen, weil nicht klar ist, ob die Unterschiede zwischen einzelnen Niveaustufen als äquidistant betrachtet werden können. Ebenso ist C als Zählvariable nicht stetig. Aus diesem Grund wurde in einem zweiten Schritt eine ordinale Regression gerechnet. Da die ordinale Regression nicht standardisierte Parameter schätzt, wurden signifikante Regressoren mit einem Kreuz in Tab. 4 markiert. Die ordinale Regression bestätigt die Mehrzahl der signifikanten Koeffizienten in der linearen Regression, gibt jedoch an vier Stellen Hinweise auf Unsicherheiten, die bei der Ergebnisinterpretation zu berücksichtigen sind.

Bei den klassifizierten Lernzuwächsen handelt es sich nicht um eine Leistung, die erst im Anschluss an die Arbeit mit dem Moodle-Kurs gemessen wurde, so wie es bisher in den meisten *learning analytics*-Studien üblich ist. Aus diesem Grund soll an dieser Stelle auch nicht von Prädiktoren gesprochen werden, vielmehr geht es um das Aufdecken von Zusammenhängen zwischen Lernfortschritten und weiteren Interaktionen innerhalb eines laufenden Kurses.

Bei den Modulen Satzglieder und Wortarten gibt es einen signifikanten bzw. tendenziell signifikanten negativen Zusammenhang mit dem Vorwissen in der linea-

Tab. 4 Zusammenhänge zwischen den klassifizierten Moodle-Interaktionen und dem Lernzuwachs in einzelnen Kursmodulen

	Groß/Klein-schreibung	Komma-setzung	Satzglieder	Wortarten	Zeitformen
A1 Erstvers. erfolg-reich	–	–	–0,22*	–0,17 (t)	–
A2 Erstvers. nicht erfolgreich	0,30 ** x	0,24* x	0,34*** x	x	0,44*** x
A3 Erfolgl. TW nach G/Ü	–	–	–	–	–
B1 Erfolgl. TW ohne G/Ü	–	–	–	–	–
B2 TW nach erfolgr. Erstv	–	0,25** x	–	–	–
D1 Passende Übung	0,19* x	–	–	0,27** x	–
D2 Passendes Glossar	x	–	0,45*** x	0,31** x	0,29** x
E1 Nicht passende Übung	–	–	–	–	–
E2 Nicht passendes Glossar	–	–	–	0,19* x	–
N	93	127	108	127	114
Varianzaufkl. lineare Regr	30,9 %	15,5 %	44,4 %	32,1 %	36,4 %
Varianzaufkl. ordinale Regr	50,4 %	27,9 %	43,9 %	44,0 %	53,6 %

Signifikanzniveaus: (t) ... $p < 0{,}10$; * ... $p < 0{,}05$; ** ... $p < 0{,}01$; *** ... $p < 0{,}001$. Standardisierte Koeffizienten, Signifikanzangaben sowie Varianzaufklärung beziehen sich auf eine lineare Regression mit Daten zu jedem Modul und der Kategorie C als abhängiger Variable. Mit x sind Regressoren markiert, die bei einer ordinalen Regression signifikant wurden ($p < 0{,}05$). Für die Varianzaufklärung der ordinalen Regression wurde das Pseudo-R-Quadrat nach Nagelkerke angegeben>

ren Regression, der durch die ordinale Regression allerdings nicht bestätigt werden konnte. Damit sind die Lernfortschritte eher unabhängig vom Vorwissensniveau.

In fast allen Modulen gibt es einen positiven Zusammenhang zwischen den gescheiterten Versuchen und dem Lernfortschritt. Dies war zu erwarten, weil erst nach einem gescheiterten Versuch Lernfortschritte möglich waren.

Erfolglose Testwiederholungen mit oder ohne passende Übungen und Glossare hängen nicht mit dem Lernfortschritt zusammen. D. h., auch die als A3 klassifizierten, systematischen erfolglosen Testwiederholungen scheinen eher wenig sinnvoll zu sein bzw. sind kein Signal für die weitere Übungsphase. Dafür gibt es zumindest einmal beim Modul Kommasetzung einen positiven Zusammenhang zwischen der ursprünglich als unsystematisch klassifizierten Wiederholung eines bereits erfolgreich absolvierten Testversuchs (B2) und den Lernzuwächsen.

Die positiven Regressionskoeffizienten bei den passenden Übungen und Glossaren wurden erwartet. Erwartungswidrig ist, dass sie nicht in allen Modulen auftreten bzw. auch sehr unterschiedlich in der Höhe ausfallen. Lediglich beim Modul Wortarten profitieren die Lernenden, wenn sie sowohl passende Übungen als auch passende Glossare bearbeiten. Beim Kommasetzungsmodul müssen Lernzuwächse mit Ursachen erklärt werden, die außerhalb des Kursdesigns liegen.

Bis auf den positiven Regressionskoeffizient für E2 im Modul Wortarten gibt es keine Zusammenhänge zwischen den als unsystematisch klassifizierten Übungen bzw. Glossaren und den Lernzuwächsen. Dies entspricht zumindest teilweise den Annahmen. Es gibt allerdings auch keine Hinweise auf negative Auswirkungen einer zeitlich nicht passenden Nutzung der Übungen und Glossare. Der positive Zusammenhang zwischen E2 und Wortartenlernzuwächsen liegt vermutlich, wie bereits im vorangehenden Abschnitt angedeutet, an der insgesamt umfangreichen Nutzung des Wortartenglossars.

6 Diskussion

Wie lassen sich die Befunde vor dem Hintergrund bisheriger *learning analytics*-Studien einordnen? Die Studie verdeutlicht, dass man aufgrund von theoretischen Überlegungen ein Kategoriensystem für die Analyse von Lernplattformereignissen entwickeln kann und die Befunde zur Modifikation des Kursdesigns beitragen können. Damit wird auf den Vorwurf reagiert, *learning analytics*-Studien seien weitgehend atheoretisch und induktiv (z. B. Clow 2013). Die Befunde geben analog zur Studie von You (2016) Einblicke in die Lernstrategien der Schülerinnen und Schüler. Dies gelingt vor allem durch eine Kategorisierung von Ereignissen, bei der die zeitliche Reihenfolge mit berücksichtigt wird. Die Kategorien könnten noch weiter verfeinert und differenziert werden. Vorbild hierfür wäre die Studie von Rupierez-Valente et al. (2015), in der Übungsaktivitäten im Hinblick auf die Nutzung von Hinweisen und die zeitliche Sequenzierung der Übungen und Hinweise (als Indikator für die Reflexionsfähigkeit) kategorisiert werden. Im vorliegenden Moodle-Kurs könnten beispielsweise die Wiederholungsraten einzelner Übungsaufgaben erfasst werden.

Ein weiteres Ziel von *learning analytics*-Studien ist die Generierung von Feedback in der Kursmitte. Dabei werden Prädiktoren gesucht, die bereits zu Beginn des Online-Kurses Hinweise auf die Lerneffektivität geben (Macfadyen und Dawson 2010; Tempelaar et al. 2015). Auch dieses Ziel müssten wir für unseren Moodle-Kurs weiter verfolgen. Ein Ansatzpunkt wäre die Darstellung der Kategorien A und B für Lehrkräfte nach drei bis vier Unterrichtsstunden mit dem Kurs. Die Lehrkräfte könnten sehen, welche Lernenden bereits auf Anhieb ein hohes Niveau erreichen und welche Lernenden ohne Übungs- oder Glossaraktivitäten erfolglos einen Test wiederholen.

Problematisch bei den Befunden zu Forschungsfrage 2 ist, dass viel Varianz unaufgeklärt bleibt. In den fünf untersuchten Klassen war der Moodle-Kurs lediglich ein Zusatzangebot für den Deutschunterricht. Diese Situation entspricht den Befunden von Agudo-Peregrina et al. (2014). Auch in dieser Studie konnten signifikante Prädiktoren nur für den Online-Kurs gefunden werden, für den alle Lernaktivitäten registriert werden. In Blended-Learning-Kursen wurde dagegen nur ein Teil der Lernaktivitäten registriert, was zu einer geringeren Varianzaufklärung führte. Für unsere Studie liegen zwar Beobachtungs- und Interviewdaten vor, die Hinweise auf den Unterricht der Lehrkräfte geben, allerdings sind diese nicht ausreichend für eine Quantifizierung von Lernaktivitäten auf Individualebene.

Eine weitere Limitation der Studie ist die Reduktion der Stichprobe auf fünf Klassen in einer Schulart. Damit haben die Befunde einen eher explorativen Charakter. In einem weiteren Schritt wird deshalb ein überarbeitetes Kategoriensystem auf alle 21 bisher teilnehmenden Klassen und alle neu dazugekommenen Klassen angewendet. Dies wird zu weiteren Analysemöglichkeiten, z. B. einer Berücksichtigung der Implementationsbedingungen (Hausaufgabe vs. Klassenunterricht), führen.

Die Befunde geben dennoch eine Reihe konkreter Hinweise für die Optimierung des Moodle-Kurses. Die Testaktivitäten können überwiegend als systematisch beschrieben werden. Der Moodle-Kurs wird von vielen Schülerinnen und Schülern in den fünf untersuchten Klassen als Vorwissens- und Lernfortschrittsdiagnostik genutzt. Erfolglose Testwiederholungen nach passenden Übungen und Glossaren wurden als systematische Diagnostik klassifiziert, hängen aber nicht mit Lernfortschritten zusammen. Nach einem zweiten erfolglosen Testversuch sollte die Lehrkraft ein Signal erhalten bzw. müssten die Lernenden durch Moodle darauf hingewiesen werden, ihre bisherige Arbeitsstrategie zu überdenken.

Die zunächst als unsystematisch klassifizierten Wiederholungen bereits erfolgreich bestandener Tests können durchaus lernförderlich sein. Eine mögliche Erklärung wäre, dass sich diese Schülerinnen und Schüler noch einmal vergewissern, welche Inhalte sie können und damit an Selbstsicherheit gewinnen, bevor sie einen Testversuch auf einer höheren Niveaustufe in Angriff nehmen. Die Testwiederholungen ohne Übung sind noch sehr hoch. Eine Reduktion wäre beispielsweise möglich, indem ein Test nur wiederholt werden darf, wenn die entsprechenden Übungen absolviert wurden. Ebenso müssten die Lehrkräfte bei der Einführung verstärkt auf diese Testwiederholungen achten.

Die Glossare in den Modulen Groß-/Kleinschreibung und Kommasetzung sollten strukturell den anderen Glossaren angepasst werden, um eine praktikable Nutzung zu ermöglichen. Um die Bedeutung der Übungen für den Lernzuwachs zu steigern, ist vermutlich eine bessere Verzahnung von Testbearbeitung und Übungen notwendig. Im Kategoriensystem wurde die Passung einer Übung mit zwei Monaten sehr großzügig bewertet. Hier ist bei der Einführung des Moodle-Kurses in den Klassen verstärkt darauf zu achten, dass Übungen zeitnah durchgeführt werden. Diese Vermutung müsste in weiteren Analyseschritten durch eine zeitlich feinere Kategorisierung geprüft werden.

Abschließend muss bemerkt werden, dass mit der hier vorliegenden Studie zwar ein Theoriekonzept für die Entwicklung eines Moodle-Kurses und die Entwicklung eines Klassifikationssystems für die Moodle-Ereignisdaten leitend war. Die Verknüpfungen zwischen Moodle-Kurs und Deutschunterricht vor, während und nach dem Einsatz des Moodle-Kurses bleibt in dieser Studie allerdings ein blinder Fleck. Uns liegen über begleitende Interviews Informationen zum Deutschunterricht vor, allerdings können diese nicht systematisch für die Analyse der Ereignisdaten genutzt werden. Aus diesem Grund mussten wir uns auch für eine enggeführte und strenge Operationalisierung von „Lernfortschritt“ entscheiden. Ein Lernfortschritt liegt nur dann vor, wenn eine Schülerin oder ein Schüler einen gescheiterten Testversuch in dem Bereich hatte. Wurden ein, zwei oder gar drei Niveaustufen auf Anhieb erfolgreich absolviert, wurde dieses Verhalten nicht als Lernfortschritt, sondern als Vorwissen registriert. Damit werden Lernverläufe außerhalb des Moodle-

Kurses (z. B. im Unterricht vor Beginn des Kurses bzw. parallel dazu) nicht abgebildet. In Folgestudien könnte man diese Problematik umgehen, indem man auf eine zeitlich festgelegte Vorwissensdiagnostik (z. B. zu Beginn des Schuljahres) zurückgreift und dann die Operationalisierung von Lernfortschritt weiter fassen kann, um Effekte des Unterrichts (nicht nur der Moodle-Interaktionen) auf die Lernverläufe prüfen zu können.

Literatur

Agudo-Peregrina, A. F., Iglesias-Pradas, S., Conde-González, M. A., & Hernández-García, A. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior*, *31*, 542–550.

Bloom, B. S. (1974). An introduction to mastery learning theory. In J. H. Block (Hrsg.), *Schools, society and mastery learning*. New York: Holt, Rinehart & Winston.

Bredel, U. (2007). *Sprachbetrachtung und Grammatikunterricht*. Paderborn: Schöningh UTB.

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, *18*(6), 683–695.

Eichler, W., & Nold, G. (2007). Sprachbewusstheit. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen - Konzepte und Messung. DESI-Studie* (S. 63–82). Weinheim: Beltz.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, *33*(2), 188–192.

Gollwitzer, M., Eid, M., & Schmitt, M. (2015). *Statistik und Forschungsmethoden* (4. Aufl.). Weinheim: Beltz.

Guskey, T. R. (2015). Mastery learning. In J. D. Wright (Hrsg.), *International encyclopedia of the social & behavioral sciences* (2. Aufl., S. 52–759). Oxford: Elsevier.

Iglesias-Pradas, S., Ruiz-de-Azcarate, C., & Agudo-Peregrina, A. F. (2015). Assessing the suitability of student interactions from Moodle data logs as predictors of cross-curricular competencies. *Computers in Human Behavior*, *47*, 81–89.

Kulik, C. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: a meta-analysis. *Review of Educational Research*, *60*(2), 265–299.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588–599.

Maier, U. (2014). Formative Leistungsdiagnostik in der Sekundarstufe – Grundlegende Fragen, domänenspezifische Verfahren und empirische Befunde. In M. Hasselhorn, W. Schneider, & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* Test und Trends, (Bd. 12, S. 19–40). Göttingen: Hogrefe.

Nunnery, J. A., Ross, S. M., & McDonald, A. (2006). A randomized experimental evaluation of the impact of accelerated reader/reading renaissance implementation on reading achievement in grades 3 to 6. *Journal of Education for Students Placed At Risk*, *11*(1), 1–18.

Ossner, J. (2006). *Sprachdidaktik Deutsch*. Paderborn: Schöningh UTB.

Ruipérez-Valiente, J. A., Munoz-Merino, P. J., Leony, D., & Delgado Kloos, C. (2015). ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform. *Computers in Human Behavior*, *47*, 139–148.

Souvignier, E., Förster, N., & Salaschek, M. (2014). quop: ein Ansatz internet-basierter Lernverlaufsdiagnostik und Testkonzepte für Mathematik und Lesen. In M. Hasselhorn, W. Schneider, & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* Test und Trends, (Bd. 12, S. 239–256). Göttingen: Hogrefe.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: review of research. *Psychology in the Schools*, *42*(8), 795–819.

Strathmann, A. M., & Klauer, K. J. (2010). Lernverlaufsdiagnostik: Ein Ansatz zur längerfristigen Lernfortschrittsmessung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *42*, 111–122.

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, *47*, 157–167.

Topping, K. J., & Fisher, A. M. (2003). Computerised formative assessment of reading comprehension: field trials in the UK. *Journal of Research in Reading*, *26*(3), 267–279.

Walter, J. (2011). Die Entwicklung eines auch computerbasiert einsetzbaren Instruments zur formativen Messung der Lesekompetenz. *Heilpädagogische Forschung*, *2011*(3), 106–126.

Walter, J. (2013). *VSL: Verlaufsdiagnostik sinnerfassenden Lesens*. Göttingen: Hogrefe.
Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior*, *47*, 168–181.
Yeh, S. S. (2010). The cost effectiveness of 22 approaches for raising student achievement. *Journal of Education Finance*, *36*(1), 38–75.
You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *Internet and Higher Education*, *29*, 23–30.
Zimmerman, B. J., & Dibenedetto, M. K. (2008). Mastery learning and assessment: implications for students and teachers in an era of high-stakes testing. *Psychology in the Schools*, *45*(3), 206–216.