Modellierung der Effekte lernzielorientierter Tests nach dem Prinzip des *mastery measurement* innerhalb einer digitalen, formativen Leistungsmessung

Uwe Maier

Online publiziert: 13. August 2020
© Der/die Autor(en) 2020

Zusammenfassung *Mastery measurements* sind kurze, kriterienorientierte, formative Tests zur Messung von Schülerleistungen in hierarchisch strukturierten Lern-domänen. Obwohl viele digitale Lernmaterialien dieses Testprinzip nutzen, sind Zusammenhänge zwischen *mastery measurments* und weiterführenden Lernprozessen bisher kaum Gegenstand empirischer Forschung. In dieser Studie werden deshalb die Effekte lernzielorientierter Tests innerhalb eines Moodle-Kurses zu Rechtschreibung und Grammatik untersucht. Als Datengrundlage dienen die im Moodle-Kurs anfallenden Bewertungen und Log-Daten von 407 Schülerinnen und Schülern aus 18 Realschul- und Gymnasialklassen (Jahrgangsstufe 6–9). Die Effekte des *mastery measurement* auf Feedbacknutzung, Lernaktivitäten im Kurs und Lernfortschritt wurden mit einem Bayes'schen Netz auf Ebene einzelner Test-Folgetest-Sequenzen modelliert. Es zeigte sich, dass eine Feedbacknutzung zusammen mit dem Nachlesen von Regeln und gezielten Übungen die Wahrscheinlichkeit eines Lernfortschritts innerhalb einer Test-Folgetest-Sequenz deutlich erhöht.

Schlüsselwörter Formative Leistungsmessung · Lernzielorientierte Tests · Feedback · Digitales Lernen · Lernmanagementsystem · Bayes'sches Netz

Zusatzmaterial online Zusätzliche Informationen sind in der Online-Version dieses Artikels (<https://doi.org/10.1007/s11618-020-00958-6>) enthalten.

Prof. Dr. U. Maier (✉)
Pädagogische Hochschule Schwäbisch Gmünd, Oberbettringerstraße 200, 73525 Schwäbisch Gmünd, Deutschland
E-Mail: Uwe.Maier@ph-gmuend.de

Modeling the effects of mastery measurements in a digital formative assessment system with a Bayesian network

Abstract Mastery measurements are short, criterion-referenced formative tests for assessing student performance in hierarchically-structured learning domains. Although many digital learning materials have built-in mastery measurements, few systems have been subject to empirical research. This study investigated the effects of mastery measurements on feedback usage, learning activities and learning progression within a Moodle course on grammar and spelling. Moodle log-files and scores from 407 students in 18 classrooms (Grades 6–9) were analyzed on the level of test-retest-sequences. A Bayesian network approach was used to model dependencies within the sequences and characteristics of the students and context. Data analyses revealed a pattern of feedback and learning activities that increased the probability of learning progress within a test-retest-sequence.

Keywords Bayesian network · Digital Learning · Feedback · Formative assessment · Learning management system · Mastery measurement

1 Einleitung

Unter formativer Leistungsmessung versteht man Verfahren oder Programme, um Leistungsmessungen im Unterricht so einzusetzen, dass die Schülerinnen und Schüler von den Rückmeldungen profitieren und die nachfolgenden Lehr- und Lernprozesse optimiert werden können (Black und Wiliam 2009; Maier 2010; Rakoczy 2012; Souvignier et al. 2018). Die Forschungsbefunde sind einerseits ermutigend, weil immer wieder gezeigt werden konnte, dass formative Leistungsrückmeldungen zu Lernzuwächsen führen (z.B. Black und Wiliam 1998; Kingston und Nash 2011). Andererseits gibt es auch viele Studien mit nur geringen Effektstärken sowie Studien, die aufzeigen, dass nicht jede Art von Rückmeldung hilfreich ist (Bennett 2011; McMillan et al. 2013; Maier et al. 2016).

Auch die digitalen Versionen der formativen Leistungsmessung sind vielversprechend (vgl. Russell 2010; Gikandi et al. 2011; McLaughlin und Yan 2017). Die Befundlage ist allerdings aufgrund der vielen Variationsmöglichkeiten bei der informationstechnischen Umsetzung, der Anwendungsgebiete sowie der Einsatzszenarien sehr umfangreich. Diese Studie beschäftigt sich mit einer digitalen, formativen Leistungsmessung im Schulbereich. Die grundlegende Idee ist, eine einfach in den Unterricht zu integrierende Diagnostik und Förderung von Grundlagenwissen zu entwickeln und zu erproben. Hierfür wurden für inhaltlich gut abgrenzbare Kurse im Bereich Sprachbetrachtung (Grammatik, Rechtschreibung) in der Sekundarstufe 1 (z.B. Kommasetzung, Wortarten) lernzielorientierte Tests nach dem Prinzip des *mastery measurement*, Trainings und Regelglossare auf unterschiedlichen Niveaustufen entwickelt und mithilfe des Lernmanagementsystems Moodle den teilnehmenden Schulen zur Verfügung gestellt. In bisherigen Studien wurden Testgüte sowie Effekte der Trainings und Regelglossare auf den Lernzuwachs innerhalb der Kurse untersucht (Maier 2017; Maier et al. 2017). Das Ziel dieser Studie ist ein genauer

Blick auf die Effekte der nach dem Prinzip des *mastery measurement* konzipierten Tests innerhalb einer weiterentwickelten Variante der Kurse.

2 Theorie und Forschungsstand

2.1 Lernfortschrittsdiagnostik und *mastery measurement*

Bereits seit Jahrzehnten werden *curriculum-based measurements* in den USA vor allem im sonderpädagogischen Förderbereich genutzt. Kurze und häufig wiederholbare Paralleltests für Grundfertigkeiten im Primarbereich werden in regelmäßigen Abständen eingesetzt, um den Lernfortschritt von Schülerinnen und Schülern in Förderprogrammen zu überwachen. Dabei wird zwischen zwei grundlegend unterschiedlichen Prinzipien des *curriculum-based measurement* unterschieden (Fuchs und Deno 1991): Beim sog. *general outcome measurement* wird ein kurzer, formativer Test mit vielen Paralleltestversionen über einen längeren Zeitraum als Indikator eingesetzt, um die Entwicklung einer komplexeren Fertigkeit (z. B. Lesen) abzuschätzen. Beim sog. *specific subskill mastery measurement* (hier kurz: *mastery measurement*) werden die Lerninhalte in Teilfertigkeiten gegliedert und als einzelne Lernziele geordnet (Fuchs und Deno 1991). Jeder Teilfertigkeit wird ein Set an Testitems zugeordnet. Ebenso wird bei jeder Teilfertigkeit ein *mastery criterion* für die Zielerreichung festgelegt. Das *mastery measurement* ermöglicht damit eine Lernverlaufsdiagnostik für deklarativ und hierarchisch organisierte Lerndomänen (Regeln, Begriffe, Konzepte) wie man sie vor allem in der Sekundarstufe findet.

Das Testprinzip des *general outcome measurement* dominiert die US-amerikanische Forschungstradition zum *curriculum-based measurement* und wird seit einigen Jahren verstärkt in Deutschland rezipiert und weitergeführt (z. B. Walter 2010; Hasselhorn et al. 2014; Sikora und Voß 2017). Neben der Testentwicklung stehen verstärkt Fragen der Implementation in den Unterricht und Effekte des Feedbacks bzw. von Fördermaterialien, die sich auf das Feedback beziehen, im Fokus (z. B. Hebbecker und Souvignier 2018). Formative Assessment-Programme, die im Sinne von Fuchs und Deno (1991) als *mastery measurement* bezeichnet werden können, gehen letztendlich auf das Konzept *mastery learning* zurück (Bloom 1974; Zimmermann und DiBenedetto 2008). Obwohl die Begriffe *mastery learning* und *mastery assessment* in der heutigen Forschung kaum noch verwendet werden, findet sich das grundlegende Prinzip in vielen praktischen Anwendungen (z. B. online-Tests zu Schulbüchern). Im Vergleich zu dieser unterrichtspraktischen Bedeutsamkeit gibt es nur sehr wenige Studien, die sich explizit mit den Effekten von *mastery measurements* im Kontext formativer Leistungsmessung beschäftigen.

2.2 Die Rolle des Feedbacks beim *mastery measurement*

Ein Vorteil formativer Leistungsmessung mit *mastery assessments* ist die Möglichkeit, inhaltlich elaborierte Rückmeldeformate zu nutzen. Die Feedbackforschung zeigt generell, dass zeitnahe und an Lernzielen orientierte Rückmeldungen effektiver sind als Rückmeldungen, die wertend sind, sich auf die Person beziehen oder

einer sozialen Bezugsnorm folgen (Bangert-Drowns et al. 1991; Kluger und DeNisi 1996). Zudem gibt es zahlreiche Hinweise, dass die Effektivität des Feedbacks mit dem Elaborationsgrad steigt (Van der Kleij et al. 2015).

Formative Leistungsmessung mit *mastery measurements* kann sich diesen Umstand zunutze machen. Wenn sich die formativen Tests auf sehr spezifische Lernzielhierarchien beziehen, dann ist bereits die Rückmeldung, ob ein Test erfolgreich absolviert wurde oder nicht, eine auf das konkrete Lernziel bezogene Rückmeldung. Orientiert man sich an der Klassifikation von elaboriertem Feedback nach Shute (2008), sind weitere Möglichkeiten der Elaboration einer Rückmeldung im Zuge der formativen Leistungsmessung denkbar (Wiederholung der Instruktion, Verweis auf Regeln usw.). Allerdings zeigen Studien auch, dass nicht jede Form der Elaboration gleich effektiv ist (Kluger und DeNisi 1996; Hattie und Timperley 2007; Van der Kleij et al. 2012). Ein wichtiges Bündel an Mediatoren sind motivationale Aspekte. Timmers et al. (2013) konnten beispielsweise zeigen, dass das Einsehen der elaborierten Rückmeldung mit einer hohen Erfolgserwartung und einer hohen Bewertung der Aufgabennützlichkeit korreliert.

Die Nutzung von elaboriertem Feedback ist zudem eine Funktion der domänen-spezifischen Schülerleistungsfähigkeit sowie der Korrektheit von Tests oder Aufgaben, auf die sich das Feedback bezieht. Hancock et al. (1995) zeigten, dass leistungsstarke Schülerinnen und Schüler das Feedback dann in Anspruch nehmen, wenn die Frage nicht korrekt war oder sie sich auch bei einer korrekt gelösten Frage unsicher fühlen. Die leistungsschwächsten Schülerinnen und Schüler nutzen das Feedback unabhängig von der Korrektheit einer Antwort am wenigsten, weil sie das Lernmaterial möglichst schnell bearbeiten wollten. Dieser Befund konnte von Timmers und Veldkamp (2011) bestätigt werden. Das elaborierte Feedback wurde vor allem nach falsch gelösten Aufgaben genutzt (*corrective feedback*), weniger nach richtig gelösten Aufgaben (*confirmatory feedback*). Bei Aufgaben mit höherem Schwierigkeitsgrad nahm die Nutzung des elaborierten Feedbacks allerdings ab. Gleiches galt für Studierende mit geringeren Gesamtpunkten beim formativen Assessment.

Auch situative Aspekte der Einbindung des Feedbacks in die Lernumgebung bzw. das Lernmaterial wurden untersucht. In der Studie von Timmers und Veldkamp (2011) wurde die Nutzung des Feedbacks nicht von der Supervision der Lernaktivitäten beeinflusst. Es spielte keine Rolle, ob das Lernmaterial zuhause in Alleinarbeit oder unter Aufsicht einer Lehrperson bearbeitet wurde. Allerdings gab es einen Effekt der Testlänge. Bei kürzeren formativen Tests stieg der Prozentsatz der Studierenden, die das elaborierte Feedback einsahen. Zudem gibt es einen positiven Zusammenhang zwischen Nutzungsintensität des formativen Assessments und Häufigkeit der Feedbacknutzung (Timmers et al. 2013).

2.3 Studien zu digitaler, formativer Leistungsmessung mit *mastery measurements*

Das Prinzip des *mastery measurement* wird in sehr vielen digitalen Lernmaterialien angewendet, ohne dass es als solches bezeichnet wird. Nahezu jedes Schulbuch hat mittlerweile digitale Zusatzmaterialien, die unter anderem Kontrollen und kurze Wiederholungstests, die sich auf Lehrbuchkapitel und damit spezifische Lernziele

beziehen. Auch kommerzielle Lernplattformen bis hin zu den großen, internationalen online-Kursanbietern arbeiten mit dem Prinzip des *mastery measurement*. Nach kleinen, überschaubaren Lerneinheiten (z. B. Lernvideos) sind Minitest, mit denen das Verständnis geprüft wird, zu absolvieren. Erst danach kann man im Kurs zur nächsten Lerneinheit wechseln. Trotz dieser Fülle an schulpraktischen Umsetzungen ist die Anzahl der empirischen Studien, die sich mit den Effekten von *mastery measurement* im Rahmen digitaler, formativer Leistungsmessung beschäftigen, relativ begrenzt.

Jia et al. (2012) konnten in einer quasi-experimentellen Studie zeigen, dass Schülerinnen und Schüler von einem Moodle-Kurs zur Übung und Wiederholung von Englischvokabeln profitieren. Die Studie liefert allerdings keine Einblicke in die Wirkmechanismen der Tests, des Feedback und der Lernaktivitäten im Rahmen des Moodle-Kurses. Bokhove und Drijvers (2012) evaluierten ein Lernsystem für Algebra in der Sekundarstufe II und beschreiben u. a. die Nutzungsdauer der formativen und summativen Tests innerhalb des Systems als wichtige Prädiktoren für den Lernerfolg. Weitere Aspekte der Nutzung der formativen Leistungsmessungen (z. B. Arbeit in den Trainingsmodulen, Fehler, Wiederholungen) wurden nicht ausgewertet bzw. in die Analysen einbezogen. Ein weiteres Beispiel für eine formative Leistungsmessung nach dem Prinzip des *mastery measurement* ist Snappet, eine für holländische Grundschulen entwickelte Lern-App für Grundfertigkeiten in Mathematik (Faber et al. 2017) und Rechtschreibung (Faber und Visscher 2018). Auch hier zeigte sich, dass der Lernzuwachs vor allem von der Nutzungsintensität (Testwiederholungen, adaptive Trainings) abhängt. Zudem profitierten eher die leistungstärkeren Schülerinnen und Schüler von der Lern-App.

In Anlehnung an das Konzept des *mastery measurement* entwickelten Maier et al. (2017) eine Moodle-Lernumgebung für Rechtschreibung und Grammatik in der Sekundarstufe I. Innerhalb einzelner Kurse (z. B. Wortarten, Kommasetzung) waren curricular gestufte, formative Tests zu absolvieren, um die nächsthöhere Niveaustufe im Kurs zu erreichen. In Abhängigkeit der Testrückmeldung konnten die Schülerinnen und Schüler wichtige Regeln in Glossaren nachschlagen oder Trainingsaufgaben machen. In den Analysen zeigte sich, dass die Nutzung der Trainings und Regelglossare in fast allen Kursen positiv mit dem Lernzuwachs zusammenhing. Analog zu den Befunden von Bokhove und Drijvers (2012) oder Faber und Visscher (2018) gab es einen deutlichen Zusammenhang zwischen intensiver Kursnutzung sowie häufiger Testwiederholung und Lernzuwachsen.

3 Untersuchungskontext und Forschungsfragen

Kontext der Studie ist eine in Moodle realisierte, formative Leistungsmessung zu grundlegenden Themenfeldern der deutschen Grammatik und Rechtschreibung für die Sekundarstufe 1 (Tab. 1). Die einzelnen Kurse orientieren sich an inhaltlichen Facetten (Regeln, Teilbegriffe) und sind jeweils in drei bis vier Niveaustufen gegliedert, die eine curriculare Reihung der Lerninhalte abbilden. Niveau 1 sind die in der Grundschule bereits behandelten Themengebiete. Alle Kurse sind strukturgleich aufgebaut. Es gibt pro Kurs ein Buch (Moodle-Element) mit den Regeln, z. B.: „Was

Tab. 1 Curriculare Struktur der Kurse

Kurs	Niveau 1	Niveau 2	Niveau 3	Niveau 4
Wortarten	Adjektiv, Artikel, Verb, Nomen	Adverb, Konjunktion, Präposition, Pronomen	Interjektion, Eigennamen, Numerale	–
Kommasetzung	Aufzählung, einfache Satzreihen	Satzreihe, wörtliche Rede	Konjunktionalsätze, Relativsätze	Apposition, Infinitivkonstruktion, schwere Konjunktionalsätze
Zeitformen	Gegenwart (Präsens), einfache Vergangenheit (Präteritum), zusammengesetzte Vergangenheit (Perfekt), Zukunft (Futur I)	Plusquamperfekt, Futur II	Anwendung der Zeiten in verschiedenen Textsorten	–
Satzglieder	Subjekt, Prädikat, Akkusativobjekt, Dativobjekt	Adverbiale Bestimmung der Zeit, des Ortes, des Grundes, der Art und Weise	Präpositionales Objekt, Genitivobjekt	–
Groß- und Kleinschreibung	Einfache und zusammengesetzte Nomen, Eigennamen, Satzanfänge	Abstrakte Nomen und Fremdwörter, Eigennamen, Nominalisierungen, Anredepronomen, Zeitangaben	Nominalisierungen nach Pronomen, Mengenangaben, Artikel oder Präposition, Anredepronomen, Nomen in Verb. mit „sein“ und „haben“	–
Dehnung und Schärfung	Dehnung bei einfachen Wörtern mit gedehnt gesprochenem Vokal: sieht, geht, Moor, froh	Schärfung bei kurz gesprochenen Vokalen: nass, Butter, backen	Wörter mit Dehnung und Schärfung, Ausnahmen: Augenlid, Kamel, s-ss-ß (Fass, Fuß, Fluss)	Längere Wörter mit mehr als einer Lupeinstelle
Getrennt- und Zusammenschreibung	Nomenverbindungen: Nomen + Nomen, Nomen + Adjektiv, Nomen + Verb mit Begleiter, Nomen + Verb ohne Begleiter	Verbverbindungen: Verb + Nomen, Verb + Adjektiv, Verb + Verb mit Begleiter, Verb + Verb ohne Begleiter, Verb + Verb übertragene Bedeutung	Adjektivverbindungen: Adjektiv + Nomen, Adjektiv + Adjektiv, Adjektiv + Verb mit nicht wörtlicher Bedeutung, Adjektiv + Verb mit wörtlicher Bedeutung	–
Dass-das	Dass-das im Satz; in einfachen Sätzen	Dass-das am Satz-anfang, dass-das in längeren Sätzen	Dass-das mit Kommasetzung; im Satz und am Satzanfang; längere Sätze	–

du über die Wortarten wissen solltest“. Über Logfiles kann die Regellesedauer (in Minuten) erfasst werden. Die formativen Tests sind an den Niveaustufen orientiert. Zunächst ist nur der Test auf der Niveaustufe 1 sichtbar. Wird ein formativer Test mit 90 % korrekten Antworten oder mehr abgeschlossen (*mastery criterion*), wird der formative Test auf der nächsten Niveaustufe sichtbar. Das *mastery criterion* wurde

bei 90 % angesetzt, sodass man in Anbetracht der hohen Ratewahrscheinlichkeit bei geschlossenen Testaufgaben auch mit Sicherheit sagen kann, dass eine Fähigkeit beherrscht wird.

Nach jedem Testversuch erhalten die Schülerinnen und Schüler ein einfaches, mit einem passenden Smiley versehenes Feedback, ob der Test bestanden ist oder nicht. Direkt im Anschluss daran können die einzelnen Testaufgaben noch einmal detailliert eingesehen werden. Über die Logfiles kann ermittelt werden, wie lange sich eine Schülerin bzw. ein Schüler mit diesem elaborierten Feedback beschäftigt (Feedbacklesedauer).

Ein nicht erfolgreicher Testversuch kann frühestens nach 15 min wiederholt werden. Damit wird verhindert, dass Schülerinnen und Schüler die Tests ohne Lernaktivitäten so oft wiederholen, bis sie ihn bestehen. Vor der nächsten Testwiederholung können Trainings durchgeführt werden. Auch die Trainings sind nach Niveaustufen geordnet und werden erst sichtbar, wenn ein erster Testversuch gemacht wurde. Analog zu den Tests variiert auch die Anzahl der Aufgaben pro Training von Kurs zu Kurs, um ungefähr zeitlich gleich umfangreiche Trainings gewährleisten zu können. Bei jedem Training können maximal 100 Trainingspunkte erreicht werden. Wenn ein Training beispielsweise aus 10 Aufgaben bestand, ergab die auf Anhieb korrekte Bearbeitung aller 10 Trainingsaufgaben 100 Punkte. Falsch oder nicht auf Anhieb korrekt bearbeitete Trainingsaufgaben führten zu Abzügen. Eine erst im zweiten Anlauf korrekte Trainingsaufgabe ergibt in diesem Beispiel noch 5 Punkte. Für weitere Lösungsversuche gab es keine Punkte mehr. Pro Niveaustufe werden in der Regel mehrere, thematisch geordnete Trainings angeboten.

Die Testwiederholungen sind Pseudo-Paralleltests (Schmidt-Atzert und Amelang 2012). Für jeden formativen Test werden die Testitems aus einer Itemdatenbank zufällig gezogen und angeordnet. Damit wird verhindert, dass die Schülerinnen und Schüler durch bloße Erinnerungseffekte erfolgreich sind. Aufgrund der hohen Menge an Einzeltests bei 8 Kursen mit 3–4 Niveaustufen ist die Anzahl der Items in der Datenbank ca. zwei- bis dreimal so hoch wie die Items pro Test (z. B. je 10 Testitems pro formativem Test im Kurs Satzglieder bei 27, 29 und 29 Items in der Datenbank für Niveau 1, 2 und 3). Jeder Wiederholungstest enthält damit im Schnitt ein Drittel der Items des vorangehenden Tests.

In dieser Studie soll untersucht werden, wie Schülerinnen und Schüler die elaborierten Rückmeldungen der formativen Tests nach dem Prinzip des *mastery measurement* nutzen und wie sich dies auf den Lernfortschritt innerhalb der Kurse auswirkt. Im Vergleich zu den in Abschn. 2.3 skizzierten Arbeiten soll dabei eine Mikroperspektive eingenommen werden. Wie agieren die Schülerinnen und Schüler innerhalb einzelner Test-Folgetest-Sequenzen mit einem bewusst streng gewählten *mastery criterion* von 90 %? Aufgrund der bisherigen Befunde der Feedbackforschung sowie der Designprinzipien für formative Leistungsmessungen werden folgende Vermutungen geprüft:

1. Schülerinnen und Schüler, die nach einem nicht bestandenen Test das Feedback auf Itemebene (Feedbacklesedauer) nutzen (*corrective feedback*), werden vor einer erneuten Testwiederholung eher Regeln nachlesen und Trainings absolvieren.

2. Die Feedbacknutzung (Feedbacklesedauer) führt in Kombination mit den nachfolgenden Lernaktivitäten (Regeln nachlesen und Trainings absolvieren) zum Lernfortschritt innerhalb einer Test-Folgetest-Sequenz (Test nicht bestanden, Folgetest bestanden).

Um diese Forschungsfragen beantworten zu können, müssen die Zusammenhänge zwischen den einzelnen Interaktionen innerhalb einer Test-Folgetestsequenz (Test, Folgetest, Feedback, Regelnutzung, Trainingsnutzung), Kursmerkmalen (Kursthema, Niveaustufe, Anzahl bisheriger Testversuche) und Schülermerkmalen (Geschlecht, Jahrgangsstufe, Schulart) modelliert werden. Viele dieser Variablen sind bereits kategorial und auch bei metrischen Variablen wie den Test- und Folgetestwerten ist vor dem Hintergrund des *mastery measurements* eine Dichotomisierung (bestanden vs. nicht bestanden) und damit eine Behandlung als kategoriale Variable sinnvoll. Gleiches gilt für die Nutzerinteraktionen mit der Lernumgebung. Hinzu kommen weitere, auswertungstechnische Schwierigkeiten, die sich in den Vorstudien bereits deutlich zeigten. Beispielsweise führte das sehr heterogene Nutzungsverhalten zu extrem schiefen Verteilungen der Variablen (Maier et al. 2017). Feedback wird entweder nicht genutzt und wenn es genutzt wird, ist die Nutzungsdauer schief verteilt.

Eine Datenauswertung mit herkömmlichen Verfahren, die sehr hohe Anforderungen an die Verteilung der Variablen und die Beziehungen zwischen den Variablen (Linearität) stellen (z. B. multiple Regressionsanalyse, Mehrebenenanalyse) kommt deshalb nicht in Frage. Aus diesem Grund ist es ein methodisches Ziel dieser Studie, die Forschungsfragen aufgrund einer Modellierung aller Variablen mit einem Bayes'schen Netzwerk (BN) zu beantworten. Eine Besonderheit von BNs ist die mathematisch stringente und informationstechnisch effiziente Modellierung von vielen Variablen mit komplexen, nichtlinearen Beziehungen. BNs werden deshalb bereits seit Jahrzehnten in unterschiedlichsten Forschungs- und Anwendungsfeldern, in denen komplexe Modellierungen mit unterschiedlichen Variablenarten und nicht-linearen Beziehungen notwendig sind, eingesetzt (Uusitalo 2007; Nagarajan et al. 2013). Gerade auch für die Analyse von Nutzerdaten aus digitalen Lernumgebungen (z. B. Garcia et al. 2007; DiCerbo 2016) oder zur Modellierung von Wissenszuwächsen bzw. zur Steuerung der Lernaktivitäten in intelligenten tutoriellen Systemen (z. B. Hooshyar et al. 2015) haben sich BNs bewährt.

4 Untersuchungsmethoden

4.1 Stichprobe

Die hier analysierten Daten entstanden im Schuljahr 2017/18. In diesem Zeitraum haben 3 Realschulen und 2 Gymnasien mit insgesamt 18 Klassen in den Jahrgangsstufen 6, 7, 8 und 9 den Moodle-Kurs genutzt. Nach der Datenbereinigung (z. B. Löschen angemeldeter Schülerinnen und Schüler ohne spätere Kursteilnahme) konnten Daten von 407 Individuen (220 Schülerinnen und 187 Schüler) in die Auswertung einbezogen werden. Für diese Personen liegen insgesamt 1754 einzelne Kursbear-

beitungen vor. Das sind im Schnitt etwas über vier Kursbearbeitungen pro Person. Die Kursbearbeitungen verteilen sich prozentual wie folgt: Wortarten (18 %), Groß- und Kleinschreibung (15 %), Kommasetzung (15 %), Satzglieder (12 %), Zeitformen (12 %), Dass-Das (11 %), Getrennt- und Zusammenschreibung (9 %), Dehnung und Schärfung (9 %).

4.2 Erhebungsinstrumente

Vorwissenstest und Deutschnote Als Maße für die Einschätzung des Vorwissens wird die Deutschnote ($M: 2,68$, $SD: 0,81$) und ein auf die Inhalte des Moodle-Kurses bezogener, selbst erstellter Vorwissenstest genutzt. Der Vorwissenstest musste vor Kursbeginn von allen Schülerinnen und Schülern absolviert werden. Er bestand aus insgesamt 20 Aufgaben, mit denen das Niveau 2 aller Kurse abgebildet wird (Gesamtpunktzahl: 129, $M: 62,3$ Punkte, $SD: 16,1$ Punkte, $Max: 90$ Punkte). Die interne Konsistenz ist ausreichend hoch (Cronbach's Alpha = 0,93). Die Punktzahl aus dem Vorwissenstest und die Deutschnote hängen erwartungsgemäß negativ miteinander zusammen ($r = -0,42$; $p < 0,001$). Ebenso gibt es deutliche Vorwissensunterschiede je nach Schulart und Jahrgangsstufe. Aufgrund der geringen Klassenzahl wurden drei Gruppen gebildet: Gruppe 1 sind die vier 6. Klassen aus den beiden Gymnasien. Gruppe 2 sind sieben Klassen aus der Jahrgangsstufe 6 und 7 der Realschulen. Gruppe 3 sind 8 Realschulklassen der Jahrgangsstufe 8 und 9. Eine einfaktorielle Varianzanalyse zeigt deutliche Mittelwertunterschiede beim Vorwissen ($F = 14,94$, $p < 0,001$). Am besten schneiden die Sechstklässler der beiden Gymnasien ab ($M = 67,4$), gefolgt von den Acht- und Neuntklässlern ($M = 64,7$) und den Sechst- bzw. Siebtklässlern ($M = 57,2$).

Formative Tests Die innerhalb des Moodle-Kurses zu absolvierenden formativen Tests sollen möglichst kurz sein, sodass die Schülerinnen und Schüler schnell eine Rückmeldung erhalten und innerhalb einer Unterrichtsstunde Regeln nachschlagen, üben und eventuell eine Testwiederholung durchführen können. Um die Testdauer möglichst gering zu halten, wurden pro Test zwischen 10 und 20 geschlossene Items eingesetzt. Abb. 1 zeigt ein Histogramm zur Testdauer aller durchgeführten, formativen Tests. Die Nutzung des Feedbacks gehört dabei nicht zur Testzeitdauer. Die mittlere Testdauer liegt bei 4,33 min ($N = 4984$, $SD = 3,45$). Es gibt jedoch Unterschiede zwischen den Kursen. Die höchste mittlere Testdauer mit fast 7 min findet sich beim Kurs Wortarten ($N = 878$ Tests, $M = 6,84$ min, $SD = 4,43$ min). Für alle Kurse liegen damit formative Kurzttests vor, deren Bearbeitung innerhalb von 5 min und bei sehr gründlicher Bearbeitung bzw. erhöhtem Schwierigkeitsgrad innerhalb von 10 min möglich ist.

Durch die konsequent an kleinschrittigen, curricular relevanten Lernzielen orientierte Entwicklung der Lernumgebung entstanden viele, sehr kurze Tests, deren Reliabilität im Vergleich zu Testverfahren mit größerer Itemanzahl geringer sein muss. Dennoch waren die internen Konsistenzen der formativen Tests in der Vorgängerversion zumindest zufriedenstellend (Maier 2017; Maier et al. 2017). Hinzu kommt, dass jedes Item über mehrere Entwicklungsjahre hinweg im Hinblick auf Schwierigkeit und Item-Test-Korrelation überprüft und wenn notwendig ausgetauscht oder

Abb. 1 Testdauer aller formativen Tests

verändert wurde. Die Validität der formativen Tests ist ebenfalls schwierig zu bestimmen. Es gibt kein standardisiertes Testverfahren, das die in den Kursen behandelten Lerninhalte abprüft. Zwei Analysen geben jedoch Hinweise auf die curriculare Validität der formativen Tests. Einmal wird ermittelt, ob die in den Kursen auf Anhieb (d.h. Test beim ersten Versuch bestanden) erzielten Niveaustufen mit dem Vorwissen sowie der Deutschnote korrelieren. Das Ergebnis des Vorwissenstest korreliert mit $r=0,27$ (Spearman Rangkorrelation, $p<0,001$) mit dem auf Anhieb erzielten Niveau. Die Korrelation des Vorwissens mit dem am Ende der Kursbearbeitung maximal erreichten Niveau (unter Niveau 1: 23,5 %, N1: 28,5 %, N2: 20,5 %, N3: 24,0 %, N4: 3,5 %) ist zwar immer noch signifikant, jedoch wesentlich geringer ($r=0,16$, $p<0,001$), was bei einem Lernzuwachs innerhalb des Moodle-Kurses zu erwarten war.

Am Beispiel des Kurses Wortarten wird die curriculare Stufung der Niveaus geprüft (Tab. 2). Die relative Häufigkeitsverteilung zeigt, dass ungefähr ein Drittel der Sechstklässler im Gymnasium aber auch der Acht- und Neuntklässler in der Realschule den Wortartentest auf Niveaustufe 1 (Grundschulniveau) nicht bestehen. Dagegen ist dieser Prozentsatz bei den Sechst- und Siebtklässlern in der Realschule doppelt so hoch. Ähnlich ist die Verteilung bei den auf Anhieb erreichten Niveau 1-Tests. Die curriculare Stufung des Kurses wird dann bei Niveau 2 deutlich. Hier gibt

Tab. 2 Auf Anhieb erreichte Niveaustufen im Kurs Wortarten nach Schülergruppen

Auf Anhieb erzielte Niveaustufe	Realschule JGS 6, 7 ($n=82$) (in %)	Realschule JGS 8, 9 ($n=161$) (in %)	Gymnasium JGS 6 ($n=66$) (in %)
Keine	64,6	34,1	34,8
N1	26,8	47,2	54,5
N2	7,3	15,5	7,6
N3	1,2	3,1	3,0

es keine Unterschiede zwischen leistungs- und vorwissensstarken Gymnasiasten und den Realschülern in der Jahrgangsstufe 6. Dafür wirkt sich der curriculare Vorsprung der Acht- und Neuntklässler in der Realschule aus.

4.3 Aufbereitung der Daten aus dem Moodle-Kurs

Das Lernmanagementsystem Moodle stellt Lehrenden recht übersichtlich Daten zu Testbewertungen zur Verfügung. Für eine Analyse der Lernaktivitäten im Rahmen einer Forschungsarbeit sind diese Formate jedoch sehr ungünstig, weil z. B. Logfiles zu Lernaktivitäten, Testbewertungen und Bewertungen auf Itemebene in verschiedenen Formaten vorliegen und zunächst mühsam aufbereitet und aggregiert werden müssen. Dem Moodle-Kurs wurden deshalb zunächst folgende Daten entnommen:

- Für jeden der 8 Kurse die Testergebnisse für jede Niveaustufe (Schülerkürzel, Zeit, Dauer, Punkte von 100)
- Für jeden der 8 Kurse alle Trainingsergebnisse (Schülerkürzel, Zeit, Dauer, Punkte von 100)
- Files mit den Log-Daten für jede Klasse, um die Lesedauer der Regeln rekonstruieren zu können (für jeden Kurs wurden die Regeln in einer Moodle-Komponente „Bücher“ hinterlegt)
- Datei zur Eingangsbefragung: Geschlecht, Deutschnote, Vorwissenstest

Die Datenaufbereitung erfolgte über mehrere Schritte im Programm Python und kann aus Platzgründen hier nicht näher dargestellt werden (siehe Anlage 1). Das Ergebnis der Datenaufbereitung sind drei Dateien auf je unterschiedlichen Aggregationsebenen:

- Schülerdatei ($N=407$) mit Angaben zu Geschlecht, Schulart, Jahrgangsstufe, Vorwissen, Deutschnote (Grundlage der Stichprobenbeschreibung)
- Kursverlaufsdatei mit Angaben zu den einzelnen Bearbeitungen ($N=1754$): z. B. auf Anhieb erreichtes Niveau, maximal erzielttes Niveau
- Datei mit den Test-Folgetest-Sequenzen ($N=5405$): Für jede Testwiederholung innerhalb einer Niveaustufe innerhalb eines Kurses werden Testwert, Folgetestwert (wenn vorhanden), Informationen zu Feedback, Übungen und Regeln sowie alle Kontextvariablen der Stichproben- und Verlaufsdatei aufgelistet.

Im Mittel wurden pro Niveaustufe 2,3 Testwiederholungen durchgeführt. Die Bandbreite der Testwiederholungen lag zwischen einem und 5 Testversuchen. Ca. zwei Drittel der Fälle in dieser Datei sind echte Test-Folgetest-Sequenzen. Bei einem Drittel liegt kein Folgetest vor, z. B. weil der Test bestanden wurde und die Schülerin bzw. der Schüler zur nächsten Niveaustufe übergeht. Um die Feedbacknutzung und die Nutzung der Folgeaktivitäten im Falle eines bestandenen Tests zu untersuchen, werden diese Fälle dennoch mit aufgeführt.

Die Variablen zur Beschreibung der Lernaktivitäten sind extrem rechtsschief verteilt. Die mittlere Sequenzdauer liegt bei ca. 5 min. Es gibt jedoch zahlreiche Fälle, in denen die Schülerinnen bzw. Schüler eine Testwiederholung erst nach Tagen oder Wochen durchführten. Die schiefen Verteilungen bei der Feedbacklesedauer sowie der Regellesedauer und den Trainingspunkten ist darauf zurückzuführen, dass bei

dem überwiegenden Teil der Sequenzen diese Elemente nicht genutzt wurden. Am geringsten ist die Nutzung der Regelbücher. Im Fall der Nutzung des elaborierten Feedbacks, der Regelbücher und der Trainings sind die Mittelwerte in einem für den Kursaufbau zu erwartenden Bereich.

4.4 Analyse der Daten

Die Analyse der Daten erfolgt mit einem Bayes'schen Netz (BN). Hierfür wurde das R-Paket *bn-learn* verwendet (Scutari 2010). Ein BN ermöglicht die Modellierung komplexer Zusammenhänge zwischen sehr vielen Variablen und gehört zu den Standardverfahren künstlicher Intelligenz (Pearl 1988). Die Abhängigkeiten zwischen Variablen (*nodes*) werden durch Pfeile (*arcs*) symbolisiert. Eine Variable kann dabei sowohl von verschiedenen anderen Variablen (*parents*) abhängen als auch weitere Variablen (*childs*) beeinflussen. Ausgeschlossen sind lediglich zyklische Netzstrukturen. Im Vergleich zu anderen Verfahren für die Modellierung komplexer Zusammenhänge (z. B. Strukturgleichungsmodelle) werden die Zusammenhänge zwischen Variablen in einem BN über bedingte Wahrscheinlichkeiten modelliert. Unter Anwendung des Satzes von Bayes ist es möglich, die beobachtete Wahrscheinlichkeitsverteilung einer Variablen in Abhängigkeit vorgegebener Ursachen (*parents*) in eine bedingte Wahrscheinlichkeitstabelle zu überführen. Ein vollständiges BN besteht am Ende aus den empirisch ermittelten Randverteilungen aller Variablen, einer azyklischen Netzstruktur und der aposteriori ermittelten Wahrscheinlichkeitsverteilung des BN in Form von bedingten Wahrscheinlichkeitstabellen für alle Variablen mit *parents*. Ob man die Pfeile zwischen den Netzknoten kausal interpretieren darf oder nicht, hängt von sachlogischen Erwägungen ab. Das BN beschreibt lediglich die statistischen Abhängigkeiten und Unabhängigkeiten zwischen Variablen. Bei der Modellentwicklung kann man bekannte, designbedingte oder vermutete Abhängigkeiten vorgeben und von den Strukturlernalgorithmen prüfen lassen, ob die Abhängigkeiten bestehen bleiben und welche Abhängigkeiten noch bestehen.

5 Ergebnisse

Im Ergebnisteil wird zunächst die Entwicklung des BN beschrieben. Anschließend werden die beiden Hypothesen geprüft. Hierfür werden sowohl deskriptive Befunde (Kreuztabellen) als auch auf dem BN basierende Wahrscheinlichkeitsabfragen (sog. *conditional probability queries*) berichtet. Eine inferenzstatistische Absicherung der Befunde erfolgt über Chi-Quadrat-Tests, die sowohl die Signifikanz von Häufigkeitsunterschieden in den Kreuztabellen als auch das Vorhandensein von Abhängigkeiten (Pfaden) im BN anzeigen. Eine inferenzstatistische Absicherung der Unterschiede zwischen verschiedenen Wahrscheinlichkeitsabfragen ist allerdings nicht möglich. Vielmehr muss diskutiert werden, ob bestimmte Wahrscheinlichkeiten bzw. Differenzen zwischen Wahrscheinlichkeiten für den Untersuchungsgegenstand bedeutsam sind oder nicht.

Tab. 3 Diskretisierte metrische, ordinal- und nominalskalierte Variablen mit Kategorienhäufigkeiten

Variable	Kategorien	Absolute Häufigkeiten	Relative Häufigkeiten
Schulart	Realschule	4751	0,88
	Gymnasium	654	0,12
Deutschnote	Note 5 und 6	98	0,02
	Note 4	578	0,14
	Note 3	1732	0,43
	Note 2 und 1	1620	0,41
Vorwissen (VW)	Weniger als 60 Punkte	1275	0,32
	60–68 Punkte	1243	0,32
	69–74 Punkte	691	0,17
	Mehr als 74 Punkte	819	0,20
Kurs	Dass-das	426	0,08
	Dehnung Schärfung	326	0,06
	Getrennt- und Zusammenschreibung	486	0,09
	Groß- und Kleinschreibung	891	0,16
	Kommasetzung	967	0,18
	Satzglieder	728	0,13
	Wortarten	1013	0,19
	Zeitformen	568	0,11
Testdauer (relativ zum Mittelwert)	Keine Angabe (Abgabebutton nicht geklickt)	354	0,07
	Sehr kurz	1173	0,22
	Kurz	1308	0,24
	Lang	2324	0,43
Tageszeit	Sehr lang	246	0,05
	Vor 9 Uhr	1586	0,29
	9 bis 13 Uhr	2173	0,40
	13 bis 18 Uhr	1237	0,23
Niveau	Nach 18 Uhr	409	0,08
	1	1099	0,28
	2	1364	0,33
	3	1266	0,33
Testzeitpunkt (TZP)	4	299	0,05
	1	1886	0,47
	2 und 3	1363	0,34
	4 und höher	779	0,19
Testwert (TW)	Deutlich nicht bestanden (nb0: unter 60 % korrekt)	1033	0,28
	Knapp nicht bestanden (nb1: 60–89 % korrekt)	2296	0,55
	Bestanden (best: >90 % korrekte Aufgaben)	699	0,17

Tab. 3 (Fortsetzung)

Variable	Kategorien	Absolute Häufigkeiten	Relative Häufigkeiten
Folgetestwert (FTW)	Kein Folgetest (kein)	1297	0,32
	Deutlich nicht bestanden (nb0: unter 60 P)	532	0,14
	Knapp nicht bestanden (nb1: 60–89 P)	1184	0,28
	Bestanden (best: >90 P)	1015	0,26
Feedbacklesedauer	Keine Feedbacknutzung (FB0)	1735	0,45
	Feedbacklesedauer 1 min (FB1)	1345	0,33
	Feedbacklesedauer >1 min (FB2)	948	0,22
Trainingspunkte	Keine Training bzw. Training unter 20 Punkte (TR0)	2779	0,69
	Trainingspunkte: 21–200 (TR1)	672	0,16
	Über 200 Trainingspunkte (TR2)	577	0,15
Regel-nutzung	Keine Regelnutzung (RE0)	3580	0,90
	Regelnutzung 1 min (RE1)	188	0,04
	Regelnutzung >1 min (RE2)	260	0,06

5.1 Entwicklung des Bayes'schen Netzes für den Datensatz mit den Test-Folgetest-Sequenzen

Als Vorarbeit müssen alle Variablen in möglichst sinnvoll gestufte, nominalskalierte Variablen überführt werden (Diskretisierung). Die Vorgehensweise bei der Diskretisierung hat dabei Konsequenzen für die Interpretation der Befunde (Uusitalo 2007; Nagarajan et al. 2013). Wenn die Beziehungen zwischen den metrischen Variablen linear sein sollten, handelt man sich mit der Diskretisierung einen Verlust an Testpower ein. Für die in den Forschungsfragen abgebildeten Beziehungen werden jedoch eher nicht lineare Zusammenhänge erwartet. Auch die Anzahl der Intervalle und die Intervallgrenzen wirken sich aus. Je höher die Anzahl der für eine Diskretisierung gewählten Intervalle ist, desto weniger Beziehungen findet man bei der Anwendung der Strukturlernalgorithmen bzw. desto mehr Fälle werden benötigt, um Beziehungen sichtbar machen zu können. In der Regel sollten die Intervallgrenzen plausibel nachvollziehbar sein und so gewählt werden, dass jedes Intervall auch eine bestimmte Menge an Fällen enthält.

Tab. 3 zeigt die gewählten Kategorienstufen und die absoluten sowie relativen Häufigkeiten. Die Variablen Kurs und Schulart liegen bereits nominalskaliert vor. Ordinalskalierte Variablen wie Deutschnote oder Niveau werden bei der Netzentwicklung nominalskaliert behandelt. Die Kategorisierung der Testdauer bezieht sich auf eine z-Standardisierung der Testdauer innerhalb der Tests eines Kurses und einer Niveaustufe. Zum Verständnis der Funktionsweise von Bayes'schen Netzwerken muss an dieser Stelle erwähnt werden, dass die Ordnung der Kategorienstufen bei der Diskretisierung ordinalskalierter oder metrisch skaliert Variablen für die Modellierung eines Bayes'schen Netzwerkes keine Rolle spielt. Alle Kategorienstufen, auch wenn sie in Tab. 3 zur besseren Übersicht noch geordnet dargestellt werden (z.B. für das Vorwissen oder die Deutschnote), werden in der anschließenden Modellierung nominalskaliert behandelt.

Mit dem Strukturlernalgorithmus *hill-climbing* von *bn-learn* wurde ein erstes Netz ohne Vorgabe von Pfaden erstellt. Man kann sich damit einen Überblick über die Zusammenhänge verschaffen (Abb. 2). Der Algorithmus liefert einen gerichteten, azyklischen Graphen, der aus 12 Knoten sowie 16 Vorschlägen für statistische Abhängigkeiten besteht (Mittlere Größe der Markov-Blankets: 3,0, benötigte Einzeltests: 242, BIC: -66.742,21). Die meisten Pfade gehen bereits in die richtige Richtung und bestätigen Annahmen (z.B. Pfeile vom Kurs zu Niveau, Testwert, Feedback und Regel, Pfeile von Niveau zu Testzeitpunkt, Pfeil von Testwert zu Folgetestwert, Pfeil von Feedback zu Training).

Ein Teil der Pfeile zeigt jedoch in eine sachlich nicht mögliche Richtung (z.B. von Testwert zu Niveau). Über eine sog. *blacklist* müssen deshalb bestimmte Pfeile ausgeschlossen werden. Ein weiteres Problem ist, dass nicht alle für die Beantwortung der Forschungsfragen relevanten Beziehungen vom Lernalgorithmus erkannt werden. Der Lernalgorithmus ist Score-basiert und sucht nur so lange nach Abhängigkeiten bis er ein Netz mit dem maximalen Fit-Index findet. Jeder zusätzliche Netzwerkpfad erhöht den Score und wird damit „bestraft“. Wenn es jedoch sachlogische Gründe gibt, können dem Netz über eine sog. *whitelist* weitere Pfade hinzuge-

Abb. 2 Erstes Netz ohne Einbeziehung von Expertenwissen mit teilweise sachlogisch unmöglichen Pfaden. Anmerkungen: VW Vorwissen, TW Testwert, FTW Folgetestwert, TZP Testzeitpunkt

fügt werden. Hierfür wurden sinnvolle Abhängigkeiten zwischen den diskretisierten Variablen der Test-Folgetest-Sequenz (Testwert, Feedback, Training, Regelnutzung und Folgetestwert) mittels Chi-Quadrat-Tests geprüft (vgl. Anlage 2 im Online-Anhang). Alle bilateral geprüften Abhängigkeiten sind signifikant, jedoch unterschiedlich stark (Cramer's V). Die Kontingenzkoeffizienten (Cramer's-V) der nicht in der *white-list* vorgegebenen Abhängigkeiten im Netz liegen im Bereich von schwachen bis mittleren Abhängigkeiten (0,1–0,4) (vgl. Anlage 2 im Online-Anhang). Alle signifikanten Zusammenhänge werden in einer *whitelist* für den nächsten Durchgang des Lernalgorithmus als Pfeile gesetzt. Das in Abb. 3 dargestellte Netz ist das Ergebnis eines weiteren Durchgangs des Strukturlernalgorithmus mit den Vorgaben der *blacklist* und der *whitelist*. Das Netz hat 13 Knoten, 23 Pfeile und eine mittlere Größe der Markov-Blankets von 5,23 (BIC: $-78.848,69$, benötigte Einzeltests: 216). Die Zunahme der mittleren Größe der Markov-Blankets ist ein Hinweis auf die Erhöhung der Vernetzung.

Um die Auswirkungen des Wegnehmens und Hinzufügens von Knoten und Pfeilen in der hier dargestellten Netzentwicklung beurteilen zu können, wird ein modifiziertes BN2 ohne Nutzung von Strukturlernalgorithmen modelliert (vgl. Anlage 3 im Online-Anhang). In *bnlearn* werden hierzu die Knoten und Pfade direkt vorgegeben. Im Vergleich zum BN1 hat das BN2 keine Variable Tageszeit. Zudem wurden Zusammenhänge ergänzt, die aufgrund der Befunde der Feedbackforschung hätten erwartet werden können, jedoch vom Strukturlernalgorithmus für das BN1 nicht vorgeschlagen wurden: Pfeile vom Vorwissen zu Feedbacklesedauer, Regellesedauer und Trainingspunkten, Pfeile vom Kurs zur Regellesedauer und den Trainings-

Abb. 3 Für die Analysen verwendetes Bayes'sches Netz (BN1)

punkten sowie ein Pfeil von der Testdauer zum Testwert. Dieses modifizierte BN2 (12 Knoten, 27 Pfeile, mittlere Größe der Markov-Blankets: 7,0, BIC: -95.422,15) hat eine höhere Komplexität als das BN1, sollte bei den Wahrscheinlichkeitsabfragen aber dennoch annähernd gleiche Werte liefern, weil sich das Hinzufügen nicht wirksamer Pfade zumindest nicht substanziell auf die übrige Netzstruktur auswirken darf. Im weiteren Verlauf der Ergebnisdarstellung werden deshalb zum Vergleich die bedingten Wahrscheinlichkeitsabfragen mit dem BN2 in Klammern berichtet.

In einem letzten Schritt müssen für jeden Knoten (Variablen) der beiden BN die bedingten Wahrscheinlichkeiten berechnet werden (*fitting*). Die Forschungsfragen können dann mit sog. *conditional probability queries* (*cpquery*) beantwortet werden. Bei diesen Abfragen wird die Auftretenswahrscheinlichkeit eines definierten Ereignisses unter der Bedingung einer bestimmten Evidenz ermittelt. Die für einzelne Knoten festgelegten Evidenzen (diskrete Ausprägungen dieser Variablen) werden in einem iterativen Verfahren auf das ganze Netz „hochgerechnet“. D.h. es findet eine schrittweise Anpassung der Randwahrscheinlichkeiten der von diesen Knoten abhängigen Knoten statt. Dieser iterative Prozess führt bei mehrfacher Wiederholung zu leicht unterschiedlichen Ergebnissen. Aus diesem Grund wird jeder *cpquery* hundertmal durchgeführt und es wird der Mittelwert dieser 100 Werte berichtet.

5.2 Trainings und Regelnutzung

Die für die Entwicklung des BN1 durchgeführten Chi-Quadrat Tests zeigten bereits, dass die Trainingspunkte und die Regelnutzung von der Feedbackesedauer abhängen. Die Häufigkeiten der Regelnutzung (RE1 oder RE2) nach einem nicht bestandenem Test unterscheiden sich in Abhängigkeit der Feedbacknutzung (Cramer's $V=0,09$). Deutlich stärker als die Regelnutzung hängen die Trainingspunkte (TR1 und TR2) nach einem nicht bestandenem Test von der Feedbacknutzung ab (Cramer's $V=0,24$). Tab. 4 verdeutlicht diese Abhängigkeiten. Es werden jeweils die Wahrscheinlichkeitsabfragen in beiden Bayes'schen Netzen aufgeführt. Dies spricht für eine valide Modellierung der Zusammenhänge zwischen den Variablen. Auch wenn die Regelnutzung insgesamt sehr gering ausfällt, kann man zumindest erkennen, dass Designprinzipien der formativen Leistungsmessung unter bestimmten Umständen wirksam werden.

5.3 Lernfortschritte innerhalb der Sequenz

Der Lernfortschritt innerhalb einer Test-Folgetestsequenz liegt nicht mehr als einzelne Variable vor, sondern wird im BN1 und BN2 als Kombination zwischen einem nicht bestandenem Test und einem bestandenem Folgetest operationalisiert. Die Wahrscheinlichkeit für einen Lernfortschritt im BN1 (bestandener Folgetest bei nicht bestandenem Test) liegt bei 27,5 %, den Folgetest knapp nicht zu bestehen bei 32,6 % und den Folgetest deutlich nicht zu bestehen bei 16,9 % (Vergleich BN2: 27,3 %, 32,6 % und 17,0 %). Durch die Operationalisierung des Lernfortschritts über eine Variablenkombination im BN können Abhängigkeiten zwischen Lernfortschritt sowie Regelnutzung und Trainingspunkten auch nicht mehr mit einfachen Kreuz-

Tab. 4 Wahrscheinlichkeiten für Regellesedauer (RE) und Trainingspunkte (TR) bei nicht bestandenem Test in Abhängigkeit der Feedbacklesedauer (FB)

	Kein FB	FB1 (bis 1 min)	FB2 (über 1 min)
RE1 oder RE2	BN1: 6,4 % BN2: 6,6 %	BN1: 10,3 % BN2: 10,5 %	BN1: 15,7 % BN2: 16,7 %
TR1 (bis 100P.)	BN1: 12,2 % BN2: 12,0 %	BN1: 16,2 % BN2: 16,3 %	BN1: 24,0 % BN2: 24,3 %
TR2 (über 100P.)	BN1: 6,8 % BN2: 7,3 %	BN1: 13,6 % BN2: 13,6 %	BN1: 32,4 % BN2: 33,4 %

tabellen deskriptivstatistisch untersucht werden. Die Forschungsfragen werden nur über Wahrscheinlichkeitsabfragen im BN beantwortet.

Das BN1 und einzelne Chi-Quadrat-Tests zeigen, dass der Lernfortschritt nicht mit dem Testzeitpunkt zusammenhängt. D. h., bei häufigeren Testwiederholungen wird die Wahrscheinlichkeit eines bestandenen Folgetests bei nicht bestandenem Test nicht höher. Der Lernfortschritt hängt auch nicht mit der Feedbacknutzung zusammen. Die Wahrscheinlichkeiten für das Auftreten eines Lernfortschritts liegen bei allen drei Feedbacknutzungskategorien zwischen 26 und 29 % (analog für BN2).

Der Lernfortschritt bei Feedbacknutzung hängt allerdings mit der Regelnutzung zusammen (Zeile 1 in Tab. 5). Die Wahrscheinlichkeit für einen Lernfortschritt ist höher, wenn das Feedback gelesen und anschließend die Regeln noch einmal wiederholt werden. Gleiches gilt für die Trainings (Zeile 2 in Tab. 5). Auch hier erhöht sich die Wahrscheinlichkeit für einen Lernfortschritt leicht, wenn bei Feedbacknutzung auch Trainings durchgeführt werden.

Jetzt werden alle drei Aktivitäten (Feedbacknutzung, Regelnutzung, Trainings) kombiniert, um den Lernfortschritt zu erklären (Zeile 3 in Tab. 5). Die höchste Wahrscheinlichkeit für einen Lernfortschritt ergibt sich bei Feedbacknutzung (FB1 oder FB2), Regelnutzung (RE1 oder RE2) und moderatem Training (TR1). Dieser Befund deutet darauf hin, dass ein kurzes Training sinnvoll ist, wenn zuvor das Feedback und die Regeln zur Kenntnis genommen wurden.

Diese Abfrage wird noch einen Schritt weiter verfeinert, indem die Regelnutzung variiert wird (Zeile 4 in Tab. 5). Die Wahrscheinlichkeit für einen Lernfortschritt steigt auf 43,1 % bei Feedbacknutzung (FB1, FB2), hoher Regelnutzung (RE2) und moderatem Training (TR1). Bei Feedbacknutzung (FB1, FB2), kurzer Regelnutzung (RE1) und moderatem Training (TR1) liegt die Wahrscheinlichkeit für einen bestandenen Folgetest lediglich bei 28,7 %.

6 Diskussion der Befunde und Ausblick

Die Studie hatte das Ziel zu untersuchen, ob Schülerinnen und Schüler in einer Moodle-Lernumgebung Feedback und Lernaktivitäten nutzen, um einen nicht bestandenen, formativen Test mit einer hohen Bestehenshürde (*mastery assessment*) erfolgreich wiederholen zu können. Die Analysen zeigten, dass Schülerinnen und Schüler, die das Feedback intensiver nutzen, sich dann auch verstärkt mit den Regeln und den Trainings auseinandersetzen. Dieser Befund ist mit Studien zu elaboriertem

Tab. 5 Wahrscheinlichkeiten für Lernfortschritt (Test nicht bestanden, Folgetest bestanden) bei Feedbacknutzung (FB1, FB2) und verschiedenen Kombinationen von Regellesedauer (RE) und Trainingspunkten (TR)

<i>(FB1 oder FB2) und RE0</i>	<i>(FB1 oder FB2) und RE1</i>	<i>(FB1 oder FB2) und RE2</i>
BN1: 27,6 %	BN1: 32,9 %	BN1: 34,2 %
BN2: 27,6 %	BN2: 32,2 %	BN2: 34,5 %
<i>(FB1 oder FB2) und TR0</i>	<i>(FB1 oder FB2) und TR1</i>	<i>(FB1 oder FB2) und TR2</i>
BN1: 26,8 %	BN1: 29,3 %	BN1: 31,5 %
BN2: 27,0 %	BN2: 29,9 %	BN2: 31,1 %
<i>(FB1 oder FB2) und TR0 und (RE1 oder RE2)</i>	<i>(FB1 oder FB2) und TR1 und (RE1 oder RE2)</i>	<i>(FB1 oder FB2) und TR2 und (RE1 oder RE2)</i>
BN1: 33,9 %	BN1: 36,8 %	BN1: 28,9 %
BN2: 33,8 %	BN2: 36,7 %	BN2: 29,1 %
<i>(FB1 oder FB2) und TR1 und RE0</i>	<i>(FB1 oder FB2) und TR1 und RE1</i>	<i>(FB1 oder FB2) und TR1 und RE2</i>
BN1: 28,8 %	BN1: 28,7 %	BN1: 43,1 %
BN2: 28,6 %	BN2: 28,3 %	BN2: 43,2 %

Feedback konform (z. B. Kluger und DeNisi 1996; Shute 2008; Van der Kleij et al. 2012) und ist ein wichtiger Hinweis für die Lernwirksamkeit einer formativen Leistungsmessung. Ein für die Weiterentwicklung des Lernangebots wichtiger Hinweis ist die insgesamt geringe Regelnutzung. Es gibt eine Dominanz der Trainings im Vergleich zum Nachlesen der Regeln. Dies ist insofern problematisch, weil sich bei der Prüfung der Forschungsfrage 2 gezeigt hat, dass ein moderates Training in Kombination mit einer ausführlichen Regelnutzung am ehesten zu einem Lernfortschritt innerhalb der Sequenz beiträgt. Ebenso zeigte sich, dass pure Testwiederholungsstrategien nicht erfolgreich sind (vgl. Faber et al. 2017; Maier et al. 2017; Faber und Visscher 2018).

Diese Ergebnisse haben zunächst einmal praktische Konsequenzen für die Optimierung der Moodle-Lernumgebung. In einer weiterentwickelten Variante (www.masteryx.de) wurde das Nachlesen der Regeln durch eine einfachere und ansprechendere Darstellung verbessert. Die Trainings wurden zudem im Umfang deutlich reduziert und direkt mit der Regeldarstellung gekoppelt, sodass die Schülerinnen und Schüler auch während der Trainingsaktivität die Regeln jeweils vor Augen haben. Zudem wurde ein Frühwarnsystem für Lehrkräfte etabliert. Die Lehrkräfte können in einer Echtzeitdarstellung erkennen, wann eine Schülerin oder ein Schüler einen formativen Test mehr als dreimal ohne Erfolg wiederholt hat. Somit können instruktionale Hinweise wesentlich zielgerichteter und früher erfolgen.

Die prinzipielle Struktur der hier untersuchten Lernumgebung lässt sich auf viele weitere, digitale Lernmaterialien mit Elementen formativer Leistungsmessung übertragen. Die Befunde legen die Vermutung nahe, dass diese Lernmaterialien nicht unbedingt zu effektiven Wiederholungen der Instruktion führen. Vielmehr wird es in vielen Fällen lediglich zu einer Erhöhung der Test- bzw. Übungsfrequenz kommen (vgl. Faber und Visscher 2018). Damit wären die mit formativer Leistungsmessung verknüpften Effekte auf Lernzuwächse lediglich Artefakte. Von formativer Leistungsmessung kann man nur dann sprechen, wenn die Rückmeldungen auch zu instruktionalen Konsequenzen führen (Black und Wiliam 2009). Die in der Stu-

die umgesetzte Mikroperspektive auf Lernaktivitäten innerhalb von Test-Folgetest-Sequenzen könnte in weiteren Studien für die Untersuchung der Lerneffizienz von digitalen, formativen Leistungsmessungen mit *mastery measurements* genutzt werden.

Eine methodische Stärke der Studie ist die Feinkörnigkeit der Datenanalyse und die externe Validität der Befunde. Die Daten stammen aus einem realen Einsatzszenario des Moodle-Kurses und reflektieren damit alle möglichen Einflüsse der Schulpraxis. Damit sinkt allerdings die interne Validität der Befunde, weil keine Kontrolle von Variablen durch experimentelle Variation oder aufwändige Erfassung von weiteren Variablen möglich war. Vor allem motivationale Aspekte (z. B. Selbstwirksamkeit, subjektive Bewertung des Nutzens der digitalen Lernumgebung) und situative Gegebenheiten (z. B. im Unterricht zur Verfügung stehende Zeit für die Bearbeitung der Kurse, Unterstützung durch die Lehrkraft, zusätzliche Instruktionen, Einbettung der digitalen Lernumgebung in den Unterricht usw.) spielen wahrscheinlich eine große Rolle, wenn man die Nutzung des Feedbacks, der Trainings und Regeln besser verstehen möchte.

Auch die Analyse der Zusammenhänge mit einem BN hat Vor- und Nachteile. Ein Vorteil ist die Robustheit der Methode bei sog. *noisy data* bzw. *real-world data*, wie sie hier vorliegen. Die Vergleiche der Wahrscheinlichkeitsabfragen zwischen BN1 und BN2 zeigten zudem, dass ein BN stabil bleibt, wenn nicht wirksame Pfade hinzugefügt werden. Ein Nachteil ist, dass die sehr verschachtelte Mehrebenenstruktur der Daten nur teilweise Berücksichtigung findet. Kontextvariablen wie Schulart oder Kurs fließen zwar ein, die Klassenzuordnung konnte jedoch nicht berücksichtigt werden. Damit lassen sich beispielsweise Effekte der Implementation durch die Lehrkraft nicht erfassen.

Eine weitere, methodische Einschränkung ist die Messgenauigkeit der formativen Tests. Um eine unterrichtspraktikable Implementation der Lernumgebung gewährleisten zu können, war die Itemanzahl der Tests gering, was notwendigerweise zu niedrigen Reliabilitäten der Tests führt. Damit ist aber auch die statistische Power der Signifikanzprüfungen schwächer, d. h. es werden weniger Effekte sichtbar. Dies ist ein Dilemma, das sich im Rahmen von formativen Leistungsmessungen auf der Basis von (sehr vielen) *mastery measurements* nicht prinzipiell lösen lässt.

Aufgrund der Fokussierung in diesem Artikel wurde auf eine fachdidaktische Analyse der Befunde verzichtet. In weiteren Studien soll jedoch eine Betrachtung der Effekte innerhalb einzelner Kurse und Niveaustufen erfolgen. Hierfür sind allerdings wesentlich mehr Beobachtungspunkte notwendig. Für die Entwicklung eines BN gibt es zwar aus mathematischer Sicht keine Mindeststichprobengröße. Man kann mit einer bestimmten Anzahl an Beobachtungen (Fällen) beginnen und bei zunehmenden Fällen das BN kontinuierlich optimieren. Es gibt jedoch gewisse Plausibilitätsüberlegungen bezüglich der Anzahl notwendiger Beobachtungen. Je komplexer das BN, d. h. je mehr Parameter (bedingte Wahrscheinlichkeiten) zu lernen sind, desto mehr Fälle sind notwendig.

Perspektivisch gesehen ist zudem angedacht, das BN als Frühwarnsystem innerhalb der digitalen, formativen Leistungsmessung zu nutzen (vgl. Macfadyen und Dawson 2010). Computerprogramme ermöglichen mittlerweile die Echtzeit-Abfrage von bedingten Wahrscheinlichkeiten innerhalb einer Lernumgebung. Damit könn-

te beispielsweise die Lehrkraft informiert werden, wenn eine Schülerin oder ein Schüler ein wenig effizientes Lernverhalten in einem Kurs zeigt. Dies würde dazu beitragen, dass die in formativen Leistungsmessungen generierten Rückmeldungen möglichst effektiv von den Lehrkräften für die Optimierung der nachfolgenden Lehr-Lernprozesse genutzt werden könnten.

Funding Open Access funding provided by Projekt DEAL.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

Literatur

- Bangert-Drowns, R. L., Kulik, C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, 61, 213–238.
- Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education*, 18(1), 5–25.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31.
- 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.
- Bokhove, C., & Drijvers, P. (2012). Effects of a digital intervention on the development of algebraic expertise. *Computers & Education*, 58, 197–208.
- DiCerbo, K. (2016). Integrating data across digital activities. *Learning, Media and Technology*, 41(2), 233–251.
- Faber, J. M., & Visscher, A. J. (2018). The effects of a digital assessment tool on spelling achievement. Results of a randomized experiment. *Computers & Education*, 122, 1–8.
- Faber, J. M., Luyten, H., & Visscher, A. J. (2017). The effects of a digital formative assessment tool on mathematics achievement and student motivation: results of a randomized experiment. *Computers & Education*, 106, 83–96.
- Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, 57(6), 488–500.
- García, P., Amandi, A., Schiaffino, S., & Campo, M. (2007). Evaluating bayesian networks' precision for detecting students' learning styles. *Computers & Education*, 49(3), 794–808.
- Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: a review of the literature. *Computers & Education*, 57, 2333–2351.
- Hancock, T. E., Thurman, R. A., & Hubbard, D. C. (1995). An expanded control model for the use of instructional feedback. *Contemporary Educational Psychology*, 20(4), 410–425.
- Hasselhorn, M., Schneider, W., & Trautwein, U. (Hrsg.). (2014). *Lernverlaufsdiagnostik. Jahrbuch der pädagogisch-psychologischen Diagnostik: Tests und Trends*. Göttingen: Hogrefe.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- Hebbeker, K., & Souvignier, E. (2018). Formatives Assessment im Leseunterricht der Grundschule – Implementation und Wirksamkeit eines modularen, materialgestützten Konzepts. *Zeitschrift für Erziehungswissenschaft*, 21(4), 735–765.

- Hooshyar, D., Ahmad, R.B., Yousefi, M., Yusop, M.D., & Horng, S.-J. (2015). A flowchart-based intelligent tutoring system for improving problem-solving skills of novice programmers. *Journal of Computer Assisted Learning*, 31, 345–361.
- Jia, J., Chen, Y., Ding, Z., & Ruan, M. (2012). Effects of a vocabulary acquisition and assessment system on students' performance in a blended learning class for English subject. *Computers & Education*, 58, 63–76.
- Kingston, N., & Nash, B. (2011). Formative assessment: a meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, 30(4), 28–37.
- 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, 254–284.
- Macfadyen, L.P., & Dawson, S. (2010). Mining LMS data to develop an “early warning system” for educators: a proof of concept. *Computers & Education*, 54, 588–599.
- Maier, U. (2010). Formative Assessment – Ein erfolgversprechendes Konzept zur Reform von Unterricht und Leistungsmessung? *Zeitschrift für Erziehungswissenschaft*, 13(2), 293–308.
- Maier, U. (2017). Verlaufsdiagnostik und Förderung im Lernbereich Sprachbetrachtung in der Sekundarstufe I mit Moodle. *Empirische Sonderpädagogik*, 2017(2), 98–115.
- Maier, U., Wolf, N., & Randler, C. (2016). Effects of a computer-assisted formative assessment intervention based on multiple-tier diagnostic items and different feedback types. *Computers & Education*, 95, 85–98.
- Maier, U., Ramsteck, C., & Hoffmann, K. (2017). Formative Leistungsdiagnostik und Learning Analytics: Entwicklung, Nutzung und Optimierung eines onlinebasierten Kurses für die Diagnostik und Förderung von Grundwissen im Kompetenzbereich Sprachbetrachtung. *Zeitschrift für Erziehungswissenschaft*, 20(4), 728–747.
- McLaughlin, T., & Yan, Z. (2017). Diverse delivery methods and strong psychological benefits: a review of online formative assessment. *Journal of Computer Assisted Learning*, 33(6), 562–574.
- McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement: so much more is needed. *Practical Assessment, Research & Evaluation*, 18(2), 1–15.
- Nagarajan, R., Scutari, M., & Lèbre, S. (2013). *Bayesian networks in R with applications in systems biology*. New York: Springer.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufmann.
- Rakoczy, K. (2012). Formatives Assessment – theoretische Erkenntnisse und praktische Umsetzung im Mathematikunterricht. In C. Fischer (Hrsg.), *Diagnose und Förderung statt Notengebung?* (S. 73–91). Münster: Waxmann.
- Russell, M. K. (2010). Technology-aided formative assessment of learning: new developments and applications. In H. L. Andrade & G. J. Cizek (Hrsg.), *Handbook of formative assessment* (S. 125–138). New York: Routledge.
- Schmidt-Atzert, L., & Amelang, M. (2012). *Psychologische Diagnostik* (5. Aufl.). Berlin: Springer.
- Scutari, M. (2010). Learning bayesian networks with the bnlearn R-package. *Journal of Statistical Software*, 35(3), 1–22.
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189.
- Sikora, S., & Voß, S. (2017). Konzeption und Güte curriculumbasierter Messverfahren zur Erfassung der arithmetischen Leistungsentwicklung in den Klassenstufen 3 und 4. *Empirische Sonderpädagogik*, 3, 236–257.
- Souvignier, E., Schütze, B., & Hasselhorn, M. (2018). Stichwort – Formatives Assessment. *Zeitschrift für Erziehungswissenschaft*, 21(4), 697–715.
- Timmers, C., & Veldkamp, B. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education*, 56(3), 923–930.
- Timmers, C. F., Braber-van den Broek, J., & van den Berg, S. M. (2013). Motivational beliefs, student effort, and feedback behaviour in computer-based formative assessment. *Computers & Education*, 60, 25–31.
- Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, 203(3/4), 312–318.
- Van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education*, 58, 263–272.
- Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: a meta-analysis. *Review of Educational Research*, 85(4), 475–511.

- Walter, J. (2010). Lernfortschrittsdiagnostik am Beispiel der Lesekompetenz (LDL): Messtechnische Grundlagen sowie Befunde über zu erwartende Zuwachsraten während der Grundschulzeit. *Heilpädagogische Forschung*, 36(4), 162–176.
- 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.