jero

1
2022

Journal for Educational Research Online

Journal für Bildungsforschung Online

WAXMANN

JERO
Journal for Educational Research Online
Journal für Bildungsforschung Online

Executive Editor:
Prof. Dr. Olaf Köller, IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik

Editorial Office:
Dr. Kerstin Schütte, IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik
E-Mail: editorialoffice@j-e-r-o.com

Editors:
Rolf Becker (Bern), Oliver Böhm-Kasper (Bielefeld), Wolfgang Böttcher (Münster), Martin Bonsen (Münster), Wilfried Bos (Dortmund), Hartmut Ditton (München), Cornelia Gräsel (Wuppertal), Stephan Huber (Zug), Harm Kuper (Berlin), Katharina Maag Merki (Zürich), Romain Martin (Luxemburg), Nele McElvany (Dortmund), Florian H. Müller (Klagenfurt), R. Daniel Muijs (Southampton), Jost Reinecke (Bielefeld), Kerstin Schneider (Wuppertal), Claudia Schuchart (Wuppertal), Wolfram Schulz (Camberwell), Birgit Spinath (Heidelberg), Petra Stanat (Berlin), Jan Van Damme (Leuven), Stefanie van Ophuysen (Münster), Andreas Voss (Hamburg), Horst Weishaupt (Frankfurt), Ludger Wößmann (München), Olga Zlatkin-Troitschanskaia (Mainz)

Please subscribe to the e-mail alert at www.waxmann.com/jero if you want to be informed about the publication of each new issue. There is a spring edition and an autumn edition each year. The subscription of the journal is free of charge.

Open access at www.waxmann.com/jero
The articles are published under the licence CC BY-NC-SA 4.0.

ISSN 1866-6671

Waxmann Verlag GmbH 2022
Steinfurter Straße 555, 48159 Münster, Germany
Internet: www.waxmann.com, E-Mail: info@waxmann.com

Setting: Roger Stoddart, Münster

Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01
Volume 14 (2022), No. 1

Inhalt

Liste der Gutachter:innen des 13. Jahrgangs.....5

Call for Papers
Sprachsensibilität in der sprachlichen Bildung?!
Lehrkräftebildung im Spannungsfeld zwischen
Bedarfsorientierung und Etikettierung 6

Editorial

Michael Schurig, Jörg-Tobias Kuhn & Markus Gebhardt
Entwurf, Konstruktion und Analyse von
Lernverlaufsdiagnostik in Schulen 9

Artikel Fokus

Florian Klapproth, Lucas Holzhüter & Tanja Jungmann
Die Prädiktion der Leseleistung von Schülerinnen und
Schülern im Rahmen von Lernverlaufsdiagnostik: Evidenz
für einen Geschlechter-Bias 16

Sterett H. Mercer & Joanna E. Cannon
Validität des automatisierten Learning Progress
Assessments im geschriebenen Englisch für Schüler:innen
mit Lernschwierigkeiten.....39

Christin Vanauer, Sarah Chromik, Philipp Doebler
& Jörg-Tobias Kuhn
Verlaufsdiagnostik arithmetischer Grundkompetenzen:
Messen verschiedene Booklets die gleiche Fähigkeit? 61

Gesa Brunn, Fritjof Freise & Philipp Doebler
Modellierung eines glatten Lernverlaufs und Testung
individueller Abweichungen von einem globalen Verlauf.....89

Sven Anderson, Daniel Sommerhoff, Michael Schurig, Stefan Ufer & Markus Gebhardt
Entwicklung eines Tests zur Lernverlaufsdiagnostik mit schwierigkeitsgenerierenden Merkmalen: Ein Beispiel für die Diagnose grundlegender arithmetischer Fertigkeiten in der Grundschule 122

Artikel Allgemein

Holger Gärtner, Felicitas Thiel, Christopher Kellermann & Max Nachbauer
Ein theoriebasierter Schülerfragebogen für Unterrichtsevaluation 147

Marion Händel, Svenja Bedenlier, Michaela Gläser-Zikuda, Rudolf Kammerl, Bärbel Kopp & Albert Ziegler
Verfügen Studierende über die Mittel und Wege, um während der Coronaviruspandemie zu lernen? Bedarfe für Online-Lehre in einer plötzlich digitalisierten Landschaft 174

Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01
Volume 14 (2022), No. 1

Contents

List of Reviewers of Volume 13 5

Call for Papers
Sprachsensibilität in der sprachlichen Bildung?!
Lehrkräftebildung im Spannungsfeld zwischen Bedarfsorientierung
und Etikettierung 6

Editorial

Michael Schurig, Jörg-Tobias Kuhn, & Markus Gebhardt
Observing Change to Generate Change – Addressing Challenges in
Learning Progress Monitoring 9

Article Focus

Florian Klapproth, Lucas Holzhüter, & Tanja Jungmann
Prediction of Students' Reading Outcomes in Learning
Progress Monitoring: Evidence for the Effect of a Gender Bias 16

Sterett H. Mercer & Joanna E. Cannon
Validity of Automated Learning Progress Assessment in
English Written Expression for Students with Learning Difficulties 39

Christin Vanauer, Sarah Chromik, Philipp Doebler,
& Jörg-Tobias Kuhn
Curriculum-Based Measurement of Basic Arithmetic
Competence: Do Different Booklets Represent the Same Ability? 61

Gesa Brunn, Fritjof Freise, & Philipp Doebler
Modeling a Smooth Course of Learning and Testing
Individual Deviations From a Global Course 89

Sven Anderson, Daniel Sommerhoff, Michael Schurig, Stefan Ufer, & Markus Gebhardt
Developing Learning Progress Monitoring Tests Using Difficulty-Generating Item Characteristics: An Example for Basic Arithmetic Operations in Primary Schools 122

Artikel General

Holger Gärtner, Felicitas Thiel, Christopher Kellermann, & Max Nachbauer
Development of a Theory-Driven Student Questionnaire for Teacher Evaluation 147

Marion Händel, Svenja Bedenlier, Michaela Gläser-Zikuda, Rudolf Kammerl, Bärbel Kopp, & Albert Ziegler
Do Higher Education Students Have the Means to Learn During the COVID-19 Pandemic? Student Demands for Distance Learning in a Suddenly Digital Landscape 174

List of Reviewers of Volume 13

The *Journal for Educational Research Online* (JERO) is fortunate to receive generous support from researchers willing to share their time and expertise. The Executive Editor wishes to acknowledge the following colleagues, who contributed to the processing of manuscripts throughout the year 2021 by providing reviews.

Karen Aldrup

Jürgen Baumert
Rolf Becker
Niels Berkemeyer
Andrea Bernholt
Sascha Bernholt

Bastian Carstensen
Colin Cramer

Jasmin Decristan
Hartmut Ditton
Hanna Dumont

Johanna Fleckenstein

Sabine Glock
Bernadette Gold

Tina Hascher
Marcus Hasselhorn
Doris Holzberger

Thorben Jansen
Monika Jungbauer-Gans

Gisela Kammermeyer
Thilo Kleickmann
Michaela Köller

Jennifer Meyer
Jens Möller
Barbara Muslic

Gabriel Nagy
Marko Neumann

Annika Ohle-Peters

Jan Retelsdorf
Dirk Richter
Alexander Robitzsch

Alfred Schabmann
Sabine Schlag
Tina Seidel

Annika Teerling

Esther Ulitzsch

Thamar Voss

Anne Weidinger
Linda Wirthwein
Jörg Wittwer

Fabian Zehner
Friederike Zimmermann

jero Journal for Educational Research Online
open access

Call for Papers

Schwerpunktheft des *Journals for Educational Research Online* (JERO)

Sprachsensibilität in der sprachlichen Bildung?! Lehrkräftebildung im Spannungsfeld zwischen Bedarfsorientierung und Etikettierung

Um die Bildungschancen aller Schüler:innen durch ein adäquates Maß sprachlicher Bildung zu erhöhen, werden in der Lehrkräftebildung an mehreren Hochschulen u. a. Module implementiert, die sich schwerpunktmäßig mit sprachlicher Heterogenität und sprachsensiblem Unterricht befassen. Zu solchen Modulen zählt u. a. das im Jahr 2009 verankerte Modul „Deutsch für Schülerinnen und Schüler mit Zuwanderungsgeschichte" in NRW oder das Modul „Umgang mit Heterogenität in der Schule" in Bremen. Als gemeinsame Ziele solcher Angebote gelten die Sensibilisierung für die Notwendigkeit der Fokussierung sprachlicher Bildung in allen Fächern, der Erwerb von Fachwissen zu linguistischen Grundlagen und Spracherwerbsprozessen, interkultureller Bildung und Deutsch als Zweitsprache sowie die Weiterentwicklung sprachdiagnostischer, sprachförderlicher und sprachdidaktischer Kompetenzen (Witte, 2017).

Ein gezieltes Angebot sprachförderlicher und -bildender Maßnahmen sowie sprachlicher Hilfen im Unterricht setzt eine Analyse der sprachlichen Fähigkeiten der Schüler:innen sowie eine Identifizierung individueller Stärken und Bedarfe voraus. Die präzise Benennung dieser Bedarfe – speziell im Kontext der Lehrkräftebildung – erfordert jedoch auch im Zusammenhang mit individueller Förderung besondere Sensibilität, da mit jeder personenbezogenen Bezeichnungspraktik die Gefahr einer Etikettierung oder sogar Diskriminierung einhergehen kann (vgl. Res-

Das open access erscheinende Journal for Educational Research Online (JERO) ist ein internationales und interdisziplinäres Forum für die Diskussion bildungswissenschaftlicher Fragestellungen und Themen. Die Zeitschrift veröffentlicht wissenschaftliche, im peer review akzeptierte Originalarbeiten von internationaler Relevanz zu Theorien, Methodik und zu empirischen Studien der Bildungsforschung.

sourcen-Etikettierungs-Dilemma nach Jeuk & Settinieri, 2019; Powell & Pfahl, 2012, bzw. Differenzdilemma nach Kuhn, 2014). Eine Schwierigkeit sprachlicher Bildung besteht daher darin, auf individuelle Bedarfe zu reagieren, ohne bestimmten Lernenden eine pauschale Förderbedürftigkeit zuzuschreiben. Lehrveranstaltungen bieten dabei die Möglichkeit, sowohl auf das sprachsensible Unterrichten als auch auf einen sensiblen Umgang mit Sprache aufmerksam zu machen.

Vor diesem Hintergrund soll der Frage nachgegangen werden, inwiefern die erste Phase der Lehrkräftebildung dazu beitragen kann, Studierende für die Bedeutung und den Gebrauch bestimmter Begriffe zu sensibilisieren und damit einhergehende Perspektiven auf sprachliche Heterogenität und speziell auf Mehrsprachigkeit zu verfestigen oder aufzulösen. Ziel dieses Schwerpunktheftes ist es, sich durch vielfältige Beiträge und eine interdisziplinäre Diskussion dem Feld der Sprachsensibilität in der sprachlichen Bildung zu nähern. Willkommen sind theoretische sowie empirische Beiträge, die sich mit folgenden Fragen und Inhalten beschäftigen:

- Welche Begrifflichkeiten und Konzepte dominieren den Diskurs in der Lehrkräftebildung mit Blick auf sprachliche Heterogenität und insbesondere Mehrsprachigkeit?
  - Inwiefern wird dabei für das Differenzdilemma in der sprachlichen Bildung sensibilisiert?
  - Worauf wird in offiziellen Dokumenten Wert gelegt? Wie positionieren sich dazu die Lehrenden? Was davon erreicht die Studierenden?
- Wie gehen Lehrende und Studierende mit Begrifflichkeiten mit Blick auf mögliche Selbstpositionierungen oder Fremdzuschreibungen im Kontext von Migration und Mehrsprachigkeit um?
- Welche konkreten didaktischen Herangehensweisen lassen sich an den Hochschulen bezüglich der Vorbereitung von Lehramtsstudierenden auf den Unterricht in Klassen mit hoher sprachlicher Heterogenität feststellen? Inwiefern sind diese sprachsensibel?
- Welche Rolle spielen dabei die in den unterschiedlichen Qualifizierungsmaßnahmen in der Lehrkräftebildung vermittelten Inhalte, Begriffe, Konzepte?
- Inwiefern gelingt es, eine Sensibilisierung dafür zu schaffen, dass sprachliche Bildung nicht nur monolingual zu verstehen ist und Potenziale für alle Schüler:innen bietet?
- Inwiefern lässt sich eine Umkehr von einer Defizitorientierung und ausschließlichen Deutschförderung als alleinigem Baustein sprachlicher Bildung hin zu einem holistischen Ansatz der Berücksichtigung aller sprachlichen Ressourcen erkennen?

Mit dem Schwerpunktheft soll zum einen aufgezeigt werden, wie sich die erste Phase der Lehrkräftebildung bereits mit Blick auf einen reflexiven Umgang mit Bezeichnungspraktiken entwickelt hat. Zum anderen soll diskutiert werden, wie sie sich weiter optimieren lässt, um eine Sensibilisierung für den konstruktiven Umgang mit sprachlicher Heterogenität und ihren Potenzialen zu schaffen.

Der Einreichungsprozess ist zweistufig angelegt. Abstracts mit einem Umfang von nicht mehr als 500 Wörtern werden bis **30. Juni 2022** im Dateiformat docx per E-Mail an folgende Adresse erbeten: editorialoffice@j-e-r-o.com. Bitte wenden Sie sich ebenfalls an diese Adresse, sollte Ihnen eine fristgerechte Einreichung des Abstracts nicht möglich sein. Alle eingereichten Abstracts werden von den Herausgeber:innen des Schwerpunktheftes sowie der Schriftleitung des JERO begutachtet. Auf Basis dieser Begutachtung wird die Einreichung von Manuskripten eingeladen.

Die Manuskripte sollen einen Umfang von 50 000 Zeichen (einschließlich Leerzeichen) nicht übersteigen. Formal sind die Richtlinien für Autor:innen zur Vorbereitung von Manuskripten für das *Journal for Educational Research Online* zu beachten. Die Begutachtung durch unabhängige Expert:innen erfolgt dem regulären Verfahren entsprechend im Doppelblindverfahren. Die Entscheidung über die Annahme der Manuskripte zur Veröffentlichung im JERO trifft die Schriftleitung. Die Frist zur Einreichung der vollständigen Manuskripte ist der **1. Dezember 2022**.

Wir freuen uns auf Ihre Einreichungen.
Evghenia Goltsev, Ina Kaplan, Ina-Maria Maahs, Helena Olfert, Marco Triulzi und Christina Winter

Literatur

Kuhn, M. (2014). Vom Allgemeinen und Besonderen. Wissens- und differenzkritische Überlegungen zur Professionalisierung von kindheitspädagogischen Fachkräften in Migrationsgesellschaften. In T. Betz & P. Cloos (Hrsg.), *Kindheit und Profession. Konturen und Befunde eines Forschungsfeldes* (S. 130–144). Beltz Juventa.

Powell, J. J. W. & Pfahl, L. (2012). Sonderpädagogische Fördersysteme. In U. Bauer, U. H. Bittlingmayer & A. Scherr (Hrsg.), *Handbuch Bildungs- und Erziehungssoziologie* (S. 721–739). Springer VS.

Settinieri, J. & Jeuk, S. (2019). Einführung in die Sprachdiagnostik. In S. Jeuk & J. Settinieri (Hrsg.), *DaZ-Handbücher: Band 2. Sprachdiagnostik Deutsch als Zweitsprache: Ein Handbuch*. (S. 3–20). De Gruyter Mouton.

Witte, A. (2017). Sprachbildung in der Lehrerausbildung. In M. Becker-Mrotzek & H.-J. Roth (Hrsg.), *Sprachliche Bildung: Band 1. Sprachliche Bildung – Grundlagen und Handlungsfelder* (S. 351–363). Waxmann.

Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01.01
Volume 14 (2022), No. 1, 9–15
CC BY-NC-SA 4.0 Waxmann 2022

Michael Schurig, Jörg-Tobias Kuhn, & Markus Gebhardt

Observing Change to Generate Change – Addressing Challenges in Learning Progress Monitoring

1. Introduction

Learning progress monitoring (LPM) as a form of formative assessment is an important element for the prevention and treatment of scholastic difficulties and lagging academic development (e.g., in reading or mathematics; Fuchs, 2004). Being prominent in the field of special education, it also addresses general education and psychology.

Repeated measurements of specific individual learning outcomes allow for evaluating ongoing learning processes and can promote those by giving feedback to students and/or teachers (Black & Wiliam, 2003). Thus, students are able to reflect on their learning success and teachers are supported in their effective decision making (Deno, 1985).

Different instruments have been supported by research in terms of reliability and validity (Reschly et al., 2009; Wayman et al., 2007), and have had particularly positive results for children with special educational needs (Stecker et al., 2005). More recently, computer-based instruments have additionally boosted the potential of instruments for learning progress monitoring (Russell, 2010). Computer-based instruments can reduce the time required for assessments, facilitate the creation of parallel test versions, provide automatic feedback to students and teachers, and supply additional important diagnostic information (e.g., response times).

Curriculum-based measurement (CBM), which has been developed in the United States to solve academic difficulties in special education (Deno, 2003), is

Dr. Michael Schurig, ORCID: 0000-0002-7708-0593, Research in Inclusive Education, TU Dortmund University, Emil-Figge-Straße 50, 44227 Dortmund, Germany
email: michael.schurig@tu-dortmund.de

Prof. Dr. Jörg-Tobias Kuhn, ORCID: 0000-0002-4399-9569, Methods of Educational Research, TU Dortmund University, Emil-Figge-Straße 50, 44227 Dortmund, Germany
email: jörg-tobias.kuhn@tu-dortmund.de

Prof. Dr. Markus Gebhardt, ORCID: 0000-0002-9122-0556, Learning Disability Pedagogy including Inclusive Pedagogy, University of Regensburg, Sedanstraße 1, 93055 Regensburg, Germany
email: markus.gebhardt@paedagogik.uni-regensburg.de

probably the most prominent form of LPM. It serves as a namesake for many specific tests and can serve as a prime example for the ongoing development of instruments for learning progress monitoring. However, the CBM framework only focuses on a selection of the tools that were developed to foster data-driven learning. Additionally, many of the alternative approaches and tests do not take into account curricular contents but address domain-specific subdimensions, skills, competences, or robust indicators directly and are more clearly focused on the support of learning. Therefore, we decided to use the broader term of learning progress monitoring. Though analytical frameworks of learning progress like the response-to-intervention approach (Bradley et al., 2005) or the student tracking system (van der Kooij, 2003) and test systems like easyCBM (Alonzo et al., 2006), DIBELS (Kaminski & Good, 1996), CODY (Schwenk et al., 2017), quop (Souvignier et al., 2014), or Levumi (Gebhardt et al., 2016; Mühling et al., 2019) differ with respect to the respective substantial outcomes, goals, and assessment references, it is clear that for a meaningful usage there is need for a strong conceptual framework to support learning (Bennett, 2011). This includes the identification of characteristics and components of the entity under scrutiny (over time) and a clear understanding of how those work together.

2. On the Necessity of Measurement Quality

Like other tests, learning progress monitoring instruments need to possess classical psychometric quality criteria such as objectivity, reliability, and validity (Good & Jefferson, 1998). But, additionally, further criteria have to be regarded. Depending on the test under consideration, it may be a necessity that essential unidimensionality is confirmed (e.g., Anderson et al., 2017), that the test is sensitive to change in learning, and that it is usable in an economic, easy, and simple way for educators, and that the results can be interpreted by non-statisticians (Klauer, 2011). Only in the case of psychometrically sound test instruments can learning progress assessment provide information that can be used in instructional decision-making processes (Gebhardt et al., 2019; Rohwer, 2015; Shapiro, 2013; Wilbert & Linnemann, 2011). Likewise, the tests must be short enough with good reliability so that they take up little learning time, can be used well in the classroom, and do not place too much of a burden on the children (Schurig et al., 2021).

Despite the great interest in this topic, there is only a limited number of instruments available for learning progress monitoring in schools. Both conceptual and methodological challenges can help explain this dearth of instruments. Therefore, it is essential to gather up-to-date research and conduct new studies relating to learning progress monitoring.

3. Manuscripts

The central question in the creation of appropriate measures must always be based on what is needed so that the measures can be used for decision-making. The contributions included in this special issue all address different challenges in creating and exploiting meaningful measures of progress monitoring in instructional practice.

In order to be considered for decision-making, *the practitioners must be able to evaluate the data provided in an unbiased manner*. But studies show that in-service and preservice teachers often have difficulties using the provided data presentation. Florian Klapproth, Lucas Holzhüter, and Tanja Jungmann address this with a study that examines whether preservice teachers' interpretations of measures of progress monitoring are biased by gender stereotypes, the preservice teachers' gender affects their predictions and extrapolations, and whether the insertion of a trend line or lowered data variability diminishes the gender bias. They assessed this by implementing a digital experimental design, using student vignettes of an oral reading fluency assessment over a period of 11 weeks. The evaluation of the effects was done within the framework of an analysis of variance. The results showed that the preservice teachers – female and male alike – were prone to favoring girls. This bias was attenuated when a trend line was plotted to assist in interpretation. Based on these results, possible strategies for the preparation of feedback are discussed and the need for research on data literacy is addressed.

In order to be considered for decision-making, *quality criteria of change measures must be established*. The most import quality dimension of any meaningful measure is the validity, but validity in itself is an elusive construct (Newton & Shaw, 2013). Sterett H. Mercer and Joanna E. Cannon articulate an explicit and testable scoring inference (Kane, 2013) for writing quality that is informed by developmental writing theory and may be able to capture writing quality rather than fluency, as it is often done due to the short times allowed to write in tests. This is integrated in an automated learning progress assessment in written expression. The authors present validity arguments and preliminary validity evidence for the automated approach to learning progress assessment for English written expression in Grades 2–12. To do so, the performance of the automated progress monitoring and its inference was compared to a hand-coded curriculum-based measure with four points of measurement, and measures of writing quality. Acceptable correlations with standardized writing subtests assessing spelling and grammar, but not the subtest assessing substantive quality were found and growth could be observed between fall and spring. These findings are interpreted as evidence that the automated processes can be used to efficiently score narrative writing samples for progress monitoring.

In order to be considered for decision-making, *multiple test forms for learning progress monitoring are needed*. Christin Vanauer, Sarah Chromik, Philipp Doebler, and Jörg-Tobias Kuhn address the question of the equivalence of test book-

lets. Booklet effects were related to between-child variance in generalized mixed linear models, testing the assumption that equal scores from different randomly generated booklets are representing the same latent ability.

At both grade levels considered, the between-booklet variance turned out to be very small relative to the between-child variance. This varied across task types but was not substantial compared with the variance in student performance. The findings could be replicated in two intervention groups in which a response-to-intervention approach was implemented. Thus, the booklets can be considered equivalent for typical application purposes. The implications are discussed with respect to trajectory diagnostics and the construction of equivalent test booklets.

In order to be considered for decision-making, *scores in computerized assessment systems have to be linked even with non-overlapping item sets*. Computerized assessment systems are able to include large item pools from which random items can be drawn as representations of the item universe. But without item overlap or strong theoretical assumptions simple sum scores are no longer comparable directly because the item parameters are assessed separately. Gesa Brunn, Fritjof Freise, and Phillip Doebler introduce the smooth growth and linear deviations Rasch model (SGLDRM) as an extension of Rasch's item response theory model for binary test data to address this matter. In this model, a smoothed global learning parameter is estimated based on splines and individual linear deviations from the global learning can be detected in intercept and slopes. The methodology is illustrated with data from an online dyscalculia assessment and training and recommendations on the implementation as well as possible extensions are given.

In order to be considered for decision-making, *a systematic and testable theoretical rationale has to be established*. In order to generate feedback that is compatible with the content and to facilitate the creation of parallel test forms, it is helpful to be able to map the content characteristics of the test in a differentiated manner on an item basis as well. The study by Sven Anderson, Daniel Sommerhoff, Michael Schurig, Stefan Ufer, and Markus Gebhardt investigates the item characteristics of an item-generating system for learning processes monitoring of basic arithmetic operations. Three hypothesized item characteristics are tested in a linear logistic test model. All characteristics are found to be meaningful providing important insights into how to address the challenges in the development of future LPM tests in mathematics.

4. Conclusions

In order to closely monitor changes in learning over time, strong assumptions concerning observed constructs of interest, quality criteria of psychological tests, potential dimensionality, fairness, and test economy should be taken into account. But the challenges for practically applicable procedures do not end at this point; in fact, only the foundation for the further validation and substantive interpretation

of growth values have been created. Taking a broad view, the five interrelated studies in this special issue show the various challenging issues of the research field.

The theoretical constructs must be defined sharply enough to be mapped in a short test. At the same time, however, the ability spectrum must be broad enough to adapt test content development to the target population. In this context, it must also be possible to take into account possible external factors that have an influence on the learning of all students, such as holidays (see Brunn et al.).

Outcome measures of learning processes can refer to more than just ratios to thresholds or norm values, but may include non-linear growth or even profiles. The presentation of the results and the level of detail of the results can therefore be a challenge. How specific can and should a presentation of results be for test takers and test administrators so that a benefit can be derived (see Anderson et al.)? Additionally, possible biases have to be taken into account in the interpretation of test scores. This pertains to both test takers and further stakeholders who are supposed to deal with test scores and integrate them into their professional work, such as teachers who are asked to interpret and extrapolate growth curves (see Klapproth et al.).

Several comparable tests must be provided for formative use, potentially multiplying the number of parameters to be tested. Therefore, determining necessary degrees of measurement invariance over a potentially high number of measurements is another challenge that needs to be addressed (see Vanauer et al.). Related to this is the question of how linkage can be made when test items are potentially used without item-anchoring between measurement points, as is possible with computer-based tests. Moreover, with reference to computer-based tests, the question can be asked to what extent automatic raters are able to emulate human raters in potentially interpretative domains such as language development (see Mercer & Cannon).

There are many hopes associated with systematically implementing formative assessment into individualized learning (e.g., personalized education; Tetzlaff et al., 2021). At the same time, many challenges remain in correctly interpreting and utilizing results from formative approaches. All contributors were able to present innovative approaches for one or even several of the development areas of formative assessment and we are excited that several methodological challenges have been addressed in the contributions to this special issue. The observation of change, as well as fostering change in individual learning remains a demanding task, but our methods are adapting.

References

Alonzo J., Tindal G., Ulmer K., & Glasgow A. (2006). *easyCBM® online progress monitoring assessment system*. Behavioral Research and Teaching, University of Oregon. http://easycbm.com

Anderson, D., Kahn, J. D., & Tindal, G. (2017). Exploring the robustness of a unidimensional item response theory model with empirically multidimensional data.

Applied Measurement in Education, 30(3), 163–177. https://doi.org/10.1080/08957347.2017.1316277

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Black, P., & Wiliam, D. (2003). 'In praise of educational research': Formative assessment. *British Educational Research Journal, 29*(5), 623–637. https://doi.org/10.1080/0141192032000133721

Bradley, R., Danielson, L., & Doolittle, J. (2005). Response to intervention. *Journal of Learning Disabilities, 38*(6), 485–486. https://doi.org/10.1177/00222194050380060201

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219–232. https://doi.org/10.1177/001440298505200303

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*(3), 184–192. https://doi.org/10.1177/00224669030370030801

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188–192. https://doi.org/10.1080/02796015.2004.12086241

Gebhardt, M., DeVries, J. M., Jungjohann, J., Casale, G., Gegenfurtner, A., & Kuhn, T.-J. (2019). Measurement invariance of a direct behavior rating multi item scale across occasions. *Social Sciences, 8*(2), Article 46. https://doi.org/10.3390/socsci8020046

Gebhardt, M., Diehl, K., & Mühling, A. (2016). Online Lernverlaufsmessung für alle SchülerInnen in inklusiven Klassen. www.LEVUMI.de [Online learning progress monitoring for all students in inclusive classes. www.LEVUMI.de]. *Zeitschrift für Heilpädagogik, 67*(10), 444–453.

Good, R., & Jefferson, G. (1998). Contemporary perspectives on curriculum-based measurement validity. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 61–88). Guilford.

Kaminski R. A., & Good R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 25*(2), 215–227.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Klauer, K. J. (2011). Lernverlaufsdiagnostik – Konzept, Schwierigkeiten und Möglichkeiten [Diagnosing the course of learning – Concept, difficulties and chances]. *Empirische Sonderpädagogik, 3*(3), 207–224.

Mühling, A., Jungjohann, J., & Gebhardt, M. (2019). Progress monitoring in primary education using Levumi: A case study. In H. Lane, S. Zvacek, & J. Uhomoibhi (Eds.), *CSEDU 2019. Proceedings of the 11th International Conference on Computer Supported Education* (pp. 137–144). SCITEPRESS – Science and Technology Publications.

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18*(3), 301–319. https://doi.org/10.1037/a0032969

Reschly, A. L, Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-Based Measurement Oral Reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*(6), 427–469. https://doi.org/10.1016/j.jsp.2009.07.001

Rohwer, G. (2015). Bemerkungen zu einem Testverfahren für Lernfortschritte [Remarks on a test procedure for long-term learning progress]. *Journal for Educational Research Online, 7*(2), 147–156.

Russell, M. K. (2010). Technology-aided formative assessment of learning: New developments and applications. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 125–138). Routledge.

Schwenk, C., Kuhn, J.-T., Gühne, D., Doebler, P., & Holling, H. (2017). Auf Goldmünzenjagd: Psychometrische Kennwerte verschiedener Scoringansätze bei computergestützter Lernverlaufsdiagnostik im Bereich Mathematik [We are going on a gold coin hunt: Psychometric properties of different scorings in computer-based progress monitoring of mathematics ability]. *Empirische Sonderpädagogik, 9*(2), 123–142.

Schurig, M., Jungjohann, J., & Gebhardt, M. (2021). Minimization of a short computer-based test in reading. *Frontiers in Education, 6*, Article 684595. https://doi.org/10.3389/feduc.2021.684595

Shapiro, E. S. (2013). Commentary on progress monitoring with CBM-R and decision making: Problems found and looking for solutions. *Journal of School Psychology, 51*(1), 59–66. https://doi.org/10.1016/j.jsp.2012.11.003

Souvignier, E., Förster, N., & Salaschek, M. (2014). quop: Ein Ansatz internetbasierter Lernverlaufsdiagnostik mit Testkonzepten für Mathematik und Lesen [quop: An internet-based approach to learning progress monitoring in math and reading]. In M. Hasselhorn, W. Schneider, & U. Trautwein (Eds.), *Lernverlaufsdiagnostik* (pp. 239–256). 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. https://doi.org/10.1002/pits.20113

Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing personalized education: A dynamic framework. *Educational Psychology Review, 33*(3), 863–882. https://doi.org/10.1007/s10648-020-09570-w

van der Kooij, R. (2003). Ist das niederländische Schülerfolgesystem (SFS) auch im deutschen Unterricht hilfreich? [Is the student tracking from the Netherlands helpful in German classes, too?]. *Sonderpädagogik, 33*, 106–113.

Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*(2), 85–120. https://doi.org/10.1177/00224669070410020401

Wilbert, J., & Linnemann, M. (2011). Kriterien zur Analyse eines Tests zur Lernverlaufsdiagnostik [Criteria for analyzing a test measuring learning progress]. *Empirische Sonderpädagogik, 3*(3), 225–242.

Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01.02
Volume 14 (2022), No. 1, 16–38
CC BY-NC-SA 4.0 Waxmann 2022

Florian Klapproth, Lucas Holzhüter, & Tanja Jungmann

Prediction of Students' Reading Outcomes in Learning Progress Monitoring: Evidence for the Effect of a Gender Bias

Abstract

Learning progress monitoring (LPM) is an effective tool for teachers to improve students' performance by systematically and quickly responding to achievement data. However, studies show that in-service and preservice teachers often have difficulties using LPM because of lacking graph literacy, especially with high data ambiguity. The present study examines whether (a) preservice teachers are biased by gender stereotypes when predicting students' performance based on progress data, (b) the preservice teachers' gender affected their predictions differentially depending on student gender, and (c) the insertion of a trend line or lowered data variability diminishes the gender bias in predictions. N = 134 preservice teachers received 16 experimental student vignettes online via the internet in random order which depicted the learning progress of boys and girls in oral reading fluency assessment over a period of 11 weeks. Half of the participants were presented with progress data accompanied by a trend line, the other half received progress data only. Results evidenced that preservice teachers were prone to a gender bias favoring girls. The gender bias was attenuated when a trend line was presented or when data variability was low, with male participants benefitting more from the trend line, and female participants benefitting more from low data variability. The adaptation of international training programs to enhance graph literacy and to diminish gender stereotyping in German teachers is recommendable.

Prof. Dr. Florian Klapproth (corresponding author), ORCID: 0000-0002-4598-837X · Lucas Holzhüter, Medical School Berlin, Rüdesheimer Straße 50, 14197 Berlin, Germany
email: florian.klapproth@medicalschool-berlin.de
l.holzhueter@gmx.de

Prof. Dr. Tanja Jungmann, ORCID: 0000-0001-8530-2857, Carl-von-Ossietzky Universität Oldenburg, Ammerländer Heerstraße 114–118, 26129 Oldenburg, Germany
email: tanja.jungmann@uni-oldenburg.de

We confirm that we have reported in the manuscript all measures, conditions, and data exclusions. All materials used for this study can be viewed at Open Science Framework (https://osf.io/hsf3j/).

Keywords
learning progress monitoring (LPM), graph literacy, oral reading fluency, data ambiguity, gender bias, stereotypes

Die Prädiktion der Leseleistung von Schülerinnen und Schülern im Rahmen von Lernverlaufsdiagnostik: Evidenz für einen Geschlechter-Bias

Zusammenfassung
Lernverlaufsdiagnostik stellt ein wirksames Instrument für Lehrkräfte dar, um die Leistung von Schüler:innen zu verbessern, indem Lehrkräfte systematisch und schnell auf die Schülerleistung reagieren können. Allerdings hat sich gezeigt, dass Lehrkräfte häufig Schwierigkeiten haben, Lernverlaufsdaten korrekt zu interpretieren. Wir haben mit der vorliegenden Studie untersucht, ob Lehramtsstudierende dazu neigen, Mädchen besser als Jungen zu bewerten, wenn sie auf Grundlage von Lernverlaufsdaten eine Prognose für künftige Leistungen erstellen müssen. Darüber hinaus haben wir die Hypothesen geprüft, dass der Geschlechter-Bias bei männlichen Lehramtsstudierenden zugunsten männlicher Schüler abgeschwächt ist und dass die Bereitstellung einer Trendlinie in den Lernverlaufsdaten bzw. eine geringere Variabilität der Daten zu einer Verringerung des Geschlechter-Bias führen. Insgesamt N = 134 Lehramtsstudierende erhielten 16 experimentelle Vignetten, in denen der Verlauf der Leseleistung von 8 weiblichen und 8 männlichen Grundschüler:innen über einen Zeitraum von 11 Wochen als Lernverlaufsgraph dargestellt war. Bei der Hälfte der Versuchspersonen waren die Lernverlaufsgraphen durch eine Trendlinie ergänzt. Die Ergebnisse zeigten, dass die Lehramtsstudierenden im Durchschnitt höhere Leistungen für Mädchen im Vergleich zu Jungen prognostizierten und dass der Geschlechter-Bias durch die Trendlinie und durch eine geringe Datenvariabilität abgeschwächt wurde. Wir empfehlen die Adaptation bereits international verwendeter Trainingsmaßnahmen zur Schulung von Lehramtsstudierenden und Lehrkräften in der Interpretation von Lernverlaufsdaten und zur Prävention von systematischen Verzerrungen von Interpretationen von Lernverlaufsdaten.

Schlagworte
Lernverlaufsdiagnostik, graph literacy, Lesekompetenz, Datenambiguität, Geschlechter-Bias, Stereotype

1. Introduction

Learning progress monitoring (LPM) is an increasingly popular method teachers can use to track how students are progressing in basic academic areas such as math, reading, spelling, or writing. It usually entails using quick, frequently ad-

ministered standardized measures to assess students' progress towards a long-term goal (Deno, 1985). Although LPM offers educators a sound foundation for making evidence-based decisions about whether students need help, instructions have to be modified, or teaching goals have to be changed (Hosp et al., 2007), the effectiveness of LPM for improving student achievement appears to be mixed (Ardoin et al., 2013; Christ et al., 2012; Stecker et al., 2005). One reason why LPM alone does not lead to better teaching, which in turn could improve student achievement, is that teachers may have difficulties using progress data to make reasonable predictions about students' future achievement (Van den Bosch et al., 2017). Prediction of student achievement within LPM may be meaningful when teachers want to identify students at risk (Strathmann et al., 2010), or when ineffective interventions should be identified before the end of the school semester (Van Norman & Parker, 2018). Actually, the interpretation of progress data is affected by several factors, for instance data variability (Klapproth, 2018), or the presence versus absence of a trend line (Van Norman et al., 2013). Although several studies showed that girls are expected to have higher competence in reading than boys although performance of both is actually the same (e.g., Plante et al., 2013), little is known about whether teachers are also biased by gender stereotypes when interpreting students' learning progress. In the present study in Germany, we sought to examine whether preservice teachers are prone to a gender bias when predicting students' achievement based on their learning progress. Additionally, we investigated whether the preservice teachers' gender affected their predictions differentially depending on student gender, and we finally assessed two factors that might lower a gender bias in predictions.

2. Theoretical and Empirical Background

2.1 Learning Progress Monitoring

Learning progress monitoring (LPM) is a broad term for progress monitoring systems designed to track the learning achievement progress of students within an academic domain. The measurement of the academic progress of an individual student or a group of students (e.g., a school class) usually serves the purpose of the evaluation of instructions for those students in order to judge whether students have reached a learning goal or instead need additional support (Deno, 2003). LPM involves the frequent administration of short measures of performance in an academic area of interest (e.g., reading, writing, or mathematics). The student's achievement usually is visualized as a graph representing the achievement trajectory of the student over a predetermined period of time. Teachers may inspect the graph to gauge whether the instructional program is effective, whether a student has mastered a topic, or whether a student is expected to perform according to prespecified teaching goals. The latter judgment needs the extrapolation of future

achievements of students from their current LPM progress graphs. This is necessary when teachers want to decide whether or not they should change instruction (Good & Shinn, 1990) to achieve long-term goals in education (Shinn et al., 1989), or when they want to identify students at risk (Hosp et al., 2007; Strathmann et al., 2010). By systematically responding to achievement data with instructional adaptations, LPM could be an effective tool for teachers to improve students' performance. However, teachers often have difficulties to improve their instruction when using LPM (Ardoin et al., 2013; Christ et al., 2012; Stecker et al., 2005). One possible reason is their lack of skills to correctly read and interpret data (Van den Bosch et al., 2017, 2019). Even the use of computer software aiming to help teachers in interpreting the graphs by providing statistics like, for example, the linear trend of the graph, does not result in adequate understanding of the progress data. Teachers often do not use these statistics, even if they have gained experiences with LPM (Espin et al., 2017). Instead, they more frequently rely on their visual inspection of the data (Van Norman et al., 2013). Visual inspection, however, is prone to error (Klapproth, 2006), and, consequently, teachers make mistakes when interpreting visualized progress data.

2.2 Understanding Progress Data

The capability of reading and interpreting progress data depicted as a graph has been called graph literacy (Friel et al., 2001). Graph literacy is defined as an individual's ability to derive meaning from a graph and includes three components: the ability to extract data from the graph ("read the data"), the ability to interpret visualized data ("read between the data"), and the ability to evaluate data ("read beyond the data"). This approach has also been applied to data obtained from LPM (e.g., Van den Bosch et al., 2017). In this context, the ability to read data means that teachers are able to correctly describe the scores and the growth rate of a graph. Reading between data corresponds with interpreting the relation between the actual growth rate and the expected growth rate, whereas reading beyond data means linking data to the instructional context.

All these components of graph literacy matter when teachers use LPM data to adjust their instruction. However, even if teachers can read beyond data, effects on student achievement might still be absent if teachers do not implement modifications in their instructional programs (Stecker et al., 2005).

Graph literacy can be fostered by providing visual aids in the graphical depiction of progress data, which in turn would make judgments in LPM more accurate. A common tool in LPM to facilitate visual judgments of progress data is the presentation of a trend line. This is calculated by ordinary least squares regression such that it fits best through all data points (Van Norman & Christ, 2016). However, whereas visual analysis supplemented with trend lines tends to outperform pure visual analysis, it is still prone to incorrect interpretations (Nelson et al., 2017; Van Norman & Christ, 2016).

2.3 Biases in Interpreting and Predicting Progress Data

Several sources of bias have been identified when teachers use LPM data to interpret or predict student achievement. For instance, LPM data patterns can be ambiguous (Deno, 2013). Ambiguity can occur when features of the graph do not consistently speak in favor or against the existence of a positive or negative trend, for instance, when data cyclicity is present (Brossart et al., 2006). It may also be increased when achievement scores are contaminated by construct-irrelevant variance (Christ et al., 2012), which may make it difficult to visually identify a trend in the data. With LPM data in particular, teachers have difficulty to accurately estimate the rate of improvement when progress data are highly variable (Klapproth, 2018; Nelson et al., 2017; Tindal et al., 1983) or include extreme values (Klapproth, 2018; Nelson et al., 2017). However, if variability in the data is beyond what is generally expected, the confidence that the data allow for judgment and prediction may decrease (Horner & Odom, 2014).

2.4 Is There a Gender Bias in LPM?

Reading skills are essential for individuals to gain an understanding across subject domains in school and hence are an important predictor of their future socioeconomic status (e.g., Ritchie & Bates, 2013). Oral reading fluency directly measures phonological segmentation and recoding as well as fast word recognition and serves as an indicator of overall reading competence (Fuchs et al., 2001). It is one of the skills that is predominantly assessed by LPM (Tindal, 2013). Although boys have consistently been found to obtain lower scores in reading competence than girls across different countries and languages (e.g., Chiu & McBridge-Chang, 2006; McElvany et al., 2017; Mullis et al., 2017), little is known about gender differences in LPM oral reading fluency. Some studies, however, indicated that girls obtained higher scores in curriculum-based measurement of reading fluency than boys (e.g., Kranzler et al., 1999; Yeo et al., 2011).

Boys' lower attainments in reading competence are discussed as being partly a result of a bias due to teachers' gender stereotypes according to which reading is suited more for girls rather than for boys (e.g., Lorenz et al., 2016). Teachers' gender stereotypes adhere to assumptions about student motivation and student working habits (Glock & Kleen, 2017; Jussim & Eccles, 1992).

Stereotypes can be defined as "shared [...] beliefs about traits that are characteristic of members of a social category" (Greenwald & Banaji, 1995, p. 14). Thus, they result from a categorization of individuals into groups according to presumed commonalities. On the one hand, stereotypes may serve as schemas that facilitate social interactions with unknown individuals. On the other hand, stereotypes can also serve as a social norm affecting expectations and behavior toward members of a particular social group (e.g., Schneider, 2004). When a target is categorized

as belonging to a given group, these beliefs are activated (e.g., Macrae et al., 1994; Van Knippenberg et al., 1999).

2.5 Stereotypes Might Reduce Uncertainty in Judgments

According to dual process theories of social judgment (e.g., Fiske et al., 1999; Fiske & Neuberg, 1990), people's judgments of other people occur on a continuum of two concurrent processes. On one end of the continuum, judgments are automatic, quick, effortless, and follow social categories ("girl", "boy", "poor", "rich", etc.); on the other end of the continuum, a process is assumed that is slow, effortful, and voluntarily initiated to overcome and enrich the automatic process by integrating all available and relevant information about the to-be-judged person. However, the use of stereotypic categories may be facilitated if there are salient attributes of a person that fit a certain stereotype (Fiske & Neuberg, 1990), if the motivation or accountability to correctly judge the person is low (Tetlock & Kim, 1987), or if a judging person is uncertain about the correct interpretation of someone's behavior (Campbell, 1967). For instance, if a teacher wants to judge the reading fluency progress of a student who shows ups and downs with no clear linear trend visible in the data, the teacher might be uncertain about the "true" achievement development of the student. Social stereotypical knowledge can act as an interpretive frame when evaluating members of a social group (Hicklin & Wedell, 2005; Hilton & von Hippel, 1996). As a consequence, evaluations of individual group members might be biased in the direction of the knowledge we have about the group as a whole (Spieß & Bekkering, 2020). The more complex or ambiguous a task is, the more likely is the application of stereotypes (Darley & Gross, 1983; Kunda & Sherman-Williams, 1993). Hence, when teachers also know about the student's gender, they might judge a girl performing higher or a boy performing lower on reading skills, respectively. Thus, the gender of the student could provide information, which in turn might reduce uncertainty and ambiguity in judgments (Kossak & Johnson, 2001).

Evidence that coping with ambiguity might result in the use of stereotypes stems, for example, from studies in the medical context where researchers found that students who had a tendency to perceive ambiguous situations as threatening were more likely to make use of stereotypes about patients (Geller et al., 1990) or were more accepting the use of stereotyping in everyday situations (Valutis, 2015) than those who were, comparatively, highly tolerant of ambiguity. Further evidence comes from studies in perception. For instance, both Bar (2003) and Correll et al. (2015) have shown that stereotypes even guided visual processing when the perceived object was ambiguous.

2.6 Teacher-Student Gender Interaction

Student assessment might also depend on teachers' gender. According to the gender-stereotypic model (Martin & Marsh, 2005), boys achieve higher scores in classes taught by males, and girls are better when instructed by female teachers. Teachers might favor students of their own gender (Holmlund & Sund, 2008). Some studies support this hypothesis. For instance, Dee (2007) revealed that in secondary school, boys and girls were evaluated more positively when they were taught by a same-gender teacher rather than by a teacher of the opposite gender. Other studies found that female teachers generally evaluated both boys and girls more positively than male teachers (Ehrenberg et al., 1995), or did not find a teacher-gender bias in assessing male and female students (e.g., Driessen, 2007). In a large-scale study conducted by Neugebauer et al. (2011) on data from IGLU-E (an expansion of the Progress in International Reading Literacy Study, PIRLS, in Germany), no same-gender effect on student outcomes was obtained. Boys did not benefit from male teachers and girls did not benefit from female teachers. However, in a recent experimental study, Klapproth and Fischer (2019) found that female participants evaluated students with more caution and precision and were less optimistic than male participants when recommending them to the tracks of secondary school. Male participants, however, showed a preference for boys, indicating a same-gender bias.

2.7 Research Questions and Hypotheses

To the authors' knowledge, this is the first experimental study that examined effects of social stereotypes on the interpretation of progress data within the frame of LPM. The rationale of the present study was as follows. Since teachers (or preservice teachers) may have stereotypical beliefs about girls and boys (Muntoni & Retelsdorf, 2018), they would presumably expect higher achievements in oral reading fluency for girls than for boys. Additionally, since ambiguity in achievement data should raise the likelihood of the use of stereotypes (Correll et al., 2015; Kossak & Johnson, 2001), we investigated whether data ambiguity would affect achievement prediction, and whether a gender bias in achievement prediction would be smaller when data ambiguity was reduced. Thus, we aimed at testing the following hypotheses:

1) We hypothesized that within LPM predicted achievements in oral reading fluency would be higher for girls than for boys, even when girls and boys show the same progress data.
2) We assumed that when data ambiguity is low, students' gender should affect predictions of achievement to a lesser degree. We sought to realize reduction of data ambiguity by (a) the insertion of trend lines in the graphical visualization of progress data, and (b) lowering data variability. Hence, we tested two interac-

tion hypotheses: the trend line × student gender interaction and the data variability × student gender interaction.

Finally, we examined whether the gender of the participants contributed differentially to oral reading fluency predictions for male and female students, that is, whether there is an interaction between participants' gender and students' gender with respect to their reading fluency predictions. However, with respect to the mixed results reported in previous research we abstained from stating a specific hypothesis.

3. Method

3.1 Participants

Based on previous investigations (cf. Klapproth, 2018, for the effect of data variability, and Klapproth & Fischer, 2019, for the effect of student and participant gender), we expected an average medium to large effect (f = .30, which translates to η^2 = .08) of the independent variables on the participants' predictions. We conducted an a priori power analysis for an analysis of variance (ANOVA) with repeated measures and four groups, using G*Power 3.1 (Faul et al., 2009). When prespecifying f = .30, α = .05, $1-\beta$ = .90, and the correlation among repeated measures with r = .50, power analysis yielded a total sample size of N = 44, which we deemed to be the minimum sample size.

Announcements on social media platforms were used to recruit a sample of preservice teachers enrolled in a primary or secondary school teacher education program at various German universities. A sample of 170 preservice teachers had been recruited for participation in the experiment. However, we excluded 31 participants from subsequent analyses because they dropped out from the study before they were presented with student vignettes or after they received the first vignette but did not continue to participate. Furthermore, five participants were not considered for analyses because they left the study during ongoing presentation of the vignettes (most of them dropped out before having received half of the vignettes). In total, n = 36 participants were dropouts and were excluded from analyses. From the remaining participants, data were complete.

Thus, a total of N = 134 preservice teachers (57.5% female, M_{age} = 24.4 years, SD = 2.5) participated in the study, which were much more than recommended by power analysis. However, we chose the larger sample because estimates of parameters are more precise with larger than with smaller samples. Most of the participants conducted their study program in universities located in North-Rhine Westphalia (35.8%) or Bavaria (32.1%), followed by participants from Berlin (11.9%), Hesse (3.7%), and remaining states (16.5%). They had studied on average for 7.5 semesters (SD = 3.5). No participant reported previous experiences with LPM. Table 1 shows the frequency of the participants' gender across the teacher education

programs, Table 2 shows participants' background information across the experimental groups (trend line versus no trend line).

Table 1: Frequency of the Participants' Gender Across Teacher Education Programs

	Teacher education program	
Participant gender	Primary	Secondary
Male	19	38
Female	34	43
Sum	53	81

Table 2: Participants' Background Information Across the Groups

	Participant gender		Teacher education program	
Group	Male	Female	Primary	Secondary
Trend line	37	30	20	47
No trend line	20	47	33	34
Sum	57	77	53	81

Whereas the participants' gender and their teacher education program were not significantly interrelated, $\chi^2(1) = 1.61$, $p = .205$, there was a significant relationship between the participants' assignments to either experimental group (trend line versus no trend line) and both their gender, $\chi^2(1) = 8.82$, $p = .003$, and their education, $\chi^2(1) = 5.28$, $p = .022$. Male participants and participants enrolled in secondary education programs were more frequently assigned to the trend line group than female participants and participants enrolled in primary education programs.

3.2 Materials and Procedure

The experiment was conducted online on www.soscisurvey.de. The participants could perform the tasks of the experiment on a computer or any other device that was connected to the internet. The study was open for 14 days.

At the beginning of the experiment, the participants were welcomed and were given a brief introduction into learning progress monitoring based on oral reading fluency that had been obtained over a period of time and is depicted visually as a learning development graph. After that, the participants received 16 experimental student vignettes in random order. The vignettes depicted the learning progress of boys and girls in oral reading fluency assessment over a period of 11 weeks. The experimental vignettes were supplemented by the presentation of six distractor vignettes, which were created arbitrarily and served to camouflage the independent variables of the study, because knowledge of the independent variables may af-

fect the participants' responses (Klein, et al., 2012). They mimicked results from curriculum-based oral reading fluency (R-CBM) tests, one of the most common in LPM, which is frequently applied in schools (Ardoin et al., 2013) and which has often been the focus of research (Reschly et al., 2009). Each vignette displayed a graph of student progress data. In line with the R-CBM, the y-axis represented the number of words read correctly (WRC), which could range from zero to 140 (Hosp et al., 2007). The x-axis represented the school week that the test was given and ranged from Week 1 to Week 17, thus spanning a typical school semester. WRC scores were given only for the first 11 weeks. The participants were told that they were to examine the development of oral reading fluency scores over a period of 11 weeks in order to make a prediction of the oral reading fluency score presumably obtained at Week 17. Half of the participants were presented with progress data accompanied by a trend line aiming at facilitating interpretation of the data, whereas the other half of participants were provided with progress data only. The trend line, depicted as a thin (1 mm) dotted black line from Week 1 to Week 11, reflected the linear trend that was estimated by ordinary least squares linear regression analysis. Hence, the occurrence of a trend line was a between-subjects variable. The participants were randomly assigned to both groups.

Independent within-subjects variables were the students' gender and the amount of data variability. Each participant received two sets of different vignettes: eight vignettes, differing in data variability and labeled with a male name, and the same eight vignettes, labeled with a female name. The names used for this study were common for either male or female German students. Common names were chosen to prevent the activation of concepts like a certain socioeconomic background by rarely used names that are especially common in certain social and economic milieus (Gerhards, 2010).

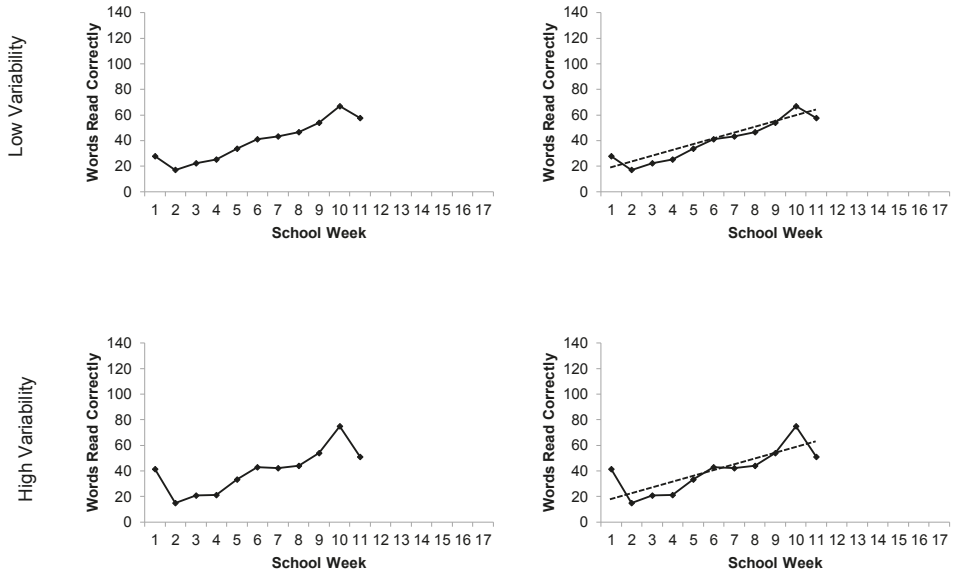
Data variability was either low or high. In the literature, variability of learning progress graphs is often quantified as the standard error of the estimate (SEE), which is defined as the average magnitude of residuals around the trend line obtained from linear regression. According to the literature (e.g., Ardoin & Christ, 2009; Van Norman & Christ, 2016), SEE values usually vary between 5 and 20, where 5 means very low, and 20 very high variability. In our study, the SEE of progress data with high variability was 10.0, whereas in the low variability condition SEE was 5.0.

We used a mixture between a within-subjects design and a between-subjects design because a complete crossing of all variables (meaning that each participant would have received both vignettes with trend lines and without trend lines) would have lacked ecological validity, as in reality teachers are seldom presented with different types of graphs in LPM.

Both low data variability and the presence of a trend line were supposed to reduce the ambiguity of the data. The dependent variable was the participants' prediction of the WRC score in Week 17.

Figure 1 shows four experimental vignettes, either with a trend line or without a trend line for both low and high data variability.

Figure 1: Experimental Vignettes Used in the Study.



Note. The horizontal axis indicates the time (school week) a LPM test was administered; the vertical axis depicts the number of words read correctly within a minute. On the left panel, vignettes are shown without a trend line, on the right panel, vignettes are accompanied by a trend line. Upper vignettes show low data variability, lower vignettes show high data variability.

3.3 Data Analyses

The constant error (i.e., the difference between participants' predictions and the prediction of a linear regression equation) served as a measure of bias (cf. Makridakis et al., 1998). To calculate the constant errors, ordinary least squares linear regression analysis of the given data was used to predict the WRC score at Week 17 in each of the experimental vignettes. The result of the linear regression analysis was then subtracted from participants' prediction. A positive constant error indicates that participants predicted higher achievement relative to the linear regression equation (i.e., positive bias), while a negative constant error indicates that participants predicted lower achievement relative to the linear regression equation (i.e., negative bias). A repeated measures ANOVA with trend line as between-subject factor, and student gender and data variability as within-subject factors was then conducted. Additionally, the participants' gender was included in our analysis as a further independent variable. With that analysis we were able to estimate the extent to which student gender, participant gender, data variability, and the presence or absence of a trend line affected the constant error of prediction. In order to assess whether the participants' teacher education program affected their predictions, we additionally ran correlational analyses between the teacher educa-

tion program (primary vs. secondary education) and the constant error of predictions obtained from all realized conditions.

4. Results

Table 3 displays the means, standard errors, and the 95%-confidence intervals of the constant errors of the predicted WRC scores of each condition. In Table 4, these descriptive statistics are given for each independent variable.

Table 3: Means, Standard Errors (in Parentheses), and 95%-Confidence Intervals (in Brackets) of the Constant Errors of Predicted WRC Scores

Trend line	Participant gender	Student gender			
		Male		Female	
		Variability			
		Low	High	Low	High
No	Male *n* = 20	−0.38 (3.05) [−6.77, 6.01]	−6.45 (2.32) [−11.30, −1.60]	4.85 (3.03) [−1.50, 11.20]	7.47 (3.44) [0.27, 14.67]
	Female *n* = 47	2.36 (1.76) [−1.18, 5.90]	−5.85 (1.52) [−8.90, −2.80]	4.07 (1.72) [0.61, 7.53]	8.89 (1.93) [5.01, 12.77]
Yes	Male *n* = 37	−0.97 (1.25) [−3.51, 1.56]	−3.26 (0.77) [−4.81, −1.70]	−2.03 (1.07) [−4.20, 0.14]	0.25 (1.04) [−1.85, 2.35]
	Female *n* = 30	2.88 (2.70) [−2.65, 8.40]	−3.64 (1.78) [−7.27, 0.00]	2.19 (2.50) [−2.92, 7.31]	5.82 (3.07) [−0.46, 12.10]

Table 4: Means, Standard Errors, and 95%-Confidence Intervals of the Constant Errors of Predicted WRC Scores per Independent Variable

IV		*M*	*SE*	CI
Trend line	No	1.87	1.38	−0.85, 4.59
	Yes	0.16	1.27	−2.35, 2.66
Participant gender	Male	−0.07	1.43	−2.89, 2.76
	Female	2.09	1.20	−0.29, 4.47
Student gender	Male	−1.91	0.87	−3.64, −0.19
	Female	3.94	1.06	1.84, 6.04
Data variability	Low	1.62	1.02	−0.40, 3.64
	High	0.41	0.91	−1.39, 2.20

The results of the ANOVA yielded two significant main effects and four significant interaction effects. All results are displayed in Table 5.

Table 5: Results of the Analysis of Variance

Effect	*F* ratio	*df*	*p*	η^2
Student gender	117.41	1, 130	< .001	.48
Variability	6.25	1, 130	.014	.05
Trend line	0.84	1, 130	.360	.01
Participant gender	1.33	1, 130	.251	.01
Student gender × trend line	31.84	1, 130	< .001	.20
Student gender × variability	70.00	1, 130	< .001	.35
Student gender × participant gender	0.71	1, 130	.401	.01
Variability × trend line	1.03	1, 130	.313	.01
Variability × participant gender	0.53	1, 130	.470	.00
Trend line × participant gender	0.39	1, 130	.536	.00
Student gender × trend line × participant gender	4.36	1, 130	.039	.03
Student gender × variability × trend line	2.60	1, 130	.110	.02
Student gender × variability × participant gender	5.18	1, 130	.025	.04
Variability × trend line × participant gender	0.57	1, 130	.453	.00
Student gender × variability × trend line × participant gender	0.08	1, 130	.777	.00

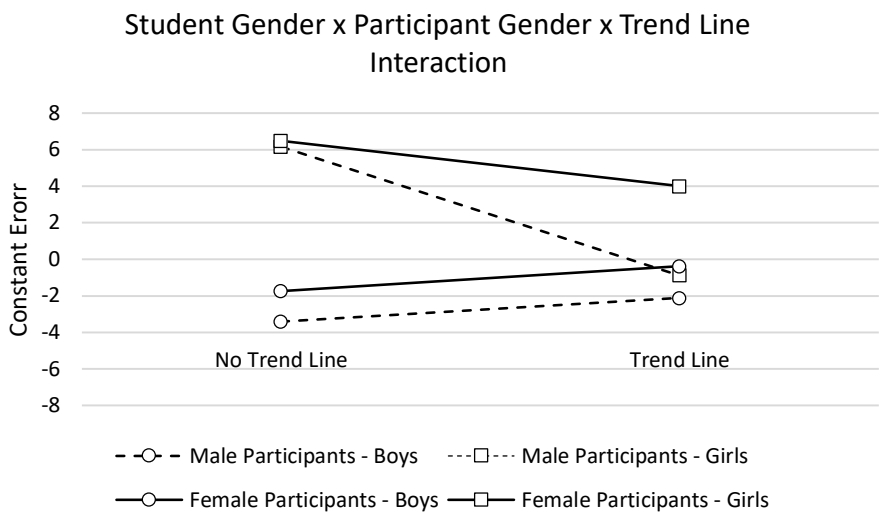
First, there was a significant main effect of student gender, $F(1, 130) = 117.41$, $p < .001$, $\eta^2 = .48$. Participants overestimated on average future reading fluency scores for female students ($M = 3.94$, $SD = 12.27$) compared to male students ($M = -1.91$, $SD = 10.07$), $d = 0.52$. Moreover, there was a significant main effect of data variability, $F(1, 130) = 6.25$, $p = .014$, $\eta^2 = .05$. When data variability was low, constant error was on average larger ($M = 1.62$, $SD = 11.81$) than with high data variability ($M = 0.41$, $SD = 10.53$), $d = 0.11$.

In addition to these main effects, the ANOVA revealed a significant student gender × trend line interaction, $F(1, 130) = 31.84$, $p < .001$, $\eta^2 = .20$. The difference in constant error between boys and girls was dependent on the presence or absence of a trend line. Simple effects test showed that when a trend line was present, the difference between boys and girls was small, $M_{\text{Diff}} = 2.80$, $d = 0.33$, yet significant, $F(1, 130) = 14.70$, $p < .001$, $h^2 = .10$, but when no trend line was shown, the difference was quite large, $M_{\text{Diff}} = 8.90$, $d = 0.97$, $F(1, 130) = 125.38$, $p < .001$, $\eta^2 = .49$. Note, that p values were already adjusted according to Bonferroni.

There was also a significant student gender × data variability interaction, $F(1, 130) = 70.00$, $p < .001$, $\eta^2 = .35$. This interaction effect means that the difference in constant error between boys and girls was affected by data variability. When data variability was high, the difference between boys and girls was larger, $M_{\text{Diff}} = 10.40$, $d = 0.98$, $F(1, 130) = 128.51$, $p < .001$, $\eta^2 = .50$, than when data variability was low, $M_{\text{Diff}} = 1.30$, $d = 0.19$, $F(1, 130) = 5.06$, $p = .026$, $\eta^2 = .35$. However, there was no significant student gender × participant gender interaction, $F(1, 130) = 0.71$, $p = .401$, $\eta^2 = .01$.

In addition to these two-way interactions, the ANOVA yielded two significant three-way interactions, which are displayed in Figures 2 and 3. The first is the student gender × participant gender × trend line interaction, $F(1, 130) = 4.36$, $p = .039$, $\eta^2 = .03$. This interaction means that the two-way student gender × trend line interaction was dependent on the participant gender. As can be seen in Figure 2, the presence of a trend line attenuated the gender bias to a larger degree, when the participants were male rather than female. Simple interaction tests revealed that when participants were female, the student gender × trend line interaction was absent, $F(1, 75) = 0.03$, $p = .865$, $\eta^2 = .00$. However, with male participants, the student gender × trend line interaction was present, but not significant, $F(1, 55) = 3.00$, $p = .089$, $\eta^2 = .05$. To get more insight into the interaction, we conducted simple effects tests. With male participants, the difference between boys and girls was still significant when a trend line was present, $F(1, 36) = 7.53$, $p = .009$, $\eta^2 = .17$, yet without a trend line, the effect was larger, $F(1, 36) = 40.50$, $p < .001$, $\eta^2 = .68$.

Figure 2: Illustration of the Student Gender × Participant Gender × Trend Line Interaction.

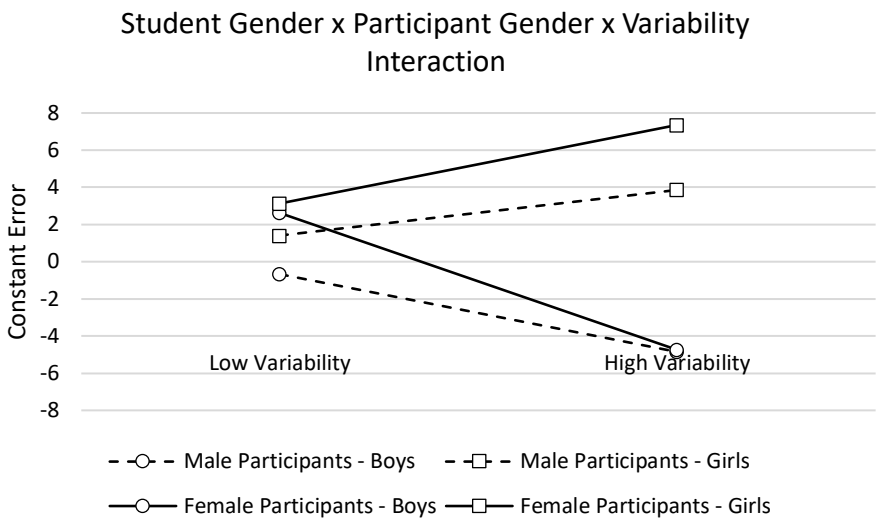


Note. The y-axis depicts the constant error, the x-axis depicts whether a trend line was present or absent.

The second three-way interaction is the student gender × participant gender × data variability interaction, $F(1, 130) = 5.18$, $p = .025$, $\eta^2 = .04$. According to Figure 3, this interaction revealed that the effect of data variability on reducing differences in constant error between girls and boys was larger for female than for male participants. Simple interaction tests showed that when participants were female, the student gender × variability interaction was significant, $F(1, 75) = 48.05$, $p < .001$, $\eta^2 = .39$. With male participants, the student gender × variance interaction was also significant, $F(1, 75) = 37.13$, $p < .001$, $\eta^2 = .40$. Simple effects tests addition-

ally shed more light on the interactions and confirmed the visual impression. With male participants, the difference between boys and girls was significant when data variability was low, $F(1, 55) = 8.35$, $p = .006$, $\eta^2 = .13$, and high, $F(1, 55) = 88.26$, $p < .001$, $\eta^2 = .62$. With female participants, however, the difference between boys and girls was only significant when data variability was high, $F(1, 75) = 74.06$, $p < .001$, $\eta^2 = .50$.

Figure 3: Illustration of the Student Gender × Participant Gender × Variability Interaction.



Note. The y-axis depicts the constant error, the x-axis depicts whether data variability was low or high.

In order to assess whether the participants' teacher education program affected the predictions of the participants, we additionally conducted correlational analyses between the teacher education program (primary vs. secondary education) and the constant error of predictions obtained from all realized combinations of variables that were used to construct the student vignettes. We obtained positive, but weak and insignificant associations, with all rs < .18 and all ps > .154.

5. Discussion

This study evidences that within LPM preservice teachers were prone to a gender bias. When presented with a graph depicting the development of reading fluency of both boys and girls over a period of 11 weeks, the predictions of reading fluency in Week 17 were clearly higher for girls than for boys. In particular, preservice teachers predicted on average a higher score for girls compared to boys, which was 0.52 times the standard deviation of the distribution of constant errors obtained in this study. Therefore, we could find support for our hypothesis that preservice teachers

were stereotyping boys and girls when making judgments about their reading fluency development.

Moreover, as predicted, the gender bias was attenuated when a trend line was presented or when data variability was low. We assumed that when we would reduce ambiguity of the data and hence increased preservice teachers' subjective certainty about the trend inherent in the data, the participants would be less prone to use stereotypical beliefs about reading skills of boys and girls for making judgments. Both the student gender × trend line interaction and the student gender × data variability interaction supported our hypothesis. Notably, the sizes of the interaction effects were quite large ($\eta^2 = .20$, and $\eta^2 = .35$, respectively), which indicates that the reduction of the gender bias was quite effective.

Furthermore, we could not find a same-gender bias in male participants in our data. Both male and female participants favored girls over boys with respect to predicted reading fluency scores. However, the participants' preference for girls reduced when ambiguity of the data was decreased or when a trend line was present. These effects, albeit small, were affected by the participants' gender. Male participants benefitted more from a trend line than female participants, as males made similar predictions for both boys and girls and with only small constant errors. When data ambiguity was lowered by reducing data variability, female participants seemed to benefit more than males, since low data variability let the gender bias almost completely disappear in female participants.

Why did a gender bias in LPM occur, why was it so strong, and why did different ways of decreasing ambiguity result in different reductions of gender bias between male and female participants? In LPM, teachers or preservice teachers use the graphical depiction of a student's achievement development to make several decisions. They could use the achievement trajectory to judge whether lessons had been effective, whether a student needs additional support, or whether their instruction has to be adjusted in order to better fit the needs of the students. Whatever the decisions to be made are, teachers have to understand the data depicted as a kind of a learning curve. Several studies (e.g., Espin et al., 2017; Van den Bosch et al., 2017; Van Norman et al., 2013) have called into question that teachers have adequate skills to fully understand what is presented in achievement trajectories, especially when relations between parameters of the curve have to be determined (e.g., how well students achieved their achievement goals) or when conclusions have to be drawn with respect to educational interventions. Since most of the learning progress data are variable (Van Norman & Christ, 2016), teachers necessarily would encounter difficulties in detecting the trend of the data, which in turn would make it difficult for them to both "read between the data" and "read beyond the data" (Friel et al., 2001).

Data ambiguity might encourage teachers to look for additional information that would facilitate their judgements. However, when there is no source of additional information, or when decisions have to be reached quickly, stereotypical beliefs are likely to guide expectations about those students (Kunda & Sherman-Williams, 1993). This might be even more the case if teachers have to predict stu-

dents' achievements based on what they see on a chart showing a student's learning progress. Given current knowledge, predictions are always uncertain since there are multiple possible future states of nature (Stewart, 2000). If uncertainty cannot be reduced by the integration of additional information, heuristic and category-based processing of information is likely to occur (Fiske & Neuberg, 1990). Moreover, when teachers (or preservice teachers) are to make a prediction about a student's future achievement, they are likely to use expectations they have about that student to guide their predictions. Compared to judgments of current states, predictions of future states might be even more susceptible for stereotypical expectation, like, for instance, the expectation that on average girls would show better achievements than boys.

The amount of gender bias obtained in this study might arise from two sources, which are the prediction task, and the absence of additional prediction-relevant information about the students. In prediction tasks, uncertainty is usually higher than in tasks where the judgment concerns actual instead of future events, since predictions involve the future, and the future is unknown. In addition, the prediction task in this study can be characterized as being complex, since it required the participants to both differentiate the information presented with regard to the different interpretations they allowed, and to integrate the information to a single judgment (cf. Tetlock & Kim, 1987). Complex tasks would elicit the use of stereotypes more likely than simple tasks would do (Bodenhausen & Lichtenstein, 1987). Therefore, the prediction task itself might have facilitated the use of stereotypical beliefs. Moreover, the participants were forced to make predictions without having access to information relevant for preservice teachers' predictions besides what was presented in the graph. A lack of individuating information, however, corresponds with an increased use of stereotypes (Fiske et al., 1999; Macrae et al., 1993).

Apparently, male and female participants processed data ambiguity and its reduction in different ways. When presented with a trend line, male participants were less prone to a gender bias compared to male participants who received vignettes without a trend line. With female participants, however, the effect of the trend line on gender bias was smaller than for male participants, and they made more accurate predictions when data variability was low. The differences between male and female participants' predictions might be caused by the differences between the realized ways of reducing data ambiguity. When data variability was low rather than high, reduction of ambiguity was produced by the students themselves. That is, students showing constant progress of achievement with only minor fluctuations might be regarded as stable and therefore easy to predict. Predictions based on the trend line, however, could neglect the ups and downs in the growth curve, such that predictions for both stable and unstable students could be made with similar certainty, provided that the trend line depicted the "real" achievement development.

It appears that female participants seemed to be more skeptical against the trend line, whereas male participants did more than female participants rely on external statistical aids.

6. Limitations

People use their stereotypical beliefs about social groups as a basis for judgment and prediction whenever they lack the desire or the ability to engage in more extensive processing of information (Bodenhausen, 1993). Such a lack of desire or ability may be common under most everyday life circumstances (Fiske & Neuberg, 1990), but not always in educational decisions made by teachers. Therefore, it might be a limitation of this study that the participants made predictions about virtual students, not real students. Hence, because of the artificial character of the experiment, the participants might not have felt highly accountable for their predictions (Tetlock & Kim, 1987) and therefore engaged less in effortful information processing than they would do in real school settings.

In addition, teachers in real classrooms have access to a variety of information that could bolster their decisions and predictions about students' achievements. We could therefore expect that predictions of achievements with real students might be less affected by stereotypes and category-based information processing than in this study.

Moreover, the participants were preservice instead of in-service teachers. Although there is ample evidence that both make comparable judgments (Glock & Karbach, 2015; Mertler, 2004), preservice teachers usually lack the experience in-service teachers have when students are to be judged, and may use information about students differently (e.g., Sabers et al., 1991).

Furthermore, effects of the independent variables used in this study could be affected by the fact that the participants were inexperienced with LPM. It is possible that more experienced participants would have made less errors in prediction.

A further limitation might be that we did not establish a baseline condition where the gender of the students was not announced. This baseline condition would have shown how precise predictions are when the participants were unaffected by student gender.

Finally, we did not apply some sort of manipulation check. For instance, we could have checked for whether the participants actually had attended to the student names that were attached to the graphs by asking them at the end of the study. Manipulation checks may help figure out if participants noticed the treatment at all. However, making the manipulation check at the end of the study may compromise its validity, since the participants may not remember what they were feeling or thinking during the study (Hauser et al., 2018).

7. Conclusions

The results indicate that preservice teachers were prone to apply gender stereotypes when interpreting and predicting learning progress data. The likelihood of the occurrence of stereotyping was raised when the ambiguity of progress data was

high. Therefore, it can be concluded that caution is warranted when (preservice) teachers base their interventions on learning progress data that is highly ambiguous and not supported by visual or statistical aids.

Based on the results obtained, we recommend training of graph literacy (Friel et al., 2001) for both preservice and in-service teachers. There are a few studies on training programs for teachers in regard to the interpretation of learning progress data, showing that training of the interpretation of visualized progress data is possible and successful (Kennedy et al., 2015; Van den Bosch et al., 2019). These training programs should be adapted to teacher education programs in Germany.

References

Ardoin, S. P., & Christ, T. J. (2009). Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set. *School Psychology Review, 38*(2), 266–283. https://doi.org/10.1080/02796015.2009.12087837

Ardoin, S. P., Christ, T. J., Morena, L. S., Cormier, D. C., & Klingbeil, D. A. (2013). A systematic review and summarization of the recommendations and research surrounding Curriculum-Based Measurement of oral reading fluency (CBM-R) decision rules. *Journal of School Psychology, 51*(1), 1–18. https://doi.org/10.1016/j.jsp.2012.09.004

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience, 15*(4), 600–609. https://doi.org/10.1162/089892903321662976

Bodenhausen, G. V. (1993). Emotions, arousal, and stereotypic judgements: A heuristic model of affect and stereotyping. In D. M. Mackie & D. L. Hamilton (Eds.), *Affect, cognition, and stereotyping* (pp. 13–37). Academic Press. https://doi.org/10.1016/B978-0-08-088579-7.50006-5

Bodenhausen, G. V., & Lichtenstein, M. (1987). Social stereotypes and information-processing strategies: The impact of task complexity. *Journal of Personality and Social Psychology, 52*(5), 871–880. https://doi.org/10.1037/0022-3514.52.5.871

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*(5), 531–563. https://doi.org/10.1177/0145445503261167

Campbell, D. T. (1967). Stereotypes and the perception of group differences. *American Psychologist, 22*(10), 817–829. https://doi.org/10.1037/h0025079

Chiu, M. M., & McBride-Chang, C. (2006). Gender, context, and reading: A comparison of students in 43 countries. *Scientific Studies of Reading*, *10*(4), 331–362. https://doi.org/10.1207/s1532799xssr1004_1

Christ, T. J., Zopluoglu, C., Long, J. D., & Monaghen, B. D. (2012). Curriculum-based measurement of oral reading: Quality of progress monitoring outcomes. *Exceptional Children, 78*(3), 356–373. https://doi.org/10.1177/001440291207800306

Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of Personality and Social Psychology, 108*(2), 219–233. https://doi.org/10.1037/pspa0000015

Darley, J. M., & Gross, P. G. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44*(1), 20–33. https://doi.org/10.1037/0022-3514.44.1.20

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources, 42*(3), 528–554. https://doi.org/10.3368/jhr.XLII.3.528

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219–232. https://doi.org/10.1177/001440298505200303

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*(3), 184–192. https://doi.org/10.1177/00224669030370030801

Deno, S. L. (2013). Problem-solving assessment. In R. Brown-Chidsey & K. J. Andren (Eds.), *Assessment for intervention: A problem-solving approach* (pp. 10–36). Guilford.

Driessen, G. (2007). The feminization of primary education: Effects of teachers' sex on pupil achievement, attitudes and behaviour. *Review of Education, 53*(2), 183–203. https://doi.org/10.1007/s11159-007-9039-y

Ehrenberg, R. G., Goldhaber, D. D., & Brewer, D. J. (1995). Do teachers' race, gender, and ethnicity matter? Evidence from the National Educational Longitudinal Study of 1988. *Industrial and Labor Relations Review, 48*(3), 547–561. https://doi.org/10.1177/001979399504800312

Espin, C. A., Waymann, M. M., Deno, S. L., & McMaster, K. L. (2017). Data-based decision making: Developing a method for capturing teachers' understanding of CBM graphs. *Learning Disabilities Research & Practice, 32*(1), 8–21. https://doi.org/10.1111/ldrp.12123

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fiske, S. T., Lin, M., & Neuberg, S. (1999). The continuum model. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 321–254). Guilford.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology, 23*, 1–74. https://doi.org/10.1016/S0065-2601(08)60317-2

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal of Research in Mathematics Education, 32*(2), 124–158. https://doi.org/10.2307/749671

Fuchs, L. S., Fuchs, D., & Hosp, M. K. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239–256. https://doi.org/10.1207/S1532799XSSR0503_3

Geller, G., Faden, R. R., & Levine, D. M. (1990). Tolerance for ambiguity among medical students: Implications for their selection, training and practice. *Social Science & Medicine, 31*(5), 619–624. https://doi.org/10.1016/0277-9536(90)90098-D

Gerhards, J. (2010). *Die Moderne und ihre Vornamen: Eine Einladung in die Kultursoziologie* [Modernity and its first names: An invitation to cultural sociology] (2nd ed.). VS Verlag für Sozialwissenschaften.

Glock, S., & Karbach, J. (2015). Preservice teachers' implicit attitudes toward racial minority students: Evidence from three implicit measures. *Studies in Educational Evaluation, 45,* 55–61. https://doi.org/10.1016/j.stueduc.2015.03.006

Glock, S., & Kleen, H. (2017). Gender and student misbehavior: Evidence from implicit and explicit measures. *Teaching and Teacher Education, 67*, 93–103. https://doi.org/10.1016/j.tate.2017.05.015

Good, R. H., & Shinn, M. R. (1990). Forecasting accuracy of slope estimates for reading curriculum-based measurement: Empirical evidence. *Behavioral Assessment, 12*(2), 179–193.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology, 9*, Article 998. https://doi.org/10.3389/fpsyg.2018.00998

Hicklin, S. K., & Wedell, D. H. (2005). Learning group differences: implications for contrast and assimilation in stereotyping. *Social Cognition, 25*(3), 410–454. https://doi.org/10.1521/soco.2007.25.3.410

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology, 47*, 237–271. https://doi.org/10.1146/annurev.psych.47.1.237

Holmlund, H., & Sund, K. (2008). Is the gender gap in school performance affected by the sex of the teacher? *Labour Economics, 15*(1), 37–53. https://doi.org/10.1016/j.labeco.2006.12.002

Horner, R. H., & Odom, S. L. (2014). Constructing single-case research designs: Logic and options. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methods and statistical advances* (pp. 27–52). American Psychological Association. https://doi.org/10.1037/14376-002

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2007). *The ABCs of CBM. A practical guide to curriculum-based measurement*. Guilford.

Jussim, L., & Eccles, J. (1992). Teacher expectations: II. Construction and reflection of student achievement. *Journal of Personality and Social Psychology, 63*(6), 947–961. https://doi.org/10.1037/0022-3514.63.6.947

Kennedy, M. J., Wagner, D., Stegall, J., Lembke, E., Miciak, J., Alves, K. D., Brown, T., Driver, M. K., & Hirsch, S. E. (2015). Using content acquisition podcasts to improve teacher candidate knowledge of curriculum-based measurement. *Exceptional Children, 82*(3), 303–320. https://doi.org/10.1177/0014402915615885

Klapproth, F. (2006). Mental models of growth. In H. Helfrich, M. Zillekens, & E. Hölter (Eds.), *Culture and development in Japan and Germany* (pp. 141–153). Daedalus.

Klapproth, F. (2018). Biased predictions of students' future achievement: An experimental study on preservice teachers' interpretation of curriculum-based measurement graphs. *Studies in Educational Evaluation*, *59*, 67–75. https://doi.org/10.1016/j.stueduc.2018.03.004

Klapproth, F., & Fischer, B. D. (2019). Preservice teachers' evaluations of students' achievement development in the context of school-track recommendations. *European Journal of Psychology of Education, 34*(4), 825–846. https://doi.org/10.1007/s10212-018-0405-x

Klein, O., Doyen, S., Leys, C., Magalhães de Saldanha da Gama, P. A., Miller, S., Questienne, L., & Cleeremans, A. (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioral experiments. *Perspectives on Psychological Science, 7*(6), 572–584. https://doi.org/10.1177/1745691612463704

Kossak, S. N., & Johnson, M. M. (2001). The effects of a sensitization technique on stereotyping behavior. *Social Work Education, 20*(2), 199–207. https://doi.org/10.1080/02615470120044293

Kranzler, J. H., Miller, M. D., & Jordan, L. (1999). An examination of racial/ethnic and gender bias on curriculum-based measurement of reading. *School Psychology Quarterly, 14*(3), 327–342. https://doi.org/10.1037/h0089012

Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the construal of individuating information. *Personality and Social Psychology Bulletin, 19*(1), 90–99. https://doi.org/10.1177/0146167293191010

Lorenz, G., Gentrup, S., Kristen, C., Stanat, P., & Kogan, I. (2016). Stereotype bei Lehrkräften? Eine Untersuchung systematisch verzerrter Lehrererwartungen [Stereotypes among teachers? A study of systematic bias in teacher expectations].

Kölner Zeitschrift für Soziologie und Sozialpsychologie, 68(1), 89–111. https://doi.org/10.1007/s11577-015-0352-3

Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology, 23*(1), 77–87. https://doi.org/10.1002/ejsp.2420230107

Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology, 66*(1), 37–47. https://doi.org/10.1037/0022-3514.66.1.37

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and applications* (3rd ed.). Wiley. http://hdl.handle.net/11728/7113

Martin, A., & Marsh, H. (2005). Motivating boys and motivating girls: Does teacher gender really make a difference? *Australian Journal of Education, 49*(3), 320–334. https://doi.org/10.1177/000494410504900308

McElvany, N., Kessels, U., Schwabe, F., & Kasper, D. (2017). Geschlecht und Lesekompetenz [Gender and reading competence]. In A. Hußmann, H. Wendt, W. Bos, A. Bremerich-Vos, D. Kasper, E.-M. Lankes, N. McElvany, T. C. Stubbe, & R. Valentin (Eds.), *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 177–194). Waxmann. https://doi.org/10.25656/01:15476

Mertler, C. A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education, 33*(1), 49–64. https://www.jstor.org/stable/41064623

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016: International results in reading*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College International Association for the Evaluation of Educational Achievement (IEA).

Muntoni, F., & Retelsdorf, J. (2018). Gender-specific teacher expectations in reading – The role of teachers' gender stereotypes. *Contemporary Educational Psychology, 54*, 212–220. https://doi.org/10.1016/j.cedpsych.2018.06.012

Nelson, P. M., Van Norman, E. R., & Christ, T. J. (2017). Visual analysis among novices: Training and trend lines as graphic aids. *Contemporary School Psychology, 21*(2), 93–102. https://doi.org/10.1007/s40688-016-0107-9

Neugebauer, M., Helbig, M., & Landmann, A. (2011). Unmasking the myth of the same-sex teacher advantage. *European Sociological Review, 27*(5), 669–689. https://doi.org/10.1093/esr/jcq038

Plante, I., de la Sablonnière, R., Aronson, J. M., & Théorêt, M. (2013). Gender stereotype endorsement and achievement-related outcomes: The role of competence beliefs and task values. *Contemporary Educational Psychology, 38*(3), 225–235. https://doi.org/10.1016/j.cedpsych.2013.03.004

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement of oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*(6), 427–469. https://doi.org/10.1016/j.jsp.2009.07.001

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science, 24*(7), 1301–1308. https://doi.org/10.1177/0956797612466268

Sabers, D. S., Cushing, K. S., & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensional, and immediacy. *American Education Research Journal, 28*(1), 63–88. https://doi.org/10.3102/00028312028001063

Schneider, D. J. (2004). *The psychology of stereotyping*. Guilford.

Shinn, M. R., Good, R. H., & Stein, S. (1989). Summarizing trend in student achievement: A comparison of methods. *School Psychology Review, 18*(3), 356–370. https://doi.org/10.1080/02796015.1989.12085432

Spieß, L., & Bekkering, H. (2020). Predicting choice behavior of group members. *Frontiers in Psychology*, *11*, Article 508. https://doi.org/10.3389/fpsyg.2020.00508

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. https://doi.org/10.1002/pits.20113

Stewart, T. R. (2000). Uncertainty, judgement, and error in prediction. In D. Sarewitz, R. A. Pielke, & R. Byerly (Eds.), *Prediction: Science, decision making, and the future of nature* (pp. 41–57). Island.

Strathmann, A., Klauer, K. J., & Greisbach, M. (2010). Lernverlaufsdiagnostik – dargestellt am Beispiel der Entwicklung der Rechtschreibkompetenz in der Grundschule [Curriculum based measurement – demonstrated through the development of writing competence in primary school]. *Empirische Sonderpädagogik, 2*(1), 64–77. https://doi.org/10.25656/01:9338

Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgement processes in a personality prediction task. *Journal of Personality and Social Psychology, 52*(4), 700–709. https://doi.org/10.1037/0022-3514.52.4.700

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education,* Article 958530. https://doi.org/10.1155/2013/958530

Tindal, G., Deno, S., & Ysseldyke, J. (1983). *Visual analysis of time series data: Factors of influence and level reliability* (Research Report No. 112). University of Minnesota, Institute for Research on Learning Disabilities (IRLD).

Valutis, S. A. (2015). The relationship between tolerance of ambiguity and stereotyping: Implications for BSW education. *Journal of Teaching in Social Work, 35*(5), 513–528. https://doi.org/10.1080/08841233.2015.1088927

Van den Bosch, R. M., Espin, C. A., Chung, S., & Saab, N. (2017). Data-based decision making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learning Disabilities Research & Practice, 32*(1), 46–60. https://doi.org/10.1111/ldrp.12122

Van den Bosch, R. M., Espin, C. A., Pat-El, R. J., & Saab, N. (2019). Improving teachers' comprehension of curriculum-based measurement progress monitoring graphs. *Journal of Learning Disabilities, 52*(5), 413–427. https://doi.org/10.1177/0022219419856013

Van Knippenberg, A., Dijksterhuis, A., & Vermeulen, D. (1999). Judgement and memory of a criminal act: The effects of stereotypes and cognitive load. *European Journal of Social Psychology, 29*(2–3), 191–201. https://doi.org/10.1002/(SICI)1099-0992(199903/05)29:2/3<191::AID-EJSP923>3.0.CO;2-O

Van Norman, E. R., & Christ, T. J. (2016). How accurate are interpretations of curriculum-based measurement progress monitoring data? Visual analysis versus decision rules. *Journal of School Psychology, 58*, 41–55. https://doi.org/10.1016/j.jsp.2016.07.003

Van Norman, E. R., Nelson, P. M., Shin, J.-E., & Christ, T. J. (2013). An evaluation of the effects of graphic aids in improving decision accuracy in a continuous treatment design. *Journal of Behavioral Education, 22*(4), 283–301. https://doi.org/10.1007/s10864-013-9176-2

Van Norman, E. R., & Parker, D. C. (2018). A comparison of common and novel curriculum-based measurement of reading decision rules to predict spring performance for students receiving supplemental interventions. *Assessment of Effective Intervention, 43*(2), 110–120. https://doi.org/10.1177/1534508417728695

Yeo, S., Fearrington, J., & Christ, T. J. (2011). An investigation of gender, income, and special education status bias on curriculum-based measurement slope in reading. *School Psychology Quarterly, 26*(2), 119–130. https://doi.org/10.1037/a0023021

Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01.03
Volume 14 (2022), No. 1, 39–60
CC BY-NC-SA 4.0 Waxmann 2022

Sterett H. Mercer & Joanna E. Cannon

Validity of Automated Learning Progress Assessment in English Written Expression for Students with Learning Difficulties

Abstract

We evaluated the validity of an automated approach to learning progress assessment (aLPA) for English written expression. Participants (n = 105) were students in Grades 2–12 who had parent-identified learning difficulties and received academic tutoring through a community-based organization. Participants completed narrative writing samples in the fall and spring of 1 academic year, and some participants (n = 33) also completed a standardized writing assessment in the spring of the academic year. The narrative writing samples were evaluated using aLPA, four hand-scored written expression curriculum-based measures (WE-CBM), and ratings of writing quality. Results indicated (a) aLPA and WE-CBM scores were highly correlated with ratings of writing quality; (b) aLPA and more complex WE-CBM scores demonstrated acceptable correlations with the standardized writing subtest assessing spelling and grammar, but not the subtest assessing substantive quality; and (c) aLPA scores showed small, statistically significant improvements from fall to spring. These findings provide preliminary evidence that aLPA can be used to efficiently score narrative writing samples for progress monitoring, with some evidence that the aLPA scores can serve as a general indicator of writing skill. The use of automated scoring in aLPA, with performance comparable to WE-CBM hand scoring, may improve scoring feasibility and increase the likelihood that educators implement aLPA for decision-making.

Keywords

automated text evaluation, learning progress assessment, written expression, curriculum-based measurement, learning difficulties

Prof. Dr. Sterett H. Mercer (corresponding author), ORCID: 0000-0002-7940-4221 · Prof. Dr. Joanna E. Cannon, ORCID: 0000-0003-4552-4476, Department of Educational and Counselling Psychology and Special Education, University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada
email: sterett.mercer@ubc.ca
joanna.cannon@ubc.ca

Sterett H. Mercer & Joanna E. Cannon

Validität des automatisierten Learning Progress Assessments im geschriebenen Englisch für Schüler:innen mit Lernschwierigkeiten

Zusammenfassung

Wir evaluierten die Validität eines automatischen Ansatzes des Learning Progress Assessments (aLPA) in geschriebener Sprache. Schüler:innen der Klassen 2 bis 12 (n = 105) mit Lernschwierigkeiten, die deren Eltern festgestellt hatten, nahmen an Nachhilfeunterricht, welcher von einer gemeinnützigen Organisation durchgeführt wurde, teil. Die Schüler:innen erstellten im Herbst und Frühjahr eines Schuljahres Schreibproben. Weiterhin nahmen einige Schüler:innen (n = 33) an einer standardisierten Schreibprüfung im Frühjahr teil. Die Schreibproben wurden mit aLPA, vier handkodierten Curriculum-Based Measures (WE-CBM) und hinsichtlich deren Schreibqualität ausgewertet. Unsere Ergebnisse zeigen, (a) aLPA- und WE-CBM-Werte korrelierten hoch mit der Bewertung der Schreibqualität, (b) aLPA- und die komplexeren WE-CBM-Werte zeigten akzeptable Korrelationen mit der standardisierten Rechtschreib- und Grammatikprüfung, jedoch nicht mit tatsächlicher Qualität, und (c) aLPA-Werte zeigten geringe, statistisch signifikante Verbesserungen von Herbst zu Frühjahr. Diese Ergebnisse deuten darauf hin, dass aLPA als effiziente Methode zur Bewertung von Schreibproben verwendet werden kann und der aLPA-Wert als allgemeiner Schreibkompetenzindikator dienen kann. Automatische Bewertung im aLPA, dessen Validität mit der von WE-CBM vergleichbar ist, kann die Bewertung vereinfachen und damit ist es wahrscheinlicher, dass Lehrkräfte aLPA verwenden.

Schlagworte

Lernverlaufsdiagnostik, automatisiertes Learning Progress Assessment, geschriebenes Englisch, Schüler:innen mit Lernschwierigkeiten

1. Introduction

Many students with academic difficulties do not respond to generally effective academic interventions (McMaster et al., 2005); this failure to respond is problematic because educators and related professionals are responsible for improving the academic skills of individual students. To address this concern, data-based individualization is gaining recognition as an approach to special education service delivery (DBI; Fuchs et al., 2013). In DBI, generally effective interventions, as determined by the research literature, are first implemented with ongoing progress monitoring of student academic outcomes. When progress monitoring data indicate that the intervention is not working for a specific student, educators modify the intervention approach until monitoring data indicate improvement. Although DBI is commonly implemented and researched for reading interventions (Filder-

man et al., 2018), DBI in written expression is less commonly implemented (Jung et al., 2018), in part due to challenges related to progress monitoring assessment. The purpose of this study is to present a validity argument and preliminary validity evidence for automated learning progress assessment (aLPA) in written expression.

1.1 Curriculum-Based Measurement

In DBI, progress monitoring assessment is typically done using curriculum-based measurement (CBM; Deno, 1985). Several characteristics are emphasized in CBM: (a) simplicity and efficiency of administration, scoring, and interpretation; (b) general outcome measurement, meaning that these simple measures should serve as indicators of overall performance in an academic domain (e.g., correlate with more comprehensive assessments of academic skill); (c) cost-effectiveness so that alternate forms can be frequently administered; and (d) reliability and validity of scores to inform data-based instructional decisions. Although CBM originally used curriculum-specific materials in assessment, the emphasis on reliability and validity led to the development of standard CBM tasks and materials that are aligned to the general skills taught in the curriculum (Hosp et al., 2016). Oral passage reading (OPR) is a commonly-used CBM that consists of a 1-minute timed reading of a standardized, field-tested, grade-level passage that is scored for the number of words read correctly. Despite the brevity and simplicity, OPR scores correlate highly with more comprehensive standardized assessments of reading (Reschly et al., 2009) and are sensitive to differential skill growth during intervention (Morgan & Sideridis, 2006).

For written expression CBM (WE-CBM), typical procedures involve presenting students with a short story starter (e.g., “One day on the way to school, I ...”), allowing them to plan for 1 minute, and then having them write for 3 minutes (Hosp et al., 2016). These writing samples are then hand-scored with simple metrics such as counts of the total number of words written and the number of words spelled correctly or with more complex metrics such as the number of correct word sequences (the number of grammatically and semantically acceptable adjacent words that are spelled and punctuated correctly; Videen et al., 1982). Although there is some research to support the use of these WE-CBM procedures as part of DBI to improve elementary students’ early writing skills (McMaster et al., 2020), there are several challenges to be addressed at higher grade levels, including the need for multiple, longer duration writing samples for adequate reliability (Keller-Margulis et al., 2016; Kim et al., 2017) and the requirement to use more complex scoring metrics for adequate validity (Mercer et al., 2012; Romig et al., 2017). Complex WE-CBM scoring of longer duration writing samples may limit feasibility of implementation by educators (Espin et al., 1999), which could be a barrier to more widespread implementation of DBI for written expression beyond the early elementary grades.

1.2 Automated Learning Progress Assessment in Written Expression

To address these concerns, we describe an automated approach to learning progress assessment as an alternative to traditional hand-scored WE-CBM. We describe our approach using the more general term LPA, consistent with Förster and Souvignier's (2014, 2015) LPA in reading, for several reasons. First, we believe that scoring methods for written expression, although summarized in an overall writing skill score, should be based on a comprehensive set of writing quality indices (an extended test concept; Förster & Souvignier, 2015). In WE-CBM, scoring primarily considers text production and accuracy, whereas in aLPA, we use a wide range of word-, sentence-, and discourse-level indices provided by automated text evaluation software to generate overall writing quality scores. Second, based on research demonstrating that automated text evaluation can generate writing quality scores that are useful for screening (Keller-Margulis et al., 2021; Mercer et al., 2019; Wilson, 2018), we also anticipate that computer-based assessment will be necessary for writing samples to be scored and for such a system to be feasibly used by teachers. Third, given that we know that multiple, longer-duration writing samples will be necessary for reliability (Keller-Margulis et al., 2016), we anticipate that a reduced test frequency will be optimal, compared to typical CBM progress monitoring procedures of weekly assessments. We differ from Förster and Souvignier (2014, 2015) in our primary intended uses for aLPA, specifically, screening and progress monitoring of students with learning difficulties instead of monitoring all students, but concur that such a system should also be useful for monitoring the learning progress of most students.

1.2.1 Interpretation and Use of aLPA

Consistent with contemporary validation approaches (American Educational Research Association et al., 2014; Kane, 2013), we specify our proposed interpretation and use of aLPA scores, and also continue to contrast aLPA with WE-CBM. Specifically, we believe that aLPA scores should assess the overall writing quality of student writing samples generated under authentic instructional conditions. By authentic, we mean, at minimum, that students should have adequate time to engage in the writing process (i.e., beyond the 3-minute time limit often used in WE-CBM), and ideally the writing samples would be generated from tasks embedded in the curriculum, "where purpose and audience are clear and meaningful, where support and feedback are readily available, and where the final product has academic value for the student" (Calfee & Miller, 2007, p. 269).

1.2.1.1 Scoring Inference

In stating that aLPA scores should reflect writing quality on the evaluated writing samples, we are articulating an explicit, testable *scoring inference* (Kane, 2013) that is informed by developmental writing theory. The not-so-simple view of writing (Berninger & Amtmann, 2003) considers writing to involve three processes that are constrained by working memory capacity: (a) text generation, which involves idea generation and translation of these ideas into language; (b) transcription skills, which is the translation of these language representations into written text by handwriting or typing; and (c) strategic self-regulation of the writing process (e.g., planning, drafting, revising). Difficulties in text generation or transcription skills can contribute to poor writing fluency, and this limited automaticity in generating connected text can impair writing quality (Kim et al., 2018). For developing writers, writing quality is often defined as the extent to which ideas are developed and organized in compositions (Kim et al., 2015, 2018).

Due to short time allowed to write, WE-CBM scores are typically interpreted as assessing the construct of writing fluency more so than writing quality (Kim et al., 2018; Ritchey et al., 2016), and the extent to which WE-CBM scores assess quality on the writing samples used to generate the scores has received minimal attention in research (McMaster & Espin, 2007). By contrast, in aLPA, students have a more extended time to write, thereby increasing students' ability to engage in the processes described in the not-so-simple view of writing. Also, we explicitly test the scoring inference that automated aLPA scores assess raters' judgements of quality on the evaluated writing samples.

1.2.1.2 Generalization Inference

In addition to assessing quality on the administered writing samples, aLPA scores should also correlate with quality on writing samples administered under similar conditions; this claim is a *generalization inference* (Kane, 2013). Because a generalizability theory study of WE-CBM scores demonstrated that reliability improves as writing sample duration increases (durations of 1–7 minutes were investigated; Keller-Margulis et al., 2016), we anticipate that the longer administration time in aLPA will support generalization. That said, a longer duration may not be sufficient given that generalizability theory studies have found that multiple writing samples are needed for adequate reliability and there may be limited generalization across writing samples of different genres (Kim et al., 2017; Wilson et al., 2019).

1.2.1.3 Extrapolation Inference

Similar to WE-CBM, we expect that aLPA scores will serve as general indicators of writing skill, which is an *extrapolation inference* (Kane, 2013). This inference is

testable by determining the extent to which aLPA scores correlate with scores on other standardized writing assessments that have evidence of validity. In CBM and aLPA, scores are used for screening and progress monitoring; because these decisions involve the assessment of large groups of students (screening) or frequent assessment of skills (progress monitoring), efficiency of administration and scoring are important (Deno, 1985). For this reason, evidence of the ability to extrapolate is critical – the goal in CBM and aLPA is to identify the most efficient test format that can maintain strong correlations with more comprehensive assessments of academic skills.

1.2.1.4 Decision Inference

Several sources of evidence can support the validity of aLPA for screening and progress monitoring, which are *decision inferences* (Kane, 2013). For example, screening instruments should have evidence that scores can be used to accurately predict meaningful outcomes such as placements of students in special education programs (e.g., Fewster & Macmillan, 2002) and whether students meet proficiency standards on high-stakes assessments (e.g., Furey et al., 2016). Progress monitoring instruments should have evidence that scores improve in response to writing intervention (e.g., McMaster et al., 2017). Ultimately, a set of decision rules would need to be articulated and evaluated to use aLPA for screening and progress monitoring, which would require the establishment of aLPA norms and performance standards.

1.3 Current Study

As an early-stage investigation of the validity of aLPA for progress monitoring in written expression, we compare the performance of aLPA vs. WE-CBM metrics in relation to some of the previously detailed assumptions related to scoring, extrapolation, and decision inferences. Analyses are based on writing samples from students with parent-reported learning difficulties receiving academic tutoring in a community agency for 1 academic year. Specifically, we address the following research questions:

1. Scoring inference: How strongly do aLPA quality scores and WE-CBM metrics correlate with ratings of writing quality on the scored writing samples?
2. Extrapolation inference: How strongly do aLPA quality scores and WE-CBM metrics correlate with scores on a standardized writing assessment?
3. Decision inference: Are aLPA quality scores sensitive to student skill growth?

2. Method

2.1 Participants and Setting

All participants were enrolled in a community-based nonprofit organization that provides low-cost, after-school, one-to-one academic tutoring for approximately 2 hours per week in reading, math, and/or written expression to students with suspected or diagnosed learning disabilities. As part of typical service delivery, the organization collected picture-prompted narrative writing samples from students in the fall (September–October) and spring (April–May) of an academic year to inform instruction, and some students also completed a standardized writing assessment in the spring. Participants included 105 students in Grades 2–12 who completed at least one of the narrative writing samples; 79 participants completed both the fall and spring writing samples. Of the participants, 103 completed the fall writing sample, 83 completed the spring writing sample, and 33 completed the standardized writing assessment. The participants were in the following grade levels in school: Grade 2 (n = 11), 3 (n = 20), 4 (n = 14), 5 (n = 20), 6 (n = 13), 7 (n = 8), 8 (n = 6), 9 (n = 7), 10 (n = 2), 11 (n = 2), and 12 (n = 2). Forty percent of the participants were female. The 33 participants with standardized writing assessment data were in Grades 3–9; 39% were female.

Because we only had access to the writing samples and assessment data, we do not have detailed demographic information for participants or the specific learning difficulties of students; however, all were experiencing academic difficulties substantial enough for parents to seek community-based tutoring beyond the school-based supports available. Formal diagnoses of learning disabilities are not required by the organization due to long wait lists for assessments in the local public schools and high costs for assessments in private clinics (Werb, 2007). Approximately 50% of students served by the organization have a formal diagnosis of a learning disability, and approximately 60% of students receive financial support to access the organization's services.

Most participants attended school in a large, urban, ethnically diverse school district of approximately 52 000 students in Western Canada. In the district, 44% of students speak a language other than English at home, with 160 different languages spoken by families in the district. The top five languages other than English spoken at home are the following: Cantonese (17% of students), Mandarin (11%), Tagalog (5%), Vietnamese (4%), and Punjabi (4%). Approximately 17% of students in the district are eligible for English language supports, and 11% of students are eligible for special education services.

2.2 Measures

Measures included ratings of writing quality, aLPA quality, and multiple WE-CBM metrics based on narrative writing samples and selected subtests from a standardized writing assessment.

2.2.1 Narrative Writing Samples

To inform instruction, the non-profit agency's tutors asked students to select one picture about which they would like to write from a collection of photos from travel, recreation, and lifestyle magazines (e.g., pictures of amusement park rides, animals, restaurants). The tutors informed students that they would have 10 minutes to handwrite, and the tutors provided no assistance to students during this time period. As part of their typical planning and assessment model, tutors recorded on a checklist whether the students' writing exhibited specific characteristics consistent with grade-level expectations, for example, subject-verb agreement, capitalization, punctuation, and organization; because the current study focused on overall writing quality, we did not use the checklist data in the current study. Before scoring by hand (rated writing quality and WE-CBM) and computer (aLPA), a research team member typed all writing samples, preserving errors in spelling and grammar, and a second research team member verified the accuracy of all transcriptions.

2.2.1.1 Rated Writing Quality

We hand-evaluated writing quality for the picture-prompted writing samples using the method of paired comparisons (Thurstone, 1927). Specifically, two raters each completed 3000 comparisons of pairs of the 186 writing samples (103 from the beginning and 83 from the end of the academic year). In these comparisons, raters identified the writing sample in the pair that was of better overall quality considering idea development and organization; for writing samples of similar quality, raters were instructed to select the writing sample that they would prefer to continue reading if it were longer. Our evaluation of writing quality based on idea development (which composition has more detailed and rich ideas from unique or interesting perspectives?) and organization of ideas (which composition has a more logical sequence, with a beginning, middle, and end, and better transitioning?) is consistent with prior research on developing writers (e.g., Kim et al., 2015, 2018).

We converted the paired comparison quality data to Elo ratings (see Pelánek, 2016, for a description) using the EloChoice R package (Neumann, 2019); the ratings represent the likelihood that writing samples would be rated as of higher quality in additional paired comparisons. The strong correlation between the Elo ratings from the two raters (r = .94) and an index of the proportion of comparisons

in which the writing sample with the highest prior Elo rating was rated as of better quality (.87; see Clark et al., 2018, for computational details) provide evidence for interrater reliability and the stability of the ratings.

2.2.1.2 aLPA Writing Quality

We used the writeAlizer R package (Mercer, 2020), which applies an ensemble scoring model to indices generated from the open-source ReaderBench text complexity analysis tool (Dascalu et al., 2018), to generate aLPA writing quality scores. The writeAlizer scoring model was originally trained based on 7-minute narrative writing samples from students in Grades 2–5 (see Mercer et al., 2019). In these analyses, seven machine learning algorithms (random forest regression, cubist regression, support vector machines with a radial kernel, elastic net regression, bagged multivariate adaptive regression splines, stochastic gradient boosted trees, and partial least squares regression; for details on these algorithms, see Hastie et al., 2009) were differentially weighted to predict writing quality ratings (i.e., Elo ratings similar to the ones used in the current study). In addition to a narrower grade range than in the current study (Grades 2–5 vs. Grades 2–12), only 6% of participants in the writeAlizer training set were receiving special education services, compared to the current participants with parent-reported learning difficulties. More details of scoring model development, including the weightings of ReaderBench indices in each algorithm and overall, are available in the writeAlizer online documentation: https://github.com/shmercer/writeAlizer. To facilitate interpretability, aLPA scores, across the fall and spring time points, were standardized before analyses.

2.2.1.3 WE-CBM Scoring

We used the guidelines of Hosp et al. (2016) to hand-score four WE-CBM metrics: total words written (TWW), words spelled correctly (WSC), correct word sequences (CWS), and correct minus incorrect word sequences (CIWS). TWW are the number of letters or groups of letters that are separated by spaces, even if the words are misspelled or used incorrectly in context. WSC are the number of words spelled correctly not considering context; words are counted as correct if they appear in the English language. CWS are the number of adjacent words that are acceptable in the English language considering spelling, punctuation, syntax, and semantics. CIWS is calculated from the number of CWS minus incorrect word sequences. Forty percent of the writing samples were independently scored by two raters – agreement was high (r = .99) for TWW, WSC, CWS, and CIWS scores.

2.2.2 Standardized Writing Assessment

Two subtests of the Test of Written Language, 4th edition (TOWL-4; Hammill & Larsen, 2009), requiring students to generate one picture-prompted narrative sample (5 minutes to plan, 15 minutes to write), were administered by the non-profit organization's staff and were scored by the research team. Scoring for contextual conventions considers spelling and grammatical errors, and scoring for story composition considers quality of vocabulary, plot, and interest to the reader. The TOWL-4 is commonly used as a criterion measure in validity studies for WE-CBM (Romig et al., 2017).

2.3 Data Analysis

To address Research Questions 1 and 2 that involve the relations of scores with rated quality on the scored writing samples and with standardized writing assessment scores, we calculated Pearson r correlation coefficients. Consistent with WE-CBM validity studies, we interpreted correlation coefficients of $r = .50$ as the minimally sufficient value for validity evidence (McMaster & Campbell, 2008), with values above $r = .60$ preferred and consistent with evidence standards used by the National Center on Intensive Intervention (2018) for academic screening tools. Differences between aLPA and WE-CBM validity coefficients were evaluated with Meng et al.'s (1992) z test for dependent correlation coefficients using the cocor package (Diedenhofen & Musch, 2015), with Bonferroni-adjusted p values reported to address multiple comparisons within the same time point and criterion measure. To address Research Question 3, we used a paired sample t test to determine if mean aLPA scores changed from fall to spring, with Hedges' g reported as an indicator of effect size.

3. Results

Means and standard deviations for all scores by time point are presented in Table 1. aLPA scores were highly correlated with WE-CBM scores at both time points (see Table 2); aLPA scores were most strongly related to TWW, WSC, and CWS scores (fall: $r = .91–.94$; spring: $r = .88–.90$), and also had large magnitude correlations with CIWS scores (fall: $r = .74$; spring: $r = .76$). Results are presented below by research question.

Table 1: Means, Standard Deviations, and Sample Sizes for all Scores by Time Point

	Fall			Spring		
	M	*SD*	*n*	*M*	*SD*	*n*
aLPA quality	−0.13	1.00	103	0.16	0.98	83
TWW	58.03	40.07	103	71.93	47.20	83
WSC	50.49	40.25	103	64.80	46.72	83
CWS	38.37	36.03	103	53.95	47.53	83
CIWS	12.77	35.59	103	28.81	51.16	83
Rated quality	−0.15	1.00	103	0.18	0.96	83
TOWL-4: CC	9.67	5.03	33	10.00	5.15	30
TOWL-4: SC	7.21	3.63	33	7.40	3.71	30

Note. aLPA = automated Learning Progress Assessment; TWW = Total Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; CIWS = Correct Minus Incorrect Word Sequences; TOWL-4 = Test of Written Language (4th ed.); CC = Contextual Conventions; SC = Story Composition. Although the TOWL-4 was administered only once at the spring time point, values are reported at both time points based on participants who completed both the narrative writing sample at that time point and the TOWL-4.

Table 2: Correlations of aLPA Quality With WE-CBM Scores

	Fall (n = 103)			Spring (n = 83)		
Score	*r*	95% CI for *r*		*r*	95% CI for *r*	
TWW	.91	.87	.94	.88	.81	.92
WSC	.94	.91	.96	.90	.85	.93
CWS	.93	.90	.95	.89	.83	.93
CIWS	.74	.63	.81	.76	.65	.84

Note. TWW = Total Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; CIWS = Correct Minus Incorrect Word Sequences.

3.1 Correlations With Rated Quality

Correlations of rated quality on the scored writing samples with aLPA quality scores and WE-CBM scores in the fall and spring of the academic year are presented in Table 3. Also, a scatterplot of the relation between aLPA and rated quality scores at the fall time point is presented in Figure 1. In general, correlations with rated quality were of large magnitude (all rs $\geq$ .79), with aLPA demonstrating higher correlations than all WE-CBM metrics at both time points, although the differences between the aLPA and WE-CBM correlation coefficients were of small magnitude and not always statistically significant. In the fall, the correlation of rated quality with aLPA quality (r = .93) was significantly greater than the correlation of rated quality with TWW (r = .81; Δr = .12, z = 6.34, p < .001), WSC (r =

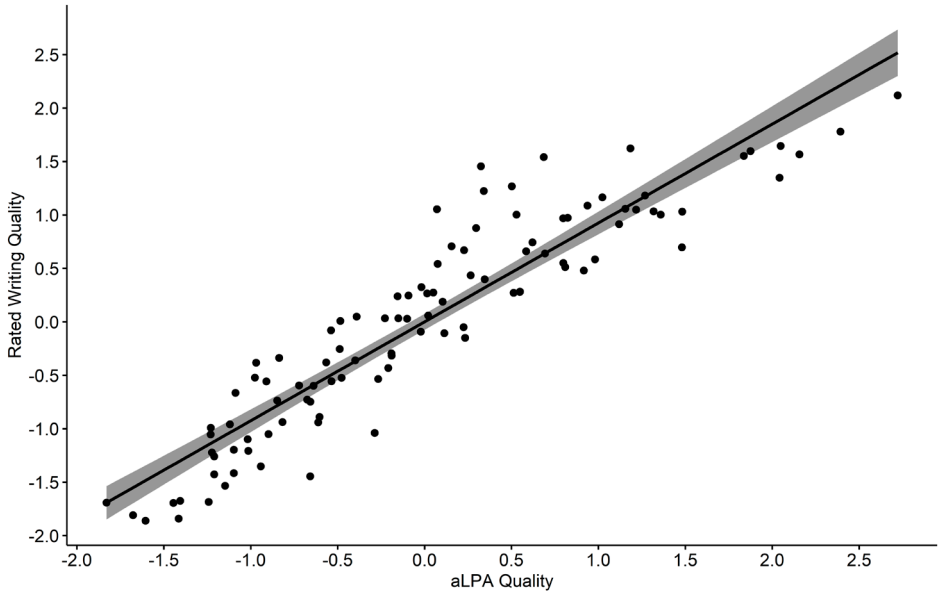
.86; Δr = .07, z = 4.43, p < .001), and CIWS (r = .79; Δr = .14, z = 4.95, p < .001), but did not statistically differ from the correlation of rated quality with CWS (r = .89; Δr = .04, z = 2.29, p = .111). In the spring, the correlation of rated quality with aLPA quality (r = .88) was significantly greater than the correlation of rated quality with TWW (r = .80; Δr = .08, z = 3.16, p = .006), but was not statistically different from the correlations of rated quality with WSC (r = .85; Δr = .03, z = 1.33, p = .739), CWS (r = .86; Δr = .02, z = 0.75, p = .999), or CIWS (r = .80; Δr = .08, z = 2.29, p = .089).

Table 3: Correlations With Rated Writing Quality by Time Point

	Fall (n = 103)			Spring (n = 83)		
Score	r	95% CI for r		r	95% CI for r	
aLPA quality	.93	.89	.95	.88	.82	.92
TWW	.81	.73	.87	.80	.70	.86
WSC	.86	.81	.90	.85	.78	.90
CWS	.89	.85	.93	.86	.80	.91
CIWS	.79	.70	.85	.80	.71	.87

Note. aLPA = automated Learning Progress Assessment; TWW = Total Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; CIWS = Correct Minus Incorrect Word Sequences.

Figure 1: Relation of aLPA and Rated Writing Quality at the Fall Time Point



Note. n = 103. aLPA and rated quality scores are standardized across the fall and spring time points (M = 0, SD = 1). The regression line is displayed as a solid line, with shading indicating its 95% confidence interval.

3.2 Correlations With a Standardized Writing Assessment

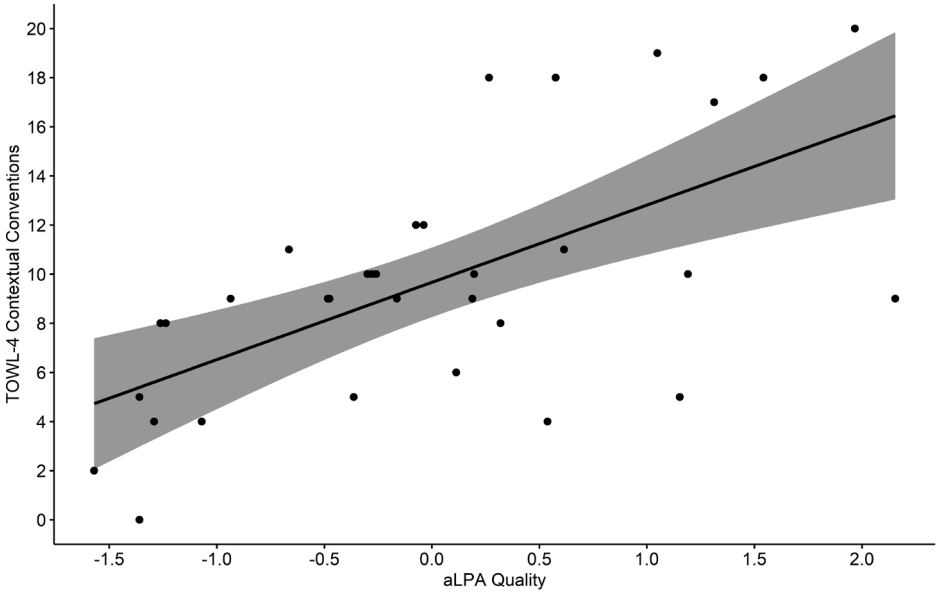
Validity coefficients in relation to the TOWL-4 subtest that assessed spelling and grammar (contextual conventions) and the subtest that assessed substantive quality (story composition) are presented below and in Table 4. Scatterplots of the relations between aLPA (fall time point) and the TOWL-4 subtest scores are presented in Figures 2 and 3.

Table 4: Correlations With a Standardized Writing Assessment

	Fall (n = 33)			Spring (n = 30)		
Score	r	95% CI for r		r	95% CI for r	
		TOWL-4 contextual conventions				
aLPA quality	.63	.36	.80	.72	.49	.86
TWW	.48	.16	.70	.59	.29	.78
WSC	.54	.24	.75	.65	.38	.82
CWS	.66	.42	.82	.68	.42	.84
CIWS	.67	.43	.83	.67	.41	.83
		TOWL-4 story composition				
aLPA quality	.44	.11	.68	.53	.20	.75
TWW	.34	.00	.61	.44	.10	.69
WSC	.36	.02	.63	.46	.12	.70
CWS	.45	.12	.69	.45	.10	.69
CIWS	.43	.10	.67	.38	.02	.65

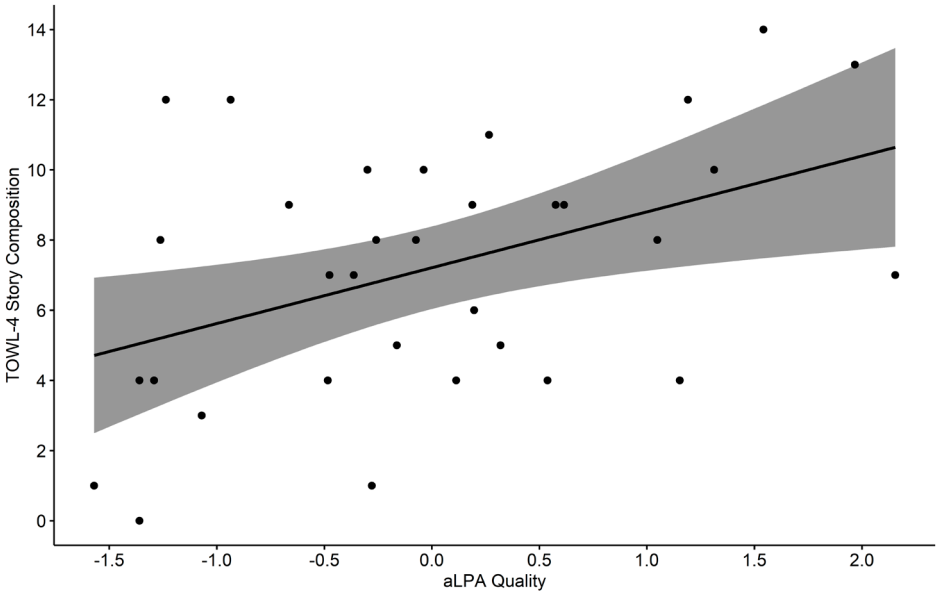
Note. TOWL-4 = Test of Written Language (4th ed.); aLPA = automated Learning Progress Assessment; TWW = Total Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; CIWS = Correct Minus Incorrect Word Sequences.

Figure 2: Relation of aLPA Quality and TOWL-4 Contextual Conventions at the Fall Time Point



Note. n = 33. aLPA quality scores are standardized across the fall and spring time points (M = 0, SD = 1). The regression line is displayed as a solid line, with shading indicating its 95% confidence interval.

Figure 3: Relation of aLPA Quality and TOWL-4 Story Composition at the Fall Time Point



The regression line is displayed as a solid line, with shading indicating its 95% confidence interval.

3.2.1 Contextual Conventions

For the fall time point, aLPA (r = .63), CWS (r = .66), and CIWS (r = .67) had correlations above the r = .60 standard for criterion-related validity coefficients. The validity coefficient for aLPA not statistically different from the coefficients for TWW (r = .48; Δr = .15, z = 2.32, p = .082), WSC (r = .54; Δr = .09, z = 1.74, p = .325), CWS (Δr = −.03, z = 0.77, p = .999), or CIWS (Δr = −.04, z = 0.49, p = .999). At spring, all coefficients were at or near the r = .60 standard, with the aLPA coefficient (r = .72) again not statistically different than the coefficients for TWW (r = .59; Δr = .13, z = 2.48, p = .053), WSC (r = .65; Δr = .07, z = 1.59, p = .448), CWS (r = .68; Δr = .04, z = 0.71, p = .999), or CIWS (r = .67; Δr = .05, z = 0.46, p = .999).

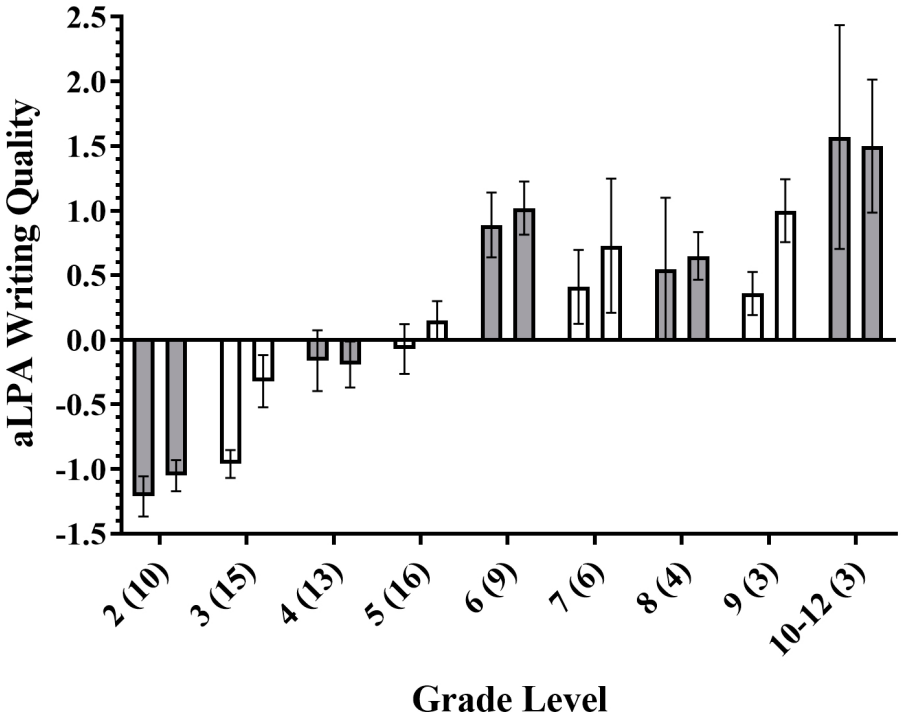
3.2.2 Story Composition

For the fall time point, the coefficients for aLPA (r = .44) and all WE-CBM scores were below the minimally sufficient standard of r = .50. There were no significant differences between the validity coefficients for aLPA and TWW (r = .34; Δr = .10, z = 1.34, p = .724), WSC (r = .36; Δr = .08, z = 1.40, p = .642), CWS (r = .45; Δr = −.01, z = 0.14, p = .999), or CIWS (r = .43; Δr = .01, z = 0.11, p = .999). For the spring time point, only the coefficient for aLPA (r = .53) was above the minimally sufficient standard; however, the aLPA coefficient again did not statistically differ from the coefficients for TWW (r = .44; Δr = .09, z = 1.31, p = .762), WSC (r = .46; Δr = .07, z = 1.22, p = .887), CWS (r = .45; Δr = .08, z = 1.13, p = .999), and CIWS (r = .38; Δr = .15, z = 1.10, p = .999).

3.3 Sensitivity to Skill Growth

For the 79 students who completed both the fall and spring narrative writing samples, aLPA scores increased from fall (M = −0.14) to spring (M = 0.11), $t(78)$ = 3.37, p = .001. Considering effect size (g = 0.25), the increase was of small magnitude. Fall and spring aLPA scores by grade level are presented in Figure 4.

Figure 4: Change in aLPA Writing Quality From Fall to Spring by Grade



Note. n = 79. aLPA scores are standardized across the fall and spring time points (M = 0, SD = 1). For each grade level, the left bar is the mean for the fall time point and the right bar is the mean for the spring time point. Values in parentheses are within-grade sample sizes.

4. Discussion

The purpose of this study was to evaluate the evidence in support of several key inferences for the use of aLPA to progress monitor the writing skills of students with parent-identified learning difficulties. First, in support of the scoring inference assumption that aLPA scores represent writing quality, we found aLPA scores to be highly correlated with quality ratings on the scored writing samples (r = .88 and .93). Although WE-CBM metrics were also highly correlated with ratings of writing quality (rs = .79 to .89), it is important to note that the writing sample duration in this study (10 minutes) was longer than in typical WE-CBM procedures (3 minutes). Second, we found partial support for the extrapolation inference assumption that aLPA scores can serve as general indicators of writing skill. In support of this assumption, aLPA scores were highly correlated, above the r = .60 criterion, with scores at both time points on a standardized writing subtest that primarily assesses spelling and grammar; in addition, the aLPA score at the spring time point was above the minimally acceptable threshold of r = .50 for the standardized sub-

test assessing substantive writing quality. There was minimal evidence of improved extrapolation for aLPA relative to WE-CBM, in large part because aLPA and WE-CBM scores were highly correlated; validity coefficients were generally very similar for aLPA and the more complex WE-CBM scores of CWS and CIWS. Third, in support of the decision inference assumption that aLPA scores are sensitive to writing skill growth for students with learning difficulties, we found small, statistically significant improvements in scores across the fall and spring of 1 academic year.

4.1 Validation of Formative Writing Assessments

Although early stage, the current study has implications for future research investigating the validity of formative writing assessments. Consistent with contemporary validation approaches (American Educational Research Association et al., 2014; Kane, 2013), we provide an interpretation and use validity argument to frame the current study and guide our future research on aLPA. In doing so, we contrast aLPA with WE-CBM; this contrast is most notable in reference to our scoring inference that scores should represent quality on the scored writing samples. Related to the behavioral orientation of foundational CBM research, CBM scores are typically assumed to be direct measures of observable behaviors (Christ et al., 2016). Although CBM scores are clearly intended to serve as general indicators of skill, which we also specify in our extrapolation inference for aLPA, there is some theoretical debate as to whether CBM scores assess the construct of fluency or indeed any latent construct (Espin & Deno, 2016). This limited attention to scoring inferences in CBM research can be problematic when WE-CBM scores are criticized as solely representing text length instead of writing quality (Gansle et al., 2002; Ritchey & Coker, 2013). Our current findings provide some evidence that WE-CBM scores assess quality on the scored writing samples, but with the caveat that we used a longer writing duration that may increase the likelihood that students had adequate time to engage in the writing process. We also provide evidence that aLPA scores are more highly correlated with quality ratings on the scored writing samples than counts of the number of total words written, thereby demonstrating that aLPA scores are assessing quality, not just composition length.

Formally specifying and evaluating the scoring inference also provides key information to interpret our extrapolation findings. Although extrapolation of aLPA to the standardized subtest assessing writing mechanics was good, extrapolation of aLPA scores to the subtest assessing substantive writing quality was not consistently adequate. Given the support for our scoring inference that aLPA scores can represent quality on the evaluated writing samples, the limited extrapolation to the substantive quality subtest appears to be an issue of limited generalization rather than problems in the aLPA scoring model – specifically, we may need to calculate aLPA scores across more than one narrative writing sample to have adequate generalization, and in turn improved extrapolation to other standardized writing assessments. We did not evaluate the generalization inference in the current study,

but other research indicates that more than one writing sample is needed for adequate generalization and reliable scores (Keller-Margulis et al., 2016; Kim et al., 2017). Also, in the current study, both the aLPA writing samples and standardized writing assessment were based on narrative compositions; evaluating generalization and extrapolation across writing genres will be necessary in future aLPA research.

4.2 Limitations

Several limitations in the current study should be considered. Because we analyzed extant data collected by the nonprofit organization's tutors as part of typical service delivery, we do not have information to support the fidelity of assessment administration procedures or have detailed information on the tutoring provided to students during the academic year. The tutors, however, did follow written assessment instructions, and we present evidence of scoring reliability for the hand scoring (WE-CBM) and quality ratings completed by the research team. In addition, although the narrative writing task used for aLPA scoring was designed by the nonprofit agency, it was a standalone assessment that was not fully embedded in their curriculum. In the future, it may be helpful to use aLPA on initial drafts from writing tasks that are later revised based on feedback and included in writing portfolios that are shared with others; by doing so, the writing task may have more value for the students, increasing motivation and eliciting better writing (Calfee & Miller, 2007). There are also some limitations related to sampling: (a) detailed demographic information for participants was not available from the nonprofit organization, which may complicate evaluations of potential generalizability to other contexts; (b) the standardized writing assessment was completed by a non-random 31% of students, contributing to wide confidence intervals around the validity coefficients that limited our ability to test differences between aLPA and WE-CBM scores; and (c) the sample size at each grade level was too small to permit within-grade level analyses, potentially leading to larger magnitude validity coefficients given that across-grade WE-CBM validity coefficients tend to be higher than within-grade coefficients (McMaster & Espin, 2007).

4.3 Conclusion

Overall, the current findings provide preliminary evidence that aLPA can be used to score narrative writing samples from students with parent-identified learning difficulties for writing quality, with preliminary evidence that aLPA scores are sensitive to student growth and some mixed evidence that aLPA scores can be extrapolated as indicators of general writing skill. Considering that correlations of aLPA scores were comparable to the best-performing hand-scored WE-CBM metrics, the use of automated text evaluation in aLPA may increase the likelihood that teach-

ers will use aLPA to monitor the written expression progress of their students due to the reduced scoring time required. Because the writeAlizer scoring model used in the current study was originally trained on writing samples from students mostly without learning difficulties, future research should investigate if training an additional scoring model on writing samples from students with learning difficulties would improve performance. In addition, future research should determine the number and duration of writing samples necessary for adequate generalization of aLPA scores, potentially improving the ability of aLPA scores to extrapolate as general indicators of writing skill. By conducting such studies in collaboration with community schools and agencies, the likelihood that aLPA can be fully embedded in the curriculum will increase, thereby making assessment more meaningful to the students and teachers.

Acknowledgments

The research reported here was supported in part by grants from the Social Sciences and Humanities Research Council of Canada and the Institute of Education Sciences, U.S. Department of Education, through Grant R305A190100. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

We would like to thank Norlan Cabot, Kate Raven, and the Learning Disabilities Association of Greater Vancouver for the research partnership that supported the current study. We would also like to thank Michèle P. Cheng, Eun Young Kwon, and Ioanna K. Tsiriotakis for their assistance in scoring writing samples and thank Daniela P. Blettner for the German translation of the abstract.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Berninger, V., & Amtmann, D. (2003). Preventing written expression disabilities through early and continuing assessment and intervention for handwriting and/or spelling problems: Research into practice. In H. L. Swanson, K. Harris, & S. Graham (Eds.), *Handbook of research on learning disabilities* (pp. 345–363). Guilford.

Calfee, R. C., & Miller, R. G. (2007). Best practices in writing assessment. In S. Graham, C. A. MacArthur, & J. Fitzgerald (Eds.), *Best practices in writing instruction* (pp. 265–286). Guilford.

Christ, T. J., van Norman, E. R., & Nelson, P. M. (2016). Foundations of fluency-based assessments in behavioral and psychometric paradigms. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 143–163). Springer.

Clark, A. P., Howard, K. L., Woods, A. T., Penton-Voak, I. S., & Neumann, C. (2018). Why rate when you could compare? Using the "EloChoice" package to assess

pairwise comparisons of perceived physical strength. *PLoS ONE, 13*(1), Article e0190393. https://doi.org/10.1371/journal.pone.0190393

Dascalu, M., Crossley, S. A., McNamara, D. S., Dessus, P., & Trausan-Matu, S. (2018). Please ReaderBench this text: A multi-dimensional textual complexity assessment framework. In S. D. Craig (Ed.), *Tutoring and intelligent tutoring systems* (pp. 251–271). Nova Science.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219–232. https://doi.org/10.1177/001440298505200303

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE, 10*(4), Article e0121945. https://doi.org/10.1371/journal.pone.0121945

Espin, C. A., & Deno, S. L. (2016). Conclusion: Oral reading fluency or reading aloud from text: An analysis through a unified view of construct validity. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 365–384). Springer.

Espin, C. A., Scierka, B. J., Skare, S., & Halverson, N. (1999). Criterion-related validity of curriculum-based measures in writing for secondary school students. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 15*(1), 5–27. https://doi.org/10.1080/105735699278279

Fewster, S., & Macmillan, P. D. (2002). School-based evidence for the validity of curriculum-based measurement of reading and writing. *Remedial and Special Education, 23*(3), 149–156. https://doi.org/10.1177/07419325020230030301

Filderman, M. J., Toste, J. R., Didion, L. A., Peng, P., & Clemens, N. H. (2018). Data-based decision making in reading interventions: A synthesis and meta-analysis of the effects for struggling readers. *The Journal of Special Education, 52*(3), 174–187. https://doi.org/10.1177/0022466918790001

Förster, N., & Souvignier, E. (2014). Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept. *Learning and Instruction, 32*, 91–100. https://doi.org/10.1016/j.learninstruc.2014.02.002

Förster, N., & Souvignier, E. (2015). Effects of providing teachers with information about their students' reading progress. *School Psychology Review, 44*(1), 60–75. https://doi.org/10.17105/SPR44-1.60-75

Fuchs, D., McMaster, K. L., Fuchs, L. S., & Al Otaiba, S. (2013). Data-based individualization as a means of providing intensive instruction to students with serious learning disorders. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (2nd ed., pp. 526–544). Guilford.

Furey, W. M., Marcotte, A. M., Hintze, J. M., & Shackett, C. M. (2016). Concurrent validity and classification accuracy of curriculum-based measurement for written expression. *School Psychology Quarterly, 31*(3), 369–382. https://doi.org/10.1037/spq0000138

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review, 31*(4), 477–497. https://doi.org/10.1080/02796015.2002.12086169

Hammill, D. D., & Larsen, S. C. (2009). *Test of written language 4 (TOWL-4)*. Pro-Ed Assessments.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement* (2nd ed.). Guilford.

Jung, P. G., McMaster, K. L., Kunkel, A. K., Shin, J., & Stecker, P. M. (2018). Effects of data-based individualization for students with intensive learning needs: A meta-analysis. *Learning Disabilities Research & Practice, 33*(3), 144–155. https://doi.org/10.1111/ldrp.12172

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Keller-Margulis, M. A., Mercer, S. H., & Matta, M. (2021). Validity of automated text evaluation tools for written-expression curriculum-based measurement: A comparison study. *Reading and Writing: An Interdisciplinary Journal, 34*(10), 2461–2480. https://doi.org/10.1007/s11145-021-10153-6

Kim, Y.-S., Al Otaiba, S., Wanzek, J., & Gatlin, B. (2015). Toward an understanding of dimensions, predictors, and the gender gap in written composition. *Journal of Educational Psychology, 107*(1), 79–95. https://doi.org/10.1037/a0037210

Kim, Y. G., Gatlin, B., Al Otaiba, S., & Wanzek, J. (2018). Theorization and an empirical investigation of the component-based and developmental text writing fluency construct. *Journal of Learning Disabilities, 51*(4), 320–335. https://doi.org/10.1177/0022219417712016

Kim, Y. G., Schatschneider, C., Wanzek, J., Gatlin, B., & Al Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Reading and Writing, 30*(6), 1287–1310. https://doi.org/10.1007/s11145-017-9724-6

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing. *The Journal of Special Education, 41*(2), 68–84. https://doi.org/10.1177/0022466907041002030 1

McMaster, K. L., & Campbell, H. (2008). New and existing curriculum-based writing measures: Technical features within and across grades. *School Psychology Review, 37*(4), 550–556. https://doi.org/10.1080/02796015.2008.12087867

McMaster, K. L., Fuchs, D., Fuchs, L. S., & Compton, D. L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children, 71*(4), 445–463. https://doi.org/10.1177/001440290507100404

McMaster, K. L., Lembke, E. S., Shin, J., Poch, A. L., Smith, R. A., Jung, P.-G., Allen, A. A., & Wagner, K. (2020). Supporting teachers' use of data-based instruction to improve students' early writing skills. *Journal of Educational Psychology, 112*(1), 1–21. https://doi.org/10.1037/edu0000358

McMaster, K. L., Shin, J., Espin, C. A., Jung, P.-G., Wayman, M. M., & Deno, S. L. (2017). Monitoring elementary students' writing progress using curriculum-based measures: Grade and gender differences. *Reading and Writing, 30*(9), 2069–2091. https://doi.org/10.1007/s11145-017-9766-9

Meng, X.-I., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*(1), 172–175. https://doi.org/10.1037/0033-2909.111.1.172

Mercer, S. H. (2020). *writeAlizer: Generate predicted writing quality and written expression CBM scores* (Version 1.2.0). Retrieved from https://github.com/shmercer/writeAlizer/

Mercer, S. H., Keller-Margulis, M. A., Faith, E. L., Reid, E. K., & Ochs, S. (2019). The potential for automated text evaluation to improve the technical adequacy of written expression curriculum-based measurement. *Learning Disability Quarterly, 42*(2), 117–128. https://doi.org/10.1177/0731948718803296

Mercer, S. H., Martínez, R. S., Faust, D., & Mitchell, R. R. (2012). Criterion-related validity of curriculum-based measurement in writing with narrative and expository prompts relative to passage copying speed in 10th grade students. *School Psychology Quarterly, 27*(2), 85–95. https://doi.org/10.1037/a0029123

Morgan, P. L., & Sideridis, G. D. (2006). Contrasting the effectiveness of fluency interventions for students with or at risk for learning disabilities: A multilevel random coefficient modeling meta-analysis. *Learning Disabilities Research & Practice, 21*(4), 191–210. https://doi.org/10.1111/j.1540-5826.2006.00218.x

National Center on Intensive Intervention. (2018). *Academic screening tools chart rating rubric*. Retrieved from https://intensiveintervention.org/sites/default/files/NCII_AcademicScreening_RatingRubric_July2018.pdf

Neumann, C. (2019). *EloChoice: Preference rating for visual stimuli based on Elo ratings* (Version 0.29.4). Retrieved from https://CRAN.R-project.org/package=EloChoice

Pelánek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers & Education, 98*, 169–179. https://doi.org/10.1016/j.compedu.2016.03.017

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*(6), 427–469. https://doi.org/10.1016/j.jsp.2009.07.001

Ritchey, K. D., & Coker, D. L., Jr. (2013). An investigation of the validity and utility of two curriculum-based measurement writing tasks. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 29*(1), 89–119. https://doi.org/10.1080/10573569.2013.741957

Ritchey, K. D., McMaster, K. L., Al Otaiba, S., Puranik, C. S., Kim, Y. G., Parker, D. C., & Ortiz, M. (2016). Indicators of fluent writing in beginning writers. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 21–66). Springer.

Romig, J. E., Therrien, W. J., & Lloyd, J. W. (2017). Meta-analysis of criterion validity for curriculum-based measurement in written language. *The Journal of Special Education, 51*(2), 72–82. https://doi.org/10.1177/0022466916670637

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286. https://doi.org/10.1037/h0070288

Videen, J., Deno, S. L., & Martson, D. (1982). *Correct word sequences: A valid indicator of proficiency in written expression*. University of Minnesota, Institute for Research on Learning Disabilities.

Werb, J. (2007, October 31). *Provincial government fails learning-disabled kids*. The Georgia Straight. https://www.straight.com/article-116366/provincial-government-fails-learning-disabled-kids

Wilson, J. (2018). Universal screening with automated essay scoring: Evaluating classification accuracy in grades 3 and 4. *Journal of School Psychology, 68*, 19–37. https://doi.org/10.1016/j.jsp.2017.12.005

Wilson, J., Chen, D., Sandbank, M. P., & Hebert, M. (2019). Generalizability of automated scores of writing quality in Grades 3–5. *Journal of Educational Psychology, 111*(4), 619–640. https://doi.org/10.1037/edu0000311

Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01.04
Volume 14 (2022), No. 1, 61–88
CC BY-NC-SA 4.0 Waxmann 2022

Christin Vanauer, Sarah Chromik, Philipp Doebler,
& Jörg-Tobias Kuhn

Curriculum-Based Measurement of Basic Arithmetic Competence: Do Different Booklets Represent the Same Ability?

Abstract

An important prerequisite of progress monitoring as one source to support instructional decision-making is the existence of equivalent booklets. This study assesses this prerequisite with respect to a German elementary school math curriculum-based measurement instrument (LVD-M 2-4; Strathmann & Klauer, 2012). Every second week of a 19-weeks period, n = 108 third and n = 109 fourth graders (regular instruction) completed one of ten parallel booklets, each containing 24 arithmetic tasks. Analyses with (generalized) linear mixed models showed that in both grades the between-booklet variance was so small in relation to the between student variance that it was practically irrelevant. This corresponds to the key assumption of the binomial model that equivalent scores from different booklets reflect the same ability. While item difficulty varied within some of the tasks, the effect was insubstantial in comparison with the variance between students. These findings were replicated in two intervention samples of an RTI study. The parallel booklets can therefore be regarded as equivalent for typical applied purposes. Implications of these findings for curriculum-based measurement and booklet design are discussed.

Keywords

curriculum-based measurement, elementary school mathematics, formative assessment, booklet equivalence

---

Dr. Christin Vanauer (corresponding author) · Prof. Dr. Jörg-Tobias Kuhn, Faculty of Rehabilitation Sciences, TU Dortmund University, Emil-Figge-Str. 50, 44227 Dortmund, Germany
email: christin.vanauer@tu-dortmund.de
tobias.kuhn@tu-dortmund.de

Sarah Chromik, M. Sc. · Prof. Dr. Philipp Doebler, Faculty of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany
email: sarah.chromik@tu-dortmund.de
doebler@tu-dortmund.de

Verlaufsdiagnostik arithmetischer Grundkompetenzen: Messen verschiedene Booklets die gleiche Fähigkeit?

Zusammenfassung

Eine wichtige Voraussetzung dafür, Lernverlaufsdiagnostik für instruktionale Entscheidungen nutzen zu können, sind äquivalente Testbooklets. Diese Studie prüft diese Voraussetzung für die „Lernverlaufsdiagnostik – Mathematik 2-4" (LVD-M 2-4; Strathmann & Klauer, 2012). Über 19 Wochen hinweg bearbeiteten n = 108 Drittklässler:innen und n = 109 Viertklässler:innen in zweiwöchigem Abstand zehn verschiedene Paralleltests mit je 24 arithmetischen Aufgaben. Mithilfe (generalisierter) gemischter linearer Modelle wurden Booklet-Effekte in Relation zur Leistungsvarianz zwischen den Kindern gesetzt. Damit wurde die Kernannahme des Binomial-Modells geprüft, dass gleiche Scores aus verschiedenen zufallsgenerierten Booklets die gleiche latente Fähigkeit abbilden sollten. In beiden Klassenstufen fiel die Between-Booklet-Varianz in Relation zur Varianz zwischen den Kindern sehr gering aus. Für einige Aufgabentypen variierte die Schwierigkeit zwar zwischen den Booklets, war verglichen mit der Varianz der Schülerleistung aber nicht substanziell. Die Befunde ließen sich in zwei Interventionsgruppen einer RTI-Studie replizieren. Die Booklets können also für typische Anwendungszwecke als äquivalent angesehen werden. Die Implikationen dieser Befunde werden vor dem Hintergrund von Lernverlaufsdiagnostik und der Konstruktion äquivalenter Testbooklets diskutiert.

Schlagworte

Lernverlaufsdiagnostik, Mathematik in der Grundschule, formatives Assessment, Booklet-Äquivalenz

1. Theoretical Background

1.1 Standardized Diagnostics as a Contribution to Instructional Decision-Making

Instructional decision-making in schools is complex. Educational psychology can assist teachers by providing and evaluating instruments to systematically assess clear-cut competences of their students. While test construction in the domains of reading and writing faces a longer tradition, mathematical learning has mainly been targeted in the last two decades. In this period, a considerable number of tests have been published to help assess mathematical achievement in general, but also to identify children at risk for dyscalculia. Test backgrounds range from neuropsychological or developmental models to curricular or even higher-order national and international standards (Köller & Reiss, 2013; Kuhn & Schwenk, 2018).

Independent of the domain two main categories of standardized diagnostics can be distinguished: summative and formative approaches (Klauer, 2014), which differ in their purpose and measurement frequency. Summative tests (status diagnostics) evaluate a student's performance at a given point in time and help to identify individuals beyond a clinically relevant threshold. Unless designed as a screening, most status diagnostics are time-consuming and not suitable for large groups in everyday school practice. By contrast, formative testing, or progress monitoring, is based on multiple measurements that usually occur repeatedly (e.g., weekly). In contrast to summative testing, progress monitoring captures learning *courses* or trajectories (Deno, 2003). Progress monitoring is applicable to any competence on which subjects are expected to make visible progress in a reasonable period of time. It is described and recommended for mainly elementary school reading, spelling, and mathematics (especially basic arithmetic operations), but also for precursor abilities like letter naming or number identification (Foegen et al., 2007; Hosp et al., 2016).

1.2 Progress Monitoring in Special Needs Education and Beyond

Progress monitoring has gained particular importance in special needs education, for example, being a core feature of the *response to intervention* (RTI) paradigm, a preventive and inclusive educational concept with roots in the USA and transfer to other countries (e.g., Germany: Huber & Grosche, 2012; Voß, 2016, or Finland: Björn et al., 2018). In RTI, formative assessment helps to dynamically keep track of key learning criteria and thereby assign students to one of three tiers of educational support between regular classroom instruction (Level 1) and intensive single-subject fostering (Level 3). In frameworks like RTI, two things are important to bear in mind regarding the informative value of progress-monitoring tests: First, aggregated progress-monitoring outcomes (e.g., number of correctly read words in 1 minute or number of correctly solved calculations in a set of 24 tasks) are not sufficiently fine-grained to unravel the educational needs of low-performing students. In this case, class-level progress monitoring can only serve as a screening tool, painting the "big picture". Qualitative diagnostics like error analysis (e.g, Ashlock, 2005; Gerster, 2012) and/or the "thinking aloud" method (e.g., Lawson & Rice, 1987) are needed to find out where a specific student struggles. Second, progress monitoring is not restricted to this subgroup with special educational needs: It potentially informs data-based instruction of students in the whole ability range, even of very high achievers (Hebbecker & Souvignier, 2016). Therefore, research on the construction and analysis of progress-monitoring assessments is relevant for all types of schools and educational practice.

1.3 Particularities of Progress-Monitoring Tests

Progress-monitoring tests must meet certain requirements which are subject of research. To systematize research on curriculum-based measurement (CBM), Lynn Fuchs (2004) describes three stages. The first stage focuses on the psychometric properties (reliability, validity) of a progress-monitoring instrument and is not much different from the evaluation of status tests. However, the question of how to generate the required *set* of equivalent test booklets is not trivial. Several psychometric challenges are specifically linked to *progress* diagnostics: For instance, the "difficulty" of a specific item as well as the difficulty of the entire test has to decrease over time if learning progress is made (Rohwer, 2015). Simply speaking, a student is expected to make fewer errors over time in case that items are very similar. Moreover, test takers' ability is likely to develop differently in between (with inter-individually different growth curves). Most importantly, as a matter of validity, an increasing test score across time should not be due to rote learning of specific, repeatedly administered items but due to an increasing mastery of the competence represented by them.

This leads to the methodologically crucial second research stage outlined by Fuchs (2004). Studies in this stage assess whether a given instrument is in fact able to measure the core construct learning *progress* (also called "sensitivity to change"; Klauer & Strathmann, 2013). In addition to performance variability across measurement points, the validity of the learning slope as a predictor of learning progress, or as a criterion for response to intervention, is investigated. Most of respective studies are based on reading (Schatschneider et al., 2008; Stage & Jacobsen, 2001), only few on mathematical instruments, and they show heterogeneous results: Some studies report substantial correlations between learning progress and later achievement (Keller-Margulis et al., 2008), others do not find any or only negligible incremental validity for learning slopes (i.e., progress lines) beyond baseline ability (e.g., Shapiro et al., 2015). This inconclusive pattern could be partially related to the lack of booklet equivalence: an important prerequisite for data-based decision-making, which is addressed in third-stage research according to Fuchs (2004).

1.4 Construction of Progress-Monitoring Instruments

To overcome psychometric dilemmas linked to progress monitoring, the framework of generic test construction (Rohwer, 2015) defines performance based on correctly solved representations of task types (e.g., mental addition as a category of similar problems), each consisting of a pool of structurally similar (and thus theoretically equivalent) items. Hence, a new test booklet generated by randomly sampling an item of each task type.

Because the idea that the *same* test needs to be repeated is more a theoretical (i.e., content validity of the tasks types) than a technical (i.e., psychometric) matter in the first place, it is important that task types represent a meaningful standard. In the context of regular schools and average student populations, the curriculum is an obvious standard, as it defines the learning goals of each grade. In line with this, learning progress monitoring is predominantly thought of as CBM since emerging in the early 1970s (Deno, 1985; Klauer, 2014). Deducing appropriate task types from curricula is referred to as curriculum-sampling. Applying this approach to elementary school mathematics, the tasks included in CBM represent the set of, for example, arithmetic problem types that children of a specific grade should master at the end of the school year. Mathematical CBMs known to the authors of this study mainly focus on arithmetic operations (addition, subtraction, multiplication, division), but there are also examples comprising other curricular mathematical domains like geometry (e.g., quop; Souvignier, 2018).

However, curriculum-sampling is just one possibility, next to concentrating on so-called robust indicators (Fuchs, 2004; Schwenk et al., 2017). Robust indicators are key competences that are empirically valid for the achievement development in a specific domain and might be more basic than the curricular goals for a given grade. The main advantage of this approach is its flexibility: Robust indicator tests are seamlessly applicable across grade boundaries. While passage reading fluency is a well-established robust indicator in reading, there is no equivalent in mathematics. Foegen (2007) explains this with the differences of curricula, with math, compared to reading, learning being more complex and better represented by curriculum-sampling instead of robust indicators.

1.5 Booklet Equivalence

A key condition to interpret learning trajectories unambiguously is the assumption of equivalent test forms. Only when different test forms are equivalent can longitudinal test scores be considered as valid measures of individual learning. However, the degree to which different booklets of progress-monitoring instruments are equivalent or influenced by design features has only been addressed explicitly by few studies, mainly in the field of reading (CBM-R): Absolute estimates of weekly growth rates and intercepts as well as standard errors were shown to vary across different texts ("passages") used for oral reading CBM (Ardoin & Christ, 2009; Francis et al., 2008). Therefore, absolute scores in CBM, that is, words read correctly per minute, should be interpreted with caution (Ardoin et al., 2013). By contrast, relative estimates of growth rates, that is rank orders of students, are more reliable.

Research in the field of mathematics (CBM-M) is comparably scarce (Montague et al., 2010). Like in CBM-R, Christ and Vining (2006) describe an "excessive reliance on research that has examined relative score interpretation" (p. 398). Stratified instead of randomly ordered booklets, that is, a presentation of items struc-

tured by the skill they assess (Christ & Vining, 2006), seem to enhance booklet equivalence.

How can cross-booklet equivalence be tested from a technical point of view? In the literature, the concept of equivalence has been investigated from varying perspectives, using at least three different yet related methodological approaches. First, in classical test theory, which operates at the level of test scores, two test forms are considered as parallel in the case that their mean test scores and test score variances are equal (e.g., McDonald, 1999). In the case of differing means and/or variances, equivalence can be established using test equating (e.g., Kolen & Brennan, 2014). In this approach, item parameters are of secondary importance. For example, Strathmann and Klauer (2010) compared means of adjacent math progress-monitoring test scores at each measurement point.

In contrast, item response theory (IRT) is a second approach that allows investigating test equivalence at the item level by analyzing differential item functioning (DIF). If no DIF exists between two test forms (i.e., if item parameters are identical), the test forms can be considered equivalent. Taking a less restrictive stance, equivalence of test forms within an IRT framework can also be assumed in the case that test information functions (TIF) are similar (Förster & Kuhn, 2021).

A third approach of analyzing test form equivalence relates to investigating the relative proportion of test score variance that can be attributed to test forms, in contrast to other factors (e.g., students, testing occasions). This is in the tradition of generalizability theory (e.g., Brennan, 2001). In this approach, the equivalence of test forms can be expressed in a relative way: The degree of equivalence is high if the variance attributable to test forms is very small, compared to the variance related to other factors. For example, Fan and Hansmann (2015) investigated a CBM of oral reading fluency, and found that 2.8% of total variance was attributable to probe variability, whereas 90.2% of variance was due to student reading skill. In the present study, we investigated test equivalence using this approach. In case the variance is solely attributable to student ability and unspecific error, but not to items/item families, there is neither DIF nor differential test functioning (DTF; i.e., when discussing Rasch models). In this sense, this is a strict approach, as the global absence of item (family) level variance implies equivalence in the IRT tradition and hence the classical test theory (CTT) tradition.

1.6 Summary: Aim of this Study

The key prerequisite of CBM instruments – a valid pool of different but psychometrically equivalent booklets – is often theoretically taken for granted. This is coupled with the assumption that identically distributed test scores stemming from different booklets are equivalent, which is part of a binomial model (cf. Klauer, 2011). Within the binomial model and under the prerequisite of random (stratified) item

sampling, an individual's ability equals the proportion of correctly solved items on a power test, that is, accuracy.

Against this background, the aim of the present study is to test the booklet equivalence assumption of the binomial model empirically. Based on a curriculum-based measurement in mathematics, more specifically: basic arithmetic (LVD-M 2-4; Strathmann & Klauer, 2012), it takes a closer look at differences between booklets as one possible source to explain the variance of third and fourth graders' scores.

2. Method

2.1 Sample

The sample of this study is part of a larger research project funded by the German Federal Ministry of Education and Research (BMBF) and approved by a local ethics committee, with the aim to evaluate the effects of dyscalculia interventions within an RTI framework. A total of n = 687 students (n = 345 Grade 3, n = 342 Grade 4) of ten elementary schools in the same German region with urban and rural parts participated in an initial screening at the beginning of school year 2015/16.

After screening, schools were assigned to one of three experimental groups of roughly the same size (> 100 per group in each of the two Grades 3 and 4): a waiting control group and two intervention groups who completed bi-weekly basic arithmetic progress-monitoring tests throughout a 19-weeks phase (more closely described in Section 2.3). During this study period, all children of the waiting control group received regular classroom instruction, irrespective of their performance in the initial screening. In the two intervention groups, those children with at-risk level performance or below (PR ≤ 25) on the initial screening (arithmetic subscale of a grade-specific German mathematics test) received trainings according to different intervention schemes: One group followed a two-tiered schemed with computer-based training (Kuhn & Holling, 2014) for all children with at- or below-risk level performance and a three-tiered group with additional within-person small group training for the weakest 10% according to the initial arithmetic screening.

As the research aim of this study applies to regular school practice, we focus on the descriptions and results of the regularly instructed waiting control group: Children who did not complete any of the ten progress-monitoring tests (cf. Section 2.3) were excluded, leading to a final (waiting control) sample of n = 108 children in Grade 3 and n = 109 in Grade 4, with eighteen classrooms of four schools (Table 1). The number of participating children per classroom (i.e., children with parental consent obtained before participation) ranged between 6 and 24 (M = 12.06, SD = 4.78, median = 11). Next to mathematical ability measured with the arithmetic subtests of DEMAT 2+ (Krajewski et al., 2004) and DEMAT 3+ (Roick et al.,

2004) as part of the screening, general intelligence (CFT 1-R; Weiß & Osterland, 2013, or CFT 20-R; Weiß, 2006) and reading speed (SLS; Wimmer & Mayringer, 2014) were captured for all children participating in the study (Table 1).

Table 1: Sample (Waiting Control Group)

School	Grade 3 (n = 108, 58.3% male)					Grade 4 (n = 109, 55.0% male)				
		A	B	C	D		A	B	C	D
	Students	26	23	40	19	Students	21	16	48	24
	Classrooms	2	3	2	2	Classrooms	2	2	3	2
Abilities	*M*	*SD*	Min	Max		*M*	*SD*	Min	Max	
Intelligence (CFT, IQ)	101.96	13.12	67.0	138.0		107.94[b]	16.52	61.0	148.0	
Reading (SLS, RQ)	99.12[a]	14.11	65.1	134.9		98.03[b]	14.43	65.1	130.81	
Arithmetic (DEMAT, *T*)	50.88[a]	8.87	28.0	71.0		49.07[c]	9.34	27.0	73.0	

Note. CFT = Culture Fair Intelligence Test; IQ = intelligence quotient; SLS = Salzburger Lese-Screening [Salzburg reading screening test]; RQ = reading quotient; DEMAT = Deutscher Mathematiktest [German mathematics test].
The sample consists of the waiting control group of a larger response to intervention study. The four schools participating in this group are referred to as "A", "B", "C" and "D". IQ and RQ values approximately follow an $N(100, 15)$ distribution, T values an $N(50, 10)$ distribution. Some statistics have a lower sample size due to missing values: [a]: n = 107; [b]: n = 108; [c]: n = 106.

2.2 Progress-Monitoring Instrument LVD-M 2-4

The progress-monitoring instrument "Learning-progress diagnostics – mathematics for grades two to four [German original: Lernverlaufsdiagnostik – Mathematik für zweite bis vierte Klassen]" (LVD-M 2-4; Strathmann & Klauer, 2012) is applicable in the respective grades of elementary school. Each LVD-M 2-4 booklet contains 24 tasks that cover the grade's arithmetic (addition, subtraction, multiplication, division) part of the curriculum – in a way that it applies to all (or at least most) German states with their federal curricula.

While Grade 3 booklets (Table 2) consist of 19 mental (addition, subtraction, multiplication, division) and five written arithmetical problems (addition and subtraction), Grade 4 booklets (Table 3) mainly contain written tasks (18 items of all four basic arithmetic operations) and only six mental addition and subtraction items at the beginning. Beyond the solution mode (mental or written) and the basic arithmetic operation, the items also differ with regard to the task structure (Do the test takers have to calculate the result of an equation, like in 926 + 53 = ?, or fill in

a blank, like in 874 + ? = 900), number range (place value structure), and the need to perform carry operations (e.g., tens-carry).

Table 2: Task Structure of the LVD-M 2-4 Booklets, Grade 3

Item(s)	Task structure	Place value structure/arithmetic operator	Mode	Example
1–2	a + b = ?	HTO + TO (with T+T ≤ 100)	m	926 + 53 = ?
3	c – b = ?	HTO – O (with ? > H, first O < second O)	m	982 – 3 = ?
4	c – b = ?	HTO – HTO	m	856 – 117 = ?
5	a + b = ?	HTO + TO (with ? ≤ 1000, TO + TO ≤ 100)	m	542 + 16 = ?
6	a + ? = H00*	HTO + TO	m	874 + ? = 900
7	c – ? = a	HTO – TO	m	967 – ? = 952
8, 9, 11	a · b = ?	times 6; times 9; times 50	m	4 · 6 = ?; ..., 7 · 50 = ?
10	a · ? = c	times 7	m	9 · ? = 63
12, 13	c = ? · b	times 8; times 20	m	24 = 8 · ?; 80 = 20 · ?
14–19	c : b = ?	divided by 6; ...by 9; ... by 8; ... by 7; ... by 20; ... by 50	m	24 : 6 = ?, ..., 200 : 50 = ?
20	a + b + c = ?	TO + TO + TO (with O + O + O > T, T + T + T > H)	w	15 + 95 + 39 =
21	a + b =?	HTO + HTO (with O + O > T, T + T < H, H + H < Th)	w	338 + 336 = ?
22	c – b = ?	HTO – HTO	w	876 – 741 = ?
23	c – b = ?	H00 – HTO	w	700 – 168 = ?
24	c – b = ?	1000 – HTO	w	1000– 439 = ?

Note. Th = thousand(s); H = hundred(s); T = ten(s); O = ones; m = mental; w = written.

*In the test manual the task structure is described as a + ? = H, we here chose "H00" instead of "H" to make clear that "full" hundreds are meant, like in Item 23.

Table 3: Task Structure of the LVD-M 2-4 Booklets, Grade 4

Item(s)	Task structure	Place value structure/arithmetic operator	Mode	Example
1	a + b = ?	ThHTO + TO (with ? < 10.000)	m	2557 + 43 = ?
2, 3	c – b = ?	10.000 – ThH > 0; ThHTO – ThHT > 0	m	10000 – 6600 = ? 6728 – 4670 = ?
4	a + ? = c	ThHTO + HT (with ? < 10.000)	m	8243 + ? = 9023
5	a + ? = 10.000	ThHTO + (Th)HTO = 10.000	m	5862 + ? = 10000
6	c – ? = a	ThHTO – ThH > 0	m	7536 – ? = 4936
7, 8	a + b = ?	ThHTO* + ThHTO (with carry over once); TThThHTO* + TThThHTO = X0 000 (with X < 10)	w	4157 + 2839 = ? 30818 + 19182 = ?
9	a + b + c = ?	ThHTO** + ThHTO + ThHTO (with carry over once or twice)	w	1092 + 3261 + 2516 =?
10–12	c – b = ?	ThHTO** – ThHTO (with no carry over); TThThHTO** – TThThHTO (with carry over once or twice); 1 000 000 – HThTThThHTO* > 0	w	6898 – 5267 = ?, ..., 1000000 – 403182 = ?
13–18	a · b = ?	T0 · O; TO · O; HTO* · O; TO · T; HTO* · T; HTO · TO	w	30 · 9 = ?, ..., 275 · 60 = ?
19–24	c : b = ?	HTO : O; ThHTO : O; TThThHTO* : O; ThHT* : T0; TThTHT : T; HThTThThHTO* : T	w	355 : 5 = ?, ..., 404580 : 60 = ?

Note. HTh = hundred thousand(s); TTh = ten thousand(s); Th = thousand(s); H = hundred(s); T = ten(s); O = ones; m = mental; w = written.

*One (but the first) place may be zero. **One or two (but the first) place(s) may be zero.

Individual booklets are generated by means of stratified item sampling, that is, each booklet follows the same task and place value structure as well as order of the 24 item families. The specific representation within each of the item families is randomly drawn, assisted by software. This means that the term "item family" refers to the structural pattern of a task (as given in the lines of Table 2). For example, Item Family 3 in Grade 3 booklets (c – b = ?, HTO – O (with ? > H, first O < second O)), could likewise be represented by the random representations "432 – 6 = ?" or "824 – 5 = ?".

LVD-M 2-4 is designed as a power test with no specific time limit. Administration in classroom settings takes approximately 15 to 20 minutes. Based on the norm sample, the test authors report split-half reliabilities of .87 (Grade 3) and .79 (Grade 4) for measurements halfway through and .81 (Grade 3) and .83 (Grade 4) at the end of the school year (Strathmann & Klauer, 2012). In terms of criterion validity, LVD-M 2-4 scores substantially correlate with results on an established curriculum-based test (DEMAT; correlations between .53 in Grade 3 and .80 in Grade 4) and math grades (correlations between –.54 and –.77; Strathmann & Klauer, 2012, p. 32).

Factor analyses of the norm sample data show that, when the 24 items are transformed to four subscores, one for each basic arithmetic operation, a strong general factor explains between 51% and 78% of the variance in Grade 3 and 4 results (Strathmann & Klauer, 2012). Therefore, to evaluate the test, the number of correctly solved items per student across the whole booklet are counted. This means that the test score x (a) aggregates the different tasks to a general "competence to perform basic arithmetic operations" (Strathmann & Klauer, 2012, p. 31) and (b) does not distinguish between incorrectly solved and unsolved items (both are scored with 0). Based on the binomial model, the ability p of a student to deal with the content represented by the test is estimated as $p = x/n$, with n denoting the number of items and x the number of correctly solved items. While n is a constant, x is a binomially distributed random variable with the variance $s^2 = n \cdot p \cdot (1-p)$, so students with either very low or very high ps have more precise ability estimates.

2.3 Design

Starting in January and ending in early June 2016 (second half-year), LVD-M 2-4 was conducted bi-weekly, adding up to ten measurements. In all participating classes, the first assessment in January was led by a student assistant of the research team who advised each math teacher to administer the consecutive nine assessments. Teachers received an information brochure with background information on LVD-M 2-4, the schedule, as well as a detailed, standardized instruction sheet. Every other week, teachers were provided with the set of tests to be completed by their class in the following assessment week (the specific assessment day was up to teachers) along with a reply envelope.

All students went through ten different booklets. The booklets (A–H) were presented in ten different orders (1–10) based on a Latin square design (Table 4), and each student was randomly assigned one of the presentation orders. To ensure that data strictly followed this design, booklets were labelled with students' names and teachers were asked to return them directly after completion. Students participated on average on 9.35 of ten measurements (SD = 0.96), 54.63% of the participants in Grade 3 and 57.8% in Grade 4 completed all ten booklets.

After five measurements, that is, in spring 2016, all teachers were provided with intermediate feedback on (a) their respective class in comparison to other classes participating in the project as well as for (b) each student compared to the specific classroom. For both levels, class and students, the development was rated as *positive trend*, *negative trend*, or *constant*. In order to classify an individual student trajectory, at least four data points had to be present, and Cohen's f^2 in a linear regression model (dependent variable: LVD-M 2-4 score, independent variable: measurement time) had to correspond at least to a small effect size (Cohen's $f^2 \geq .02$).

Table 4: Measurement Scheme

	Measurements 1–10 / Booklet versions A–J										
Order	1 Jan	2 Jan	3 Feb	4 Feb	5 Mar	Easter break	6 Apr	7 Apr	8 May	9 May	10 Jun
1	A	B	C	D	E		F	G	H	I	J
2	B	C	D	E	F		G	H	I	J	A
3	C	D	E	F	G	Feedback	H	I	J	A	B
...											
9	I	J	A	B	C		D	E	F	G	H
10	J	A	B	C	D		E	F	G	H	I

2.4 Booklet Effects

2.4.1 Booklet Effects on the Total Score

The total score y_{ijkt} (number of correctly solved items out of 24) of child i in class j at time t completing booklet k was modeled by a fixed intercept and a fixed slope estimating the linear development across the ten bi-weekly measurements, as given in Model (1). This model and its extensions are practically relevant as the total score is the parameter used for test evaluation and interpretation.

Random effects were included to represent the following variance components: First, random intercepts on the level of (a) booklets (b_{0k}), (b) classes (b_{0j}), and (c) individuals ($b_{0i|j}$) were specified to capture differences in the (baseline) test score due to each of the three levels. Second, the variances of random slopes on the class (b_{1j}) and individual level ($b_{1i|j}$) capture differences in linear growth across time between the instances of each level. Correlations between random intercepts and slopes represent the relation between (baseline) level of achievement and development across time within classrooms or individuals. Based on restricted maximum likelihood estimation (REML), Model (1) was fitted separately for each of the two grades.

$$y_{ijkt} = \beta_0 + b_{0k} + b_{0j} + b_{0i|j} + (\beta_1 + b_{1j} + b_{1i|j}) \times time_t + s_{ijkt} \quad (1)$$

2.4.2 Item (Family) Effects on Solution Probability

By means of four successive models, the equivalence of randomly sampled booklets was not only assessed for the total score as in (1) but also on the item level. All four models are hierarchical logistic regressions (i.e., a multilevel generalized mixed model with logit link), with the logit of a correct solution denoted as $\eta = log\left(\frac{p(Y=1)}{1-p(Y=1)}\right)$.

First, a baseline Model (2) ignoring possible booklet effects served as a benchmark. This model includes the average item difficulty (fixed intercept β_0), the fixed slope, as well as the random intercept (b_{0i}) and the random slope per child i across measurements (b_{1i}). Pairs of random effects (b_{0i}, b_{1i}) are assumed to follow a bivariate normal distribution, that is, their correlations are not restricted. This model is conceptually closest to the binomial model.

$$\eta_{it} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) \times time_t \quad (2)$$

Second, in a consecutive Model (3), the outcome η_{itc} was predicted by a fixed (β_1) and child-specific random slope (b_{1i}) representing the students' development over the ten measurements points. The fixed effect δ_c, with c indicating the number of

the respective family, expresses the average difficulty of each of the 24 item families and replaces the fixed intercept β_0 of Model (2).

$$\eta_{itc} = \delta_c + b_{0i} + (\beta_1 + b_{1i}) \times time_t \quad (3)$$

Third, to examine the variability of item family effects, random intercepts b_{0l} for the 240 different items l (240 = 24 item families times 10 booklets) were added to Model (3). In this Model (4), the fixed effect $\delta_{c(l)}$ treats all ten randomly drawn members of a specific item family as if they were different items. The item-level random intercept b_{0l} was assumed to be independent of all other random effects and to be $N(0,\omega_2)$-distributed. In Model (4), the variance parameter ω_2 expresses a "global variance" which is independent of the item families.

$$\eta_{ilt} = \delta_{c(l)} + b_{0l} + b_{0i} + (\beta_1 + b_{1i}) \times time_t \quad (4)$$

Fourth, to investigate even more fine-grained item effects, Model (4) was extended by a random intercept for each item clustered by booklets ($b_{0l,c(l)}$, again normally distributed). In other words, Model (5) assesses to what extent items within one family (task type) are homogeneous or heterogeneous across the ten randomly generated booklets. Thus, in Model (5), the item variance depends on the item families.

$$\eta_{ilt} = \delta_{c(l)} + b_{0i} + b_{0l,c(l)} + (\beta_1 + b_{1i}) \times time_t \quad (5)$$

3. Results

3.1 Booklet Invariance in a Sample Receiving Regular Instruction

3.1.1 Booklet Effects on the Total Score

In Grade 3 (n = 108), the variance of random intercepts (level effects) between children (var[$b_{0i|j}$] = 25.722, Table 5) was over 100 times larger than the variance between booklets (k = 10, var[b_{0k}] = 0.245, Table 5). Moreover, also the class-level effect (var[b_{0j}] = 0.163, Table 5) measured only a small proportion (0.6%) of the variance on the individual level. On the class level, random intercepts (b_{0j}) and slopes (b_{1j}) correlated positively (r = .45), meaning that higher class-level baseline scores tended to go along with steeper slopes (so-called Matthew effect). On the individual level, the correlation between random intercepts ($b_{0i|j}$) and slopes ($b_{1i|j}$) was negative, that is, individuals with a higher (baseline) achievement level showed a flatter progress (r = –.61).

In Grade 4 (n = 109), the effects reported for Grade 3 replicated in direction but varied in size: The variance of random intercepts between individuals was almost 50 times larger than between booklets. While the booklet-level variance of random intercepts was similar in both grades (var[b_{0k}] = 0.25 in Grade 3 vs. 0.28 in Grade 4), individual-level variance was lower in Grade 4 (var[$b_{0i|j}$] = 13.80 vs. 25.72), meaning that individual baseline levels were more heterogeneous in Grade 3. Again, the class-level variance of random intercepts only measured a small proportion (4.2%) of the individual-level variance. The class-level Matthew effect was more pronounced in the Grade 4 sample based on mainly written arithmetic booklets (r = .55) compared to Grade 3 where booklets contain mainly mental arithmetic tasks (r = .45, see above). These effects went along with a lower mean baseline (fixed intercept) in Grade 4 (9.31) compared to Grade 3 (15.39). On the individual level, the correlation between random intercepts and slopes was negative but smaller (r = −.43).

Table 5: Variances and Correlations of Random Effects in Model (1)

Level	Parameter	Grade 3 (n = 108)	Grade 4 (n = 109)
	Fixed effects		
	β_0	15.38 (0.54)	9.31 (0.47)
	β_1	0.18 (0.15)	0.27 (0.12)
	Random effects		
Booklet	var(b_{0k})	0.25	0.28
Class	var(b_{0j})	0.16	0.59
Class	var(b_{1j})	0.20	0.11
Class	cor(b_{0j}, b_{1j})	.45	.55
Individual	var($b_{0i|j}$)	25.72	13.80
Individual	var($b_{1i|j}$)	0.22	0.23
Individual	cor($b_{0i|j}$, $b_{1i|j}$)	−.61	−.43
Individual	var(Residual)	18.67	9.41

Note. The standard error of estimation (fixed effects) is indicated in parentheses.

3.1.2 Item (Family) Effects on Solution Probability

Table 6 displays an overview of the four models expressing item family effects: In both grades, Model (3) containing an item-family-specific fixed effect δ_c showed a better fit than the baseline Model (2), indicating that difficulty estimates vary visibly between item families (Grade 3: $\chi^2[23] = 2555.8$, $p < .001$, Grade 4: $\chi^2[23] = 2743.4$, $p < .001$). In Grade 3, Item 9 (times 9, cf. Table 2) was the easiest ($\delta_c = 2.79$, Table 7) while Item 24 (written subtraction: 1000 − HTO) was the hardest ($\delta_c = -0.96$). Transformed to the more intuitive level of solution probability this

means, for example that halfway through the measurement period (measurement $time_t = 5$) in Grade 3 items with δ_c below –0.65 had a probability of correct solution of less than 50%, items with δ_c above –.65 a solution probability of more than 50%, and items with $\delta_c > .23$ a solution probability of over 95%. In Grade 4, Item 13 (T · O) was the easiest ($\delta_c = 2.27$, Table 7) and Item 24 (HThTThThHTO* : T) the hardest ($\delta_c = -1.81$, Table 7).

The main result contributed by Model (4) is that the variance of the random intercepts clustered by item families, that is, task types, was only one eighth (Grade 3: $var[b_{0l}] = 0.09$, Table 8) or one tenth (Grade 4: $var[b_{0l}] = 0.12$, Table 8) of the variance related to individual subjects (Grade 3: $var[b_{0i}] = 0.72$, Grade 4: $var[b_{0i}] = 1.18$, Table 8). Although the additional variance component significantly improved model fit (Grade 3: $\chi^2[1] = 61.30$, $p < .001$, Grade 4: $\chi^2[23] = 137.49$, $p < .001$), the effect was relatively small (cf. AIC in Table 6). This means that general level differences in achievement are far more heterogeneous between students than between items.

In both grades, difficulty effects (random intercepts) of the 24 items representing 24 tasks differed in the extent to which they varied across booklets (comparison of Models [4] and [5]: Grade 3: $\chi^2[23] = 46.26$, $p < .01$; Grade 4: $\chi^2[23] = 70.47$, $p < .001$). While some items produced homogeneous difficulties, that is, variances estimated close to zero, across the ten booklets (Grade 3: Items 3, 8, 13, 14, 21, 23, and 24; Grade 4: Items 5, 9, 11, 19, 20, 21, and 22; Table 9), others varied more strongly (Grade 3: e.g., Items 7, 9, 10, 11, 17, and 22; Grade 4: Items 1, 3, 4, 6; Table 9). In Grade 3, the most homogeneous items comprised all four arithmetic operations, four of them were mental, three written tasks. The most heterogeneous items were mainly mental and multiplication or division tasks. In Grade 4, the most homogeneous items were all but one (Item 5) written tasks, requiring addition, subtraction, and division. The four most heterogeneous ones were mental tasks. While in Grade 4 the most difficulty-heterogeneous items could be found in the initial part of the booklet (first quarter), in Grade 3 difficulty-variant items were spread across the entire booklet.

The correlation between average difficulty and cross-booklet heterogeneity (in difficulty) of items, $r(\delta_{c(l)}, var[b_{0l,c(l)}]$, was $r = .29$ ($p = .16$) in Grade 3 and $r = .06$ ($p = .77$) in Grade 4.

Table 6: Comparison of Models Expressing Item (Family) Effects on Solution Probability

Model	Formula	Description	AIC (p model comparison)	
			Grade 3 (n = 108)	Grade 4 (n = 109)
(2)	$\eta_{it} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) \times time_t$	Baseline model ignoring possible booklet effects (benchmark).	18 208	20 904
(3)	$\eta_{itc} = \delta_c + b_{0i} + (\beta_1 + b_{1i}) \times time_t$	The fixed effect δ_c expresses the average difficulty of each of the 24 item families.	15 698 (< .001)	18 206 (< .001)
(4)	$\eta_{ilt} = \delta_{c(l)} + b_{0l} + b_{0i} + (\beta_1 + b_{1i}) \times time_t$	Random intercepts b_{0l} for the 240 different items l (240 = 24 item families times 10 booklets) express item family effects.	15 638 (< .001)	18 071 (< .001)
(5)	$\eta_{ilt} = \delta_{c(l)} + b_{0i} + b_{0l,c(l)} + (\beta_1 + b_{1i}) \times time_t$	The random intercept $b_{0l,c(l)}$ for each item clustered by booklets assesses to what extent items within one family (task type) are homogeneous or heterogeneous across the ten randomly generated booklets.	15 638 (< .01)	18 047 (< .001)

Note. Meaning of the model parameters in the order of appearance: η = logit of a correct solution (link level); β_0 = average item difficulty (fixed intercept); b_{0i} = child-specific random intercept; b_{1i} = child-specific random slope; β_1 = fixed slope, $time_t$ = measurement time (1–10); δ_c = item-specific fixed effect (average difficulty of each of the 24 item families); $\delta_{c(l)}$ = item-specific fixed effect (all 10 randomly drawn representations of a family are treated as different items); b_{0l} = item-specific random intercepts (for all 240 different items); $b_{0l,c(l)}$ = random intercept for each item clustered by booklets indices; i = individual; t = time; c = item family; l = booklet; the value in parentheses is the p value of the model comparison (χ^2 difference test) of the given model to the model before, for example of Model (3) compared to Model (2).

Table 7: Fixed and Random Effects of Model (3), Logit Link

Effect (level)	Parameter	Grade 3 (n = 108)	Grade 4 (n = 109)
Fixed slope	β_1	0.13 (0.01)	0.08 (0.01)
Fixed (item)	δ_{Item1}	1.73 (0.14)	0.95 (0.14)
Fixed (item)	δ_{Item2}	1.96 (0.14)	0.08 (0.13)
Fixed (item)	δ_{Item3}	1.24 (0.13)	−0.25 (0.13)
Fixed (item)	δ_{Item4}	0.59 (0.12)	−0.91 (0.14)
Fixed (item)	δ_{Item5}	0.95 (0.13)	−0.42 (0.13)
Fixed (item)	δ_{Item6}	0.95 (0.13)	0.18 (0.14)
Fixed (item)	δ_{Item7}	0.31 (0.13)	1.97 (0.15)
Fixed (item)	δ_{Item8}	2.71 (0.18)	1.85 (0.15)
Fixed (item)	δ_{Item9}	2.79 (0.19)	1.52 (0.14)
Fixed (item)	δ_{Item10}	2.20 (0.16)	1.72 (0.15)
Fixed (item)	δ_{Item11}	1.78 (0.15)	−0.08 (0.13)
Fixed (item)	δ_{Item12}	1.84 (0.15)	−0.56 (0.13)
Fixed (item)	δ_{Item13}	1.16 (0.13)	2.27 (0.16)
Fixed (item)	δ_{Item14}	2.25 (0.16)	0.65 (0.13)
Fixed (item)	δ_{Item15}	2.11 (0.16)	0.63 (0.14)
Fixed (item)	δ_{Item16}	1.81 (0.15)	0.25 (0.13)
Fixed (item)	δ_{Item17}	2.25 (0.16)	−0.11 (0.13)
Fixed (item)	δ_{Item18}	1.21 (0.14)	−0.91 (0.14)
Fixed (item)	δ_{Item19}	1.30 (0.13)	0.31 (0.15)
Fixed (item)	δ_{Item20}	0.17 (0.12)	−0.14 (0.15)
Fixed (item)	δ_{Item21}	1.12 (0.13)	−0.24 (0.16)
Fixed (item)	δ_{Item22}	0.05 (0.12)	−1.39 (0.18)
Fixed (item)	δ_{Item23}	-0.96 (0.12)	−1.74 (0.21)
Fixed (item)	δ_{Item24}	-0.96 (0.12)	−1.81 (0.23)
Random intercept (individual)	var(b_{0i})	0.69	1.14
Random slope (individual)	var(b_{1i})	0.01	0.01
Correlation of random intercept and slope (individual)	cor(b_{0i}, b_{1i})	.22	−.23

Note. The standard error of estimation (fixed effects) is indicated in parentheses; the structure of Items 1–24 is detailed in Tables 2 and 3.

Table 8: Random Effects of Model (4)

Level	Parameter	Grade 3 (n = 108)	Grade 4 (n = 109)
Item	var(b_{0l})	0.094	0.122
Individual	var(b_{0i})	0.723	1.176
Individual	var(b_{1i})	0.008	0.005
Individual	cor(b_{0i}, b_{1i})	.21	−.22

Table 9: Variances of Model (5)

	Variance of the random intercept for each item clustered by booklets: var($b_{0l,c(l)}$)	
Item family c	Grade 3	Grade 4
1	0.16	0.41
2	0.06	0.04
3	0.00	0.46
4	0.18	0.21
5	0.06	0.00
6	0.02	0.84
7	0.25	0.03
8	0.00	0.03
9	0.37	0.00
10	0.43	0.12
11	0.23	0.00
12	0.19	0.01
13	0.00	0.14
14	0.00	0.07
15	0.13	0.09
16	0.09	0.11
17	0.25	0.08
18	0.01	0.11
19	0.04	0.00
20	0.08	0.00
21	0.00	0.00
22	0.31	0.00
23	0.00	0.01
24	0.00	0.06
Child-level	Grade 3	Grade 4
b_{0i}	0.73	1.18

Note. The structure of Items 1–24 is detailed in Tables 2 and 3.

3.2 Replication Studies in Two Intervention Groups

So far, this paper concentrated on a group of 108 third and 109 fourth graders who received regular instruction during a 19 weeks progress-monitoring phase. This waiting control group of a larger RTI evaluation study represents classroom practice without systematic intervention.

To test and discuss the generalizability of the findings on booklet equivalence, the same analyses were also carried out with the two intervention groups of the same study. In the first of these groups (Intervention Group 1), children with arithmetic performance at risk level or below (PR ≤ 25) received individual computer-based training with a dyscalculia training app throughout the 19 weeks intervention phase. In the second group (Intervention Group 2), a three-tiered scheme was realized: Children in the risk range (25 ≤ PR ≤ 10) worked with the training app, while children with arithmetic performance below PR 10 received the training app and an additional weekly small group in-person training. Taken together, the two intervention groups represent a population of children receiving needs-based intervention beyond regular classroom instruction: In both intervention groups and grades, the ratio of between-children variance (var[$b_{0i|j}$]) and between-booklets variance (var[b_{0k}]) was even more pronounced than in the waiting control group: In Intervention Group 1 (training app), the between-children variance was 229 times (Grade 3) or 252 times (Grade 4) larger than the between-booklet variance (Table A1). In Intervention Group 2 (training app + small group intervention), this ratio was 362 (Grade 3) and 53 (Grade 4), respectively (Table A1).

Beyond booklet equivalence, results of the intervention and waiting control groups differed in the following ways: First, class-level variance of the random intercepts (var[b_{0j}]) was larger in the intervention groups compared to the waiting control group (Table A1), pointing to more heterogeneity between classes in the intervention samples. Second, the correlational patterns of random effects were different in the intervention groups: While random intercepts and slopes on class level were positively correlated in the waiting control group (cor[b_{0j}, b_{1j}] = .45 and .55, Table 5), the opposite was the case for Grade 3 students in both intervention groups (cor[b_{0j}, b_{1j}] = −.75 and −.74, Table A1), and Grade 4 students in Intervention Group 2 (cor[b_{0j}, b_{1j}] = −.57, Table A1). Item-level variances were even smaller than in the waiting control group, with a maximum of 0.08 (Item 16, Grade 3) and 0.05 (Item 10, Grade 4) in the first intervention group and 0.20 (Item 9, Grade 3) and 0.08 (Item 22, Grade 4) in the second intervention group (Table A3).

4. Discussion

4.1 Summary and Relevance

This study is based on an established German math CBM instrument which was applied in a third- and fourth-grade sample receiving regular instruction during a school half-year with ten bi-weekly measurements. The instrument, LVD-M 2-4 (Strathmann & Klauer, 2012), is broadly relevant, since it is easily applicable in classroom contexts and valid for the whole achievement range within the German elementary school curriculum.

Against this background, the present study tested booklet effects to examine to what extent the theoretically justified item-sampling approach of LVD-M 2-4 is empirically backed up. Analyses with linear mixed models showed that in both grades the variance of test scores attributable to different booklets was so small in relation to the between-student variance to be practically irrelevant. This is an important prerequisite for the practical interpretation of repeatedly measured (longitudinal) test scores valid measures of individual learning. Generalized linear mixed models revealed that item difficulty varied within some of the item families, while others showed close to zero variance across the ten booklets. Although, on a descriptive level, heterogeneity tended to correlate with item difficulty in Grade 3 and was more pronounced for mental compared to written tasks in Grade 4, no clear pattern explaining this source of variance emerged. Importantly, the effect was insubstantial in comparison with the effects due to individual differences between students. To learn more about the inconclusive pattern of differences in across-booklet heterogeneity (of difficulties) for the different item families (Table 9), quasi-experimental studies could help to unravel content (task type, arithmetic operation, etc.) from sequence effects (position, order).

Taken together, it seems legitimate to assume that the $t = 10$ measurements of LVD-M 2-4 overall express the same ability. This finding is relevant for the *conceptualization perspective* on progress-monitoring instruments and the first stage of CBM research described by Fuchs (2004): Like in previous research (Christ & Vining, 2006), it corroborates that stratified random item-sampling within predefined competences (e.g., written or mental arithmetic, with a certain task and place value structure; cf. Table 2 and Table 3) leads to comparable parallel tests, which is a key element of generic test construction (Rohwer, 2015).

The two dyscalculia intervention groups of the larger RTI study differed from the regular-instruction group in some respect, for example, showing a reversed Matthew effect on the class level (weaker start level along with higher improvement), which might be to some degree explainable by a compensation effect caused by the intervention framework. However, the main finding of this study, that is, a relatively low proportion of test score variance accounted for by different booklets compared to different individuals, clearly generalized to two intervention samples. In other words, equivalence of LVD-M 2-4 booklets can also be assumed for RTI

settings with in-depth fostering of students with special educational needs in basic arithmetic.

This is an important precondition for further research aims, for example, assessing sensitivity to change or deriving norms for the learning slope. Such follow-up goals are only reasonable if they build on solid grounds in terms of booklet equivalence.

4.2 Limitations and Further Directions

The results of this study (negligible booklet effects) are limited by its context: the methodological approach, the instrument, and possibly also the population underlying the regional sample in this study.

Regarding the methodological approach, in the current study, we did not proceed by calibrating all items a priori and then investigating DIF or measurement invariance. Rather, test booklets were randomly allocated to students, corresponding to the random groups design as outlined by Kolen and Brennan (2014). According to this approach, "If the same scaling convention [...] for ability is used [...], then the parameter estimates [...] are assumed to be on the same scale without further transformation" (p. 182). Hence, we assumed that all test scores and item parameters were on a comparable scale.

We utilized a statistical approach similar to generalizability theory (Brennan, 2001) in which we compared the magnitude of variance components (cf. Fan & Hansmann, 2015) to establish empirical evidence of test equivalence. This approach does not focus on the equivalence of single item parameters (e.g., classical DIF or measurement invariance approaches). Instead, the approach chosen here deals with the variance of relevant parameters or factors (e.g., item parameters, test scores). In the case of test equivalence, the variance component pertaining to booklets or items should be very small and negligible, compared to factors that should substantially contribute to total variance (e.g., differences in person ability, time points). In fact, the complete absence of item (family)-level and booklet-level variance would imply that parameters of items in the same position are stable between booklets. The tests would then have a strong form of measurement equivalence, in the sense that there was neither DIF nor DTF and the test would parallel in the sense of CTT.

Although the results provided here lend support to the assumption that test forms were practically equivalent, we recognize the limitations of the chosen statistical approach. Specifically, by focusing on variance components, we did not look at single item parameters, which would be helpful in identifying those items whose presence substantially enlarges the family-wise variance component. Further, it remains difficult to provide a suitable effect size of test or item equivalence in this context. At what (relative) test form variance component magnitude can test equivalence be assumed, and when does it not hold? Does test form equivalence obtained using an approach comparing variance components also hold un-

der a DIF or measurement invariance approach? These questions necessitate further research.

Regarding the instrument (LVD-M 2-4) and its construction rationale, we investigated the equivalence of math progress-monitoring tests that were constructed using a systematic rule-based item design approach. In rule-based item design, item features that should theoretically affect item difficulty are identified (Enright et al., 2002). Once these essential item features are known, design of equivalent test forms becomes more straightforward.

However, available progress-monitoring tests differ in how thoroughly essential item features were identified and/or taken into consideration in test design. Some progress-monitoring measures are based on comprehensive and detailed, rule-based item design (e.g., Förster & Kuhn, 2021; Klauer & Strathmann, 2012), providing evidence for a high degree of test equivalence. Other progress-monitoring measures are less explicit concerning the rationale for test and item design. The results obtained in this study, therefore, should not be seen as representative for progress-monitoring measures in general. Rather, they should be regarded as preliminary evidence of equivalence of a progress-monitoring measure that was constructed using rule-based item design.

Moreover, to what extent the findings can be generalized to broader content domains (like in *quop*, which includes precurricular basic numerical abilities in early-grade version and aspects like geometry or basic statistics in later grades; Souvignier, 2018) or different administration modes (computer-based instead of paper-pencil) remain empirical questions for subsequent studies.

Acknowledgements

This research was funded by the German Federal Ministry of Education and Research (grant 01 GJ 1302).

References

Ardoin, S. P., & Christ, T. J. (2009). Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set. *School Psychology Review, 38*(2), 266–283. https://doi.org/10.1080/02796015.2009.12087837

Ardoin, S. P., Christ, T. J., Morena, L. S., Cormier, D. C., & Klingbeil, D. A. (2013). A systematic review and summarization of the recommendations and research surrounding Curriculum-Based Measurement of oral reading fluency (CBM-R) decision rules. *Journal of School Psychology, 51*(1), 1–18. https://doi.org/10.1016/j.jsp.2012.09.004

Ashlock, R. B. (2005). *Error patterns in computation: Using error patterns to improve instruction* (9th ed.). Prentice Hall

Björn, P. M., Aro, M., Koponen, T., Fuchs, L. S., & Fuchs, D. (2018). Response-to-intervention in Finland and the United States: Mathematics learning support as an example. *Frontiers in Psychology, 9*, Article 800. https://doi.org/10.3389/fpsyg.2018.00800

Brennan, R. L. (2001). *Generalizability theory*. Springer. https://doi.org/10.1007/978-1-4757-3456-0

Christ, T. J., & Vining, O. (2006). Curriculum-based measurement procedures to develop multiple-skill mathematics computation probes: Evaluation of random and stratified stimulus-set arrangements. *School Psychology Review, 35*(3), 387–400. https://doi.org/10.1080/02796015.2006.12087974

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219–232. https://doi.org/10.1177/001440298505200303

Deno, S. L. (2003). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention, 28*(3–4), 3–12. https://doi.org/10.1177/073724770302800302

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education, 15*(1), 49–74. https://doi.org/10.1207/S15324818AME1501_04

Fan, C.-H., & Hansmann, P. R. (2015). Applying generalizability theory for making quantitative RTI progress-monitoring decisions. *Assessment for Effective Intervention, 40*(4), 205–215. https://doi.org/10.1177/1534508415573299

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics. *The Journal of Special Education, 41*(2), 121–139. https://doi.org/10.1177/0022466907041002010 1

Förster, N., & Kuhn, J.-T. (2021). Ice is hot and water is dry – Developing equivalent reading tests using rule-based item design. *European Journal of Psychological Assessment*. Advance online publication. https://doi.org/10.1027/1015-5759/a000691

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*(3), 315–342. https://doi.org/10.1016/j.jsp.2007.06.003

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188–192. https://doi.org/10.1080/02796015.2004.12086241

Gerster, H.-D. (2012). *Schülerfehler bei schriftlichen Rechenverfahren: Diagnose und Therapie* [Student errors in written arithmetic procedures: Diagnosis and therapy]. WTM. https://nbn-resolving.de/urn:nbn:de:hbz:6-79209639973

Hebbecker, K., & Souvignier, E. (2016). Lernverlaufsdiagnostik zur Unterstützung individueller (Begabungs-)Förderung: Internetbasierte Lernverlaufsdiagnostik mit dem System „quop" [Learning progress diagnostics to support individual (gifted) support: Internet-based learning progress diagnostics with the system "quop"]. *Journal für Begabtenförderung, 16*(1), 29–38.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABC's of CBM: A practical guide to curriculum-based measurement* (2nd ed.). Guilford.

Huber, C., & Grosche, M. (2012). Das Response-to-Intervention-Modell als Grundlage für einen inklusiven Paradigmenwechsel in der Sonderpädagogik [The response to intervention model as a basis for an inclusive paradigm shift in special education]. *Zeitschrift für Heilpädagogik, 63*(8), 312–322.

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review, 37*(3), 374–390. https://doi.org/10.1080/02796015.2008.12087884

Klauer, K. J. (2011). Lernverlaufsdiagnostik – Konzept, Schwierigkeiten und Möglichkeiten [Diagnosing the course of learning – Concept, difficulties and chances]. *Empirische Sonderpädagogik, 3*(3), 207–224.

Klauer, K. J. (2014). Formative Leistungsdiagnostik: Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik [Formative achievement assessment: Historical background and evolution into progress monitoring]. In M. Hasselhorn, W. Schneider, & U. Trautwein (Eds.), *Tests und Trends: Vol. 12: Lernverlaufsdiagnostik* (pp. 1–18). Hogrefe.

Klauer, K. J., & Strathmann, A. M. (2013). Lernverlaufsdiagnostik Mathematik: Test auf Änderungssensibilität bei rechenschwachen Grundschülern [Assessing growth in mathematical achievement: A test of sensitivity to change with underperforming elementary students]. *Psychologie in Erziehung und Unterricht, 60*(4), 241–252. https://doi.org/10.2378/peu2013.art18d

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7

Köller, O., & Reiss, K. (2013). Mathematische Kompetenz messen: Gibt es Unterschiede zwischen standardbasierten Verfahren und diagnostischen Tests? [Assessing mathematics competence: Are there differences between standard-based methods and diagnostic tests?] In M. Hasselhorn, A. Heinze, W. Schneider, & U. Trautwein (Eds.), *Diagnostik mathematischer Kompetenzen* (pp. 25–40). Hogrefe.

Krajewski, K., Liehm, S., & Schneider, W. (2004). *Deutscher Mathematiktest für zweite Klassen (DEMAT 2+)* [German mathematics test for second grades]. Hogrefe.

Kuhn, J.-T., & Holling, H. (2014). Number sense or working memory? The effect of two computer-based trainings on mathematical skills in elementary school. *Advances in Cognitive Psychology, 10*(2), 59–67. https://doi.org/10.5709/acp-0157-2

Kuhn, J.-T., & Schwenk, C. (2018). Onlinebasierte Diagnostik mathematischer Kompetenzen: Möglichkeiten und Grenzen [Online-based assessment of mathematical competencies: Possibilities and limitations]. *Lernen und Lernstörungen, 7*(4), 231–235. https://doi.org/10.1024/2235-0977/a000232

Lawson, M. J., & Rice, D. N. (1987). Thinking aloud: Analysing students' mathematics performance. *School Psychology International, 8*(4), 233–244. https://doi.org/10.1177/014303438700800404

McDonald, R. P. (1999). *Test theory: A unified treatment*. Erlbaum.

Montague, M., Penfield, R. D., Enders, C., & Huang, J. (2010). Curriculum-based measurement of math problem solving: A methodology and rationale for establishing equivalence of scores. *Journal of School Psychology, 48*(1), 39–52. https://doi.org/10.1016/j.jsp.2009.08.002

Rohwer, G. (2015). Bemerkungen zu einem Testverfahren für Lernfortschritte [Remarks on a test procedure for long-term learning progress]. *Journal for Educational Research Online, 7*(2), 147–156.

Roick, D., Görlitz, T., & Hasselhorn, M. (2004). *Deutscher Mathematiktest für dritte Klassen (DEMAT 3+)* [German mathematics test for third grades]. Hogrefe.

Schatschneider, C., Wagner, R. K., & Crawford, E. C. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning and Individual Differences, 18*(3), 308–315. https://doi.org/10.1016/j.lindif.2008.04.005

Schwenk, C., Kuhn, J.-T., Gühne, D., Doebler, P., & Holling, H. (2017). Auf Goldmünzenjagd: Psychometrische Kennwerte verschiedener Scoringansätze bei computergestützter Lernverlaufsdiagnostik im Bereich Mathematik [We are going on a gold coin hunt: Psychometric properties of different scorings in computer-based progress monitoring of mathematics ability]. *Empirische Sonderpädagogik, 9*(2), 123–142.

Shapiro, E. S., Dennis, M. S., & Fu, Q. (2015). Comparing computer adaptive and curriculum-based measures of math in progress monitoring. *School Psychology Quarterly, 30*(4), 470–487. https://doi.org/10.1037/spq0000116

Souvignier, E. (2018). Computerbasierte Lernverlaufsdiagnostik [Computer-based learning progress assessment]. *Lernen und Lernstörungen, 7*(4), 219–223. https://doi.org/10.1024/2235-0977/a000240

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*(3), 407–419. https://doi.org/10.1080/02796015.2001.12086123

Strathmann, A. M., & Klauer, K. J. (2010). Lernverlaufsdiagnostik: Ein Ansatz zur längerfristigen Lernfortschrittsmessung [Diagnosing the trajectory of learning: An approach to long term measuring of learning progress]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42*(2), 111–122. https://doi.org/10.1026/0049-8637/a000011

Strathmann, A. M., & Klauer, K. J. (2012). *LVD-M 2-4. Lernverlaufsdiagnostik – Mathematik für zweite bis vierte Klassen* [LVD-M 2-4: Learning-progress diagnostics – mathematics for grades two to four]. Hogrefe.

Voß, S. (2016). Rechengeschwindigkeit, -präzision oder -flüssigkeit? Zur Vorhersage und Förderung der Rechenleistungen von Erstklässlern [Computational speed, accuracy or fluency? The prediction and promotion of computational skills of first-grade students]. *Heilpädagogische Forschung, 42*(1), 13–24.

Weiß, R. H. (2006). *CFT 20-R – Grundintelligenztest Skala 2 mit Wortschatztest (WS) und Zahlenfolgentest (ZF)-Revision* [Culture Fair Intelligence Test, Scale 20-R]. Hogrefe.

Weiß, R. H., & Osterland, J. (2013). *CFT 1-R – Grundintelligenztest Skala 1 – Revision* [Culture Fair Intelligence Test, Scale 1]. Hogrefe.

Wimmer, H., & Mayringer, H. (2014). *Salzburger Lese-Screening für die Schulstufen 2–9* [Salzburg reading screening test for grades 2–9]. Hogrefe.

Appendix

Table A1: Variances and Correlations of Random Effects in Model (1) in the Two Intervention Groups

		Intervention group 1		Intervention group 2	
Level	Parameter	Grade 3 (n = 140)	Grade 4 (n = 118)	Grade 3 (n = 103)	Grade 4 (n = 125)
Fixed effects					
	β_0	14.52 (1.02)	8.44 (0.85)	13.18 (1.48)	8.23 (1.06)
	β_1	0.17 (0.16)	0.36 (0.14)	0.10 (0.21)	0.22 (0.12)
Random effects					
Booklet	var(b_{0k})	0.18	0.08	0.10	0.28
Class	var(b_{0j})	6.44	4.09	17.47	8.66
Class	var(b_{1j})	0.46	0.15	0.40	0.11
Class	cor(b_{0j}, b_{1j})	−.75	−.08	−.74	−.57
Individual	var($b_{0i\|j}$)	42.10	20.96	37.61	14.62
Individual	var($b_{1i\|j}$)	0.27	0.19	0.34	0.28
Individual	cor($b_{0i\|j}$, $b_{1i\|j}$)	−.45	.01	−.59	−.26
Individual	var(Residual)	13.89	8.85	23.26	15.17

Table A2: Random Effects of Model (4) in the Two Intervention Groups

		Intervention Group 1		Intervention Group 2	
Level	Parameter	Grade 3 (n = 129)	Grade 4 (n = 104)	Grade 3 (n = 103)	Grade 4 (n = 124)
Item	var(b_{0l})	0.00	0.00	0.00	0.00
Individual	var(b_{0i})	0.70	0.59	1.86	0.98
Individual	var(b_{1i})	0.01	0.00	0.09	0.00
Individual	cor(b_{0i}, b_{1i})	−.11	.20	−.51	−.10

Table A3: Variances of Model (5) in the Two Intervention Groups

	Variance of the random intercept for each item clustered by booklets: var($b_{0l,c(l)}$)			
	Intervention Group 1		Intervention Group 2	
Item family c	Grade 3	Grade 4	Grade 3	Grade 4
1	0.00	0.00	0.03	0.00
2	0.00	0.00	0.00	0.03
3	0.00	0.03	0.00	0.00
4	0.00	0.00	0.00	0.00
5	0.00	0.03	0.00	0.00
6	0.00	0.00	0.00	0.00
7	0.00	0.01	0.00	0.00
8	0.00	0.00	0.03	0.00
9	0.04	0.00	0.20	0.00
10	0.03	0.05	0.00	0.00
11	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.01
13	0.01	0.00	0.04	0.00
14	0.00	0.00	0.00	0.01
15	0.00	0.00	0.00	0.01
16	0.08	0.03	0.00	0.00
17	0.06	0.00	0.00	0.00
18	0.05	0.02	0.00	0.00
19	0.00	0.00	0.00	0.07
20	0.00	0.04	0.00	0.00
21	0.00	0.03	0.04	0.00
22	0.00	0.00	0.00	0.08
23	0.00	0.00	0.00	0.00
24	0.00	0.00	0.00	0.03
b_{0i}	0.70	0.59	1.87	0.978

Note. The structure of Items 1–24 is described in Table 2 and Table 3.

Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01.05
Volume 14 (2022), No. 1, 89–121
CC BY-NC-SA 4.0 Waxmann 2022

Gesa Brunn, Fritjof Freise, & Philipp Doebler

Modeling a Smooth Course of Learning and Testing Individual Deviations From a Global Course

Abstract

Formative assessment supplies valuable feedback for teachers and learners, and has been facilitated by computerized implementations. While longitudinal within-student assessment or within-class comparisons are useful, a normative interpretation of an individual's course of learning can only be given relative to a reference population. As current computerized assessment systems sample items from pools or adapt tests, monitored students might work on non-overlapping item sets, so that classic sum scores cannot be compared directly. To meet this challenge, the Smooth Growth and Linear Deviations Rasch Model (SGLDRM) is introduced, an extension of Rasch's item response theory model for binary test data. With the help of spline functions a smooth global course of learning is included. The model is flexible enough to accommodate increases and/or decreases of the mean ability level, which might be more or less pronounced at each measurement occasion. On the individual level, a random slope and a random intercept with amenable interpretations modify the global course of learning. Two measurement occasions suffice to estimate person-specific courses. A likelihood ratio test allows identifying students whose performance differs from the mean course. The methodology is illustrated with data from an online dyscalculia assessment and training.

Keywords

item response theory, latent growth curve model, formative assessment, random slope random intercept model, smooth growth curve

---

Gesa Brunn, ORCID: 0000-0002-0570-4520 · Prof. Dr. Philipp Doebler (corresponding author), ORCID: 0000-0002-2946-8526, Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany
email: gesa.brunn@tu-dortmund.de
doebler@statistik.tu-dortmund.de

Dr. Fritjof Freise, ORCID: 0000-0002-8493-8359, Institute for Biometry, Epidemiology and Information Processing, University of Veterinary Medicine Hannover, Bünteweg 2, 30559 Hannover, Germany
email: Fritjof.Freise@tiho-hannover.de

Modellierung eines glatten Lernverlaufs und Testung individueller Abweichungen von einem globalen Verlauf

Zusammenfassung

Formatives Assessment liefert Lernenden und Lehrenden wertvolles Feedback und ist durch computergestützte Implementationen stark vereinfacht worden. Zwar sind längsschnittliche individuelle Assessments und Vergleiche innerhalb einer Klasse nützlich, aber normative Interpretationen von individuellen Lernverläufen können nur relativ zu einer Referenzpopulation gegeben werden. Da aktuelle computergestützte Assessment-Systeme Items aus Pools zufällig auswählen oder Tests adaptieren, arbeiten die Getesteten u. U. auf sich nicht überlappenden Itemmengen, wodurch klassische Summenscores nicht direkt vergleichbar sind. Um dem zu begegnen, wird das Smooth Growth and Linear Deviations Rasch Model (SGLDRM) eingeführt, eine Erweiterung des Rasch-Modells für binäre Testdaten aus der Item-Response-Theorie. Durch Splines wird ein glatter globaler Verlauf eingebunden. Das Modell ist flexibel genug, um Anstiege und Verringerungen des mittleren Fähigkeitsniveaus abzubilden, welche je nach Messzeitpunkt unterschiedlich stark ausgeprägt sein dürfen. Auf der individuellen Ebene wird der globale Lernverlauf durch gut interpretierbare zufällige Achsenabschnitte und Steigungen modifiziert. Zwei Messzeitpunkte reichen aus, um personenspezifische Verläufe zu schätzen. Ein Likelihood-Quotienten-Test erlaubt es, Lernende zu identifizieren, die vom mittleren Lernverlauf abweichen. Die Methode wird anhand von Daten aus einem Online-System zur Diagnostik und Behandlung von Dyskalkulie illustriert.

Schlagworte

Item-Response-Theorie, latentes Wachstumskurvenmodell, formatives Assessment, Random-Slope-Random-Intercept-Modell, glatte Wachstumskurve

1. Introduction

The possibility to efficiently implement and administer test items with computerized test platforms facilitates the routine assessment of learning development (e.g., Klinkenberg et al., 2011; Kuhn et al., 2018; Mühling et al., 2017; Souvignier et al., 2014), and has progressed to a stage where system output informs school teachers rather than research scientists (e.g., Schurig et al., 2019). Next to processing test data to extract information on the overall development, it can also be of interest to track individual progress, give feedback on the performance, and initiate remedial measures if necessary, in the tradition of assessment for learning (Black & Wiliam, 1998). In order to utilize the assets of computer testing, we focus on the already highly developed item response theory (IRT) which offers strong frameworks for model extensions.

IRT models for longitudinal data are becoming increasingly important to track educational progress on a global and individual level. IRT provides several frameworks to model subject-based growth over multiple occasions, many of them being extensions of the Rasch (1960) model for binary item responses. A line of research starting with the seminal paper of Rost and Spada (1983) focuses on latent differences between occasions (Andersen, 1985; Embretson, 1991; Fischer, 1973, 1976, 1989). Initially, the approach was intended for two fixed occasions, but subsequent models were more general. Noteworthy refinements include the generalization of this approach to the two parameter logistic case by Embretson (1997), which was further extended (Andrade & Tavares, 2005).

However, there are still several issues that should be addressed and improved when modeling longitudinal data. Especially in computerized test and training systems, data collection often does not happen at fixed discrete equally spaced occasions, be it due to randomness or by design. Adequate models should allow for many if not all time values in an interval. Therefore, it is convenient to consider the time component to be continuous, and observations to be snap-shots of the current ability (e.g., Hecht et al., 2019). By using classic parametric models to describe the global progression of performance, for example, linear or quadratic, a certain shape of the global growth curve is presumed. This can severely restrict model fit and lead to biased representations.

In this study, we propose a non-parametric method to estimate a smooth global trend based on splines. This approach enables us to detect even small unexpected changes over time. Although the main focus of modeling longitudinal data in IRT has been the growth of a population or several subgroups, subject-specific performance can be of interest, for example when psychometric tests are used as formative evaluation tools parallel to educational interventions. By allowing a linear deviation from the average growth for each subject, we can track the individual starting level and gain given by a random intercept and slope, respectively. A likelihood ratio test allows identifying individual courses that significantly deviate from the average growth. As a consequence, feedback is possible and the initiation of interventions based on the intensity and direction of the deviation.

Saha (2016) presents a similar approach: The Bayesian dynamic item response model with semi-parametric and smooth ability growth (DIR-SMSG) uses B-spline functions to estimate ability growth in a dynamic IRT (Wang et al., 2013) framework. Saha's (2016) model works with a discrete time component and assumes ability growth to be monotone. The present approach avoids monotonicity assumptions, incorporating growth, set-backs, and learning boosts. In contrast to the approach presented here, the DIR-SMSG uses spline functions to estimate each person's ability separately. While a flexible model for individual growth results, the approach requires relatively dense data on the person level. To ensure accessible interpretation of both the population ability growth and person-specific deviations from the average growth the Smooth Growth Linear Deviation Rasch Model (SGLDRM) assumes person-specific linear deviations from the global trend. In this case, only two or more measurement occasions are required in order to estimate a

person-specific ability growth which can be beneficial when only relatively few data are observed on the person level. In terms of complexity, the SGLDRM is between longitudinal IRT models with strong linearity assumptions on the logit scale and the DIR-SMSG.

The population's development is an explicit part of the SGLDRM, while it is necessary to treat individual subject effects as random effects to avoid over-parametrization (Baayen et al., 2008; Hecht et al., 2019). Realizations of these random effects are not estimated when fitting the model (but their variances and covariances are). However, the estimates (in the form of conditional modes/best linear unbiased predictors, BLUPs; Robinson, 1991) can be obtained subsequently.

The remainder of this paper is organized as follows: The advantages of item-level scaling in the context of formative assessment are stressed in the Section 2 and it is explained why non-linear growth curves safeguard against potential problems. Next, the SGLDRM, an extension of the Rasch model for longitudinal data, is presented: The model features a global smooth curve and individual deviations. The model can be seen as a generalized additive mixed model, leading to an estimation approach. The subsequent Section 4 develops a likelihood ratio test which helps to discern whether individual courses of learning deviate from the global course. The test is evaluated by simulation. An application to data from an online sample of primary school students with math difficulties or dyscalculia illustrates the usefulness of the method and is contained in Section 5. We close with some remarks on the limitation of the method and its relationship to existing methods.

2. Flexible Scaling on the Item Level With Non-Linear Growth Curves

Scaling on item level is a central characteristic of IRT and it produces some clear advantages compared to scaling on (sub-)test level. This has been stressed especially in the context of longitudinal applications (Reise & Haviland, 2005), though Jabrayilov et al. (2016) caution, that sufficient test length is still needed: An IRT model with acceptable fit for at least 20 items is recommended to reach acceptable misclassification rates of latent change. More generally, an ideal item set for measuring change provides high Fisher information for relevant parts of the ability continuum, so that person-specific reliability of the change scores is acceptable. When scaling on the level of a whole test, for example, with raw scores, the comparison of test results requires the tests to be equally difficult and the response patterns to be similar in the sense that subjects with the same true score make mistakes on items with similar difficulty. Creating such parallel tests for intensive longitudinal situations is complex, but not impossible (e.g., Fuchs et al., 1984; Strathmann & Klauer, 2012).

In contrast, scaling on item level allows calibrating large item pools, each test taker only working on a subset of items (Kolen & Brennan, 2004). These subsets

can even be assembled randomly, or – for more effective and efficient testing – be drawn adaptively in computerized adaptive tests (CATs; van der Linden & Glas, 2000). Adaptive testing avoids ceiling and floor effects within the limits of the test design. With item-level scaling it is possible to model data from tests with time limits as it is not required for tests to be of the same length. Note, that this will work under the premise that the difficulty of an item is not affected by the number of items assigned in a test. However, item position effects do affect item parameters and include phenomena such as within-test learning and fatigue, and have been documented by Le (2007), Debeer and Janssen (2013), Debeer et al. (2014), Nagy et al. (2018), and Wu et al. (2019). A crucial problem is a correlation of item parameters and item position, say if a linear test design is used in estimation, but item order is varied subsequently. Hence, testing systems should either not vary item position or vary item position in the calibration system.

The validity of individual items can be examined when scaling on item level, which is beneficial for the process of item construction. These advantages of scaling on the item level require an overlap of the item subsets administered to each person, so that item parameters are properly linked (Kolen & Brennan, 2004). Randomly sampling items is a typical way to achieve this. The discussion highlights further caveats.

Next to flexibility with respect to the item subsets, the approach we introduce is flexible with respect to sampling occasions as well as with respect to the shape of the global course of learning. The use of spline functions (Wood, 2017) to model the global learning course allows flexible adjustments making it more convenient to fit real data compared to parametric alternatives. Smooth functions are especially useful to model short-term changes, for example, setbacks after school holidays or intervention effects.

3. Smooth Growth and Linear Deviation Rasch Model

We assume that $t = 0$ indexes the first measurement occasion and that there are occasions $t = 0, \ldots, T$, not necessarily evenly spaced. We do not assume that all persons are measured at the same occasions, reflecting a missing data situation that is common in longitudinal data. We index a pool of items by $i = 1, \ldots, I$ and assume that each person responds to a (maybe empty) subset of items at a measurement occasion and this subset might or might not be identical for all persons $j = 1, \ldots, J$.

The SGLDRM is introduced now. Let Y_{jit} be a dichotomous random variable that indicates whether at measurement time t person j answered item i either correctly ($Y_{jit} = 1$) or not ($Y_{jit} = 0$). Further assume that over the course of time a person's ability ϑ_{jt} can change while item difficulties β_i stay invariant. As the probability of solving an item should increase with a person's ability and decrease the more difficult an item is, the probability is assumed to increase monotonously with $\vartheta_{jt} - \beta_i$. The Rasch model assumes a logit-linear relationship:

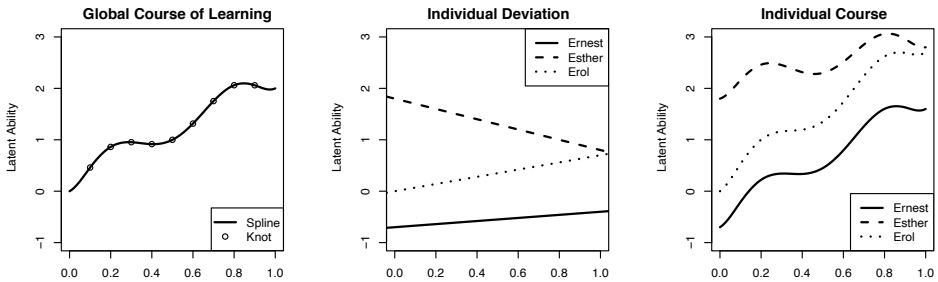
$$P(Y_{jit} = y_{jit}) = \frac{\exp(y_{jit}(\vartheta_{jt} - \beta_i))}{1 + \exp(\vartheta_{jt} - \beta_i)}. \tag{1}$$

In our extension of the Rasch model, the latent person ability at time t is assumed to be composed of a latent global trend $\lambda(t)$, in addition to a person-specific latent intercept δ_j and slope γ_j, such that

$$\vartheta_{jt} = \lambda(t) + \delta_j + \gamma_j t. \tag{2}$$

Note, that $\delta_j + \gamma_j t$ is a person-specific linear function in t that is interpreted as the deviation of person j from the global trend. We assume that the function of global mean latent ability on time, $\lambda(t)$, is smooth, that is, we rule out sudden jumps. Otherwise, the function has an arbitrary shape which has to be estimated from the data and is not specified a priori. The technical approach via spline functions is detailed below. Figure 1 illustrates the approach.

Figure 1: Illustration of the Model: A Smooth Global Course is Modified by a Linear Individual Deviation (Synthetic Data for Three Students Named Ernest, Esther and Erol)



Similar to many other IRT models, local stochastic independence is assumed: Responses made by person j are independent conditional on person ability at time points t_1 and t_2 (where $t_1 = t_2$ is possible). Hence, conditional on item abilities at t_1 and t_2, the answer to item i_1 at t_1 does not affect the answer for item i_2 at t_2 (where $i_1 = i_2$ is possible), which can be expressed as

$$P(Y_{ji_1t_1} = y_{ji_1t_1} | \vartheta_{jt_1}, \vartheta_{jt_2}) = P(Y_{ji_1t_1} = y_{ji_1t_1} | Y_{ji_2t_2} = y_{ji_2t_2}, \vartheta_{jt_1}, \vartheta_{jt_2}). \tag{3}$$

Similarly, the probability of person j_1 solving item i correctly does not depend on whether person $j_2 \neq j_1$ was able to solve item i, leading to an independence assumption for persons.

If the same item is used at two different time points, the model, as specified, ignores potential local item dependence, that is, a person's chance to correctly answer at t_2 is conditionally independent on whether the item was solved at t_1 or not.

This might be most plausible if the item index represents a whole family of items (see the sample application), or for domains like mental calculation, attention, or clerical speed, while the assumption is implausible for items requiring some sort of insight or factual knowledge. Our recommendation is hence to avoid reusing the very same items if the SGLDRM is to be applied.

To complete the model specification of the SGLDRM, a bivariate normal marginal distribution for the latent variables δ and γ is assumed:

$$(\delta, \gamma)' \sim \mathrm{N}_2(\mathbf{0}, \Sigma), \tag{4}$$

where

$$\Sigma = \begin{pmatrix} \sigma_\delta^2 & \rho\sigma_\delta\sigma_\gamma \\ \rho\sigma_\delta\sigma_\gamma & \sigma_\gamma^2 \end{pmatrix} \tag{5}$$

is a not necessarily diagonal 2 × 2 covariance matrix. It is possible to constrain $\rho = 0$ for ease of fitting the model, which might however be implausible in some contexts. The individual latent variables $(\delta_j, \gamma_j)'$ are independent copies of the bivariate random variable $(\delta, \gamma)'$.

While the interpretation of ϑ_{jt} and β_i is familiar from the Rasch model, it is important to note how to interpret the other parts of the SGLDRM: The global latent growth curve λ(t) coincides with the latent ability of an average person with $\delta_j = \gamma_j = 0$. The parameter σ_δ^2 is the variance of latent ability at $t = 0$. The standard deviation σ_δ of latent ability is on the familiar logit scale and is hence easier to interpret. The larger σ_δ, the larger the spread of ability at $t = 0$. If $\gamma_j = 0$, the person-specific course of learning is just shifted up or down by δ_j. Hence, the larger σ_γ, which is the standard deviation of the person-specific slope of the linear deviation, the less likely it is for the individual trajectories to be practically parallel to the global curve given by $\lambda(t)$. Finally, ρ, the correlation of δ and γ, is to be interpreted depending on its sign: If ρ is positive, persons starting above the latent mean at $t = 0$ ($\delta_j > 0$) have larger slopes γ_j on average than those below the latent mean at $t = 0$ ("The rich get richer and the poor get poorer."). If ρ is negative, persons starting above average ($\delta_j > 0$) tend to have smaller slopes than average ($\gamma_j < 0$), implying that, on average, persons with low ability catch up (to some extent) and that the latent variance might decrease as t increases.

3.1 The SGLDRM as a Generalized Additive Mixed Model

For purposes of parameter estimation, it is useful to understand the SGLDRM as a binomial generalized additive mixed model (binomial GAMM; Wood, 2017) with a logit link function, where $\lambda(t)$ is a smooth function of the covariate *time t*. The latent intercept δ and slope γ are random effects from the GAMM perspective

and the item difficulties β_i are treated as fixed effects. Some details on binomial GAMMs follow, so that the connection becomes apparent.

A binomial GAMM with a univariate smooth function s is of the general form (Wood, 2017)

$$g(E(y_l|\mathbf{b})) = \mathbf{X}_l\mathbf{a} + s(x_l) + \mathbf{Z}_l\mathbf{b}. \tag{6}$$

In the GAMM context, g is a differentiable link function with logit and probit being the most popular choices. The logit is chosen here, as this preserves the log-odds interpretation of many parameters familiar from Rasch models. On the right-hand side of the equation we have a model matrix of fixed effects $\mathbf{X}$ and the corresponding parameter vector $\mathbf{a}$. Here, only the item difficulties β_i are fixed effects, but the GAMM perspective allows for straightforward inclusion of other fixed person or item effects.

The smooth component $s(\cdot)$ is included for flexible modeling of the effect of time on the learning process. Polynomials of high orders seem to offer a similar amount of flexibility, but overfitting polynomials leads to implausible oscillating patterns, limiting the ability to predict ability by interpolation. Instead, we use spline functions to approximate the process. They are highly flexible, smooth, and rather simple to use. Spline functions are reviewed with some detail in Subsection 3.2.

The random person effects δ_j and γ_j for the linear deviations in the SGLDRM are represented by the vector of random effects $\mathbf{b}$ and the corresponding model matrix of random effects $\mathbf{Z}$. By not treating them as fixed effects, the person-specific coefficients $(\delta_j, \gamma_j)'$, $j = 1, \ldots, J$, are not estimated when fitting the model. From a technical point of view, the person parameters are nuisance parameters and by treating them has random effects, they do not have to be estimated. However, after model fitting, individual estimates can be obtained, which we detail below. By eliminating person parameters from the model, the model fit does not depend on the particular set of people, but applies to subjects with abilities from the same distribution.

3.2 Regression Splines

Regression spline functions are a composition of continuous functions joined together at so-called knots to fit a smooth function to a certain set of noisy data (Wood, 2017). Instead of determining the shape of the mean course of learning in advance, for example, by choosing a linear function or another polynomial of some degree, the use of spline functions provides a more flexible fit to the data points and therefore is presumably closer to the "true" course of development. The true course is, of course, unknown since it is a latent psychological construct. The existence of a smooth underlying function of the true course is an assumption.

One issue in fitting a smooth function to data is to compromise between smoothness and fit. A class of splines that are commonly used as default in implementations are thin plate regression splines (Wood, 2003). They are based on the idea of minimizing the squared distance between spline and actual data, penalized by an additive term that gets larger the "wigglier" a spline gets, the so-called penalized residual sum of squares. A detailed illustration of spline functions can be found in Green and Silverman (1994).

When fitting a model with penalized regression splines, one has to choose the dimension of the basis that spans the space of spline functions. While a dimension which is too small can deteriorate a model's fit, a higher dimensionality primarily leads to a higher computational cost of the thin plate regression spline (Wood, 2017). By choosing a basis with a sufficiently high dimension the fitted curve's flexibility is unrestricted, while overfitting (too much "wiggliness") is prevented by determining the penalty tuning parameter by a generalized cross-validation scheme. In other words, choosing a high dimension is not critical in packages with penalized splines, and this is one of the reasons we recommend the packages gamm4 (Wood & Scheipl, 2017) and mgcv (Wood, 2017) to fit the SGLDRM. Wood (2017, Chapter 5.9) provides an insight into the procedure of choosing and checking the basis dimension. A guideline given by Gu and Kim (2002) says that the basis dimension should be around $10n^{\frac{2}{9}}$ where n is the number of observations. Additional information on the estimation procedure and suggestions for implementations can be found in the Appendix.

As soon as the model is fitted to a sufficiently large set of data based on a representative set of subjects, the mean course of learning can, among other purposes, be used to compare the performances of newly tested individuals to it. This point is taken up in the next section.

4. Deviations From the Mean Course of Learning

After a baseline assessment, individual person parameters can be estimated, and point estimates together with asymptotic (e.g., Lord, 1983) or exact confidence intervals (Doebler et al., 2013; Klauer, 1991) can be used to assess initial ability. We now discuss a method for when longitudinal data is available: The aim is to compare results of an individual participant and its change over time to the mean development of the population. The SGLDRM composes a person's ability at a certain point in time by a global trend and a person-specific intercept and slope. This composition can be used to develop statistical hypotheses regarding the divergence of a person's ability from the global trend. Some approaches to hypothesis testing are presented below.

Note, that the inference in the following section is conditional given a fixed person j and, hence, the individual random effects δ_j and γ_j can be considered to be fixed parameters. Additionally, $\lambda(t)$ as well as item parameters β_i, $i = 1, \ldots, I$, are

assumed to be known. Since we consider the deviation of an individual from the mean, all inference is done conditional on this person. The person index j is omitted in this section since a particular person's data is not necessarily part of the data set that is used to obtain the global course.

Assuming that the average course of learning $\lambda(t)$ represents typical or healthy development, it is important to check whether an individual's course of learning matches the mean course of learning relating to a certain population. If this was the case this person parameter, denoted by ϑ_t, would only consist of the global trend for all points in time t:

$$\vartheta_t = \lambda(t). \tag{7}$$

With respect to the person-specific linear deviation this yields the hypothesis

$$H_0 : \delta = \gamma = 0. \tag{8}$$

If this null hypothesis is discarded the individual either would have a level of ability above or below average ($\delta_j \neq 0$), or learn faster or slower than the average ($\gamma_j \neq 0$) or a combination of both. Figure 2 shows some theoretical learning courses.

To test the hypothesis in (8), we propose to employ a likelihood ratio test (LRT), which compares the likelihood functions under H_0 and the alternative H_1. We first make the likelihood $L(\delta,\gamma \mid \mathbf{y})$ explicit for a pair $(\delta,\gamma)'$ given the data $\mathbf{y}$ of person j: Assume test performance of person j has been observed at times t_k, $k = 1, \ldots, K$. The items administered to the individual at hand might not be the same at each point in time, so denote the index sets containing the item indices by A_{t_k}, $k = 1, \ldots, K$. Thus, one observes $y_{it_k} \in \{0,1\}$ for $i \in A_{t_k}$ and $k = 1, \ldots, K$. Local independence assumptions then imply

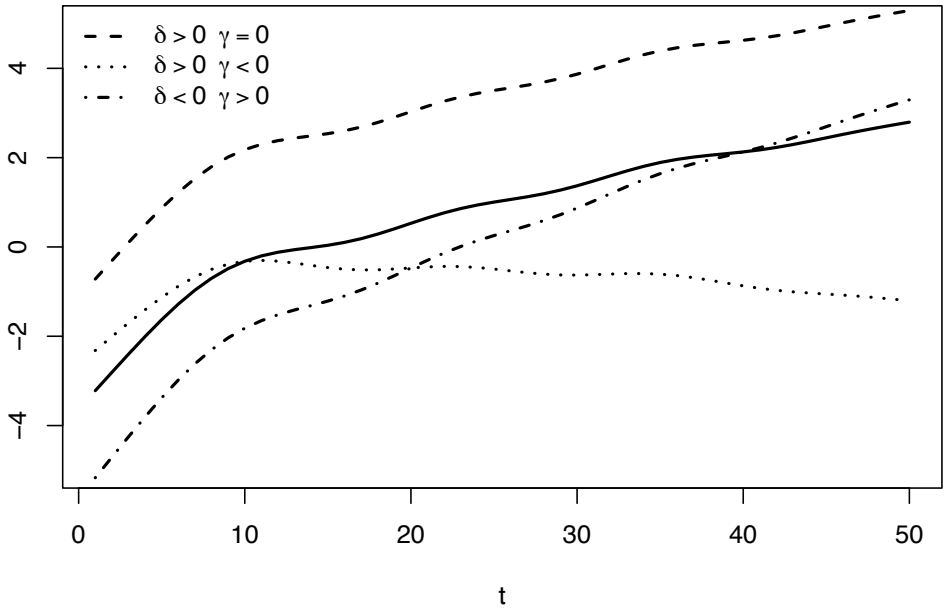
$$L(\delta,\gamma|\mathbf{y}) = \prod_{k=1}^{K} \prod_{i \in A_{t_k}} \frac{\exp(y_{it_k}(\vartheta_{t_k} - \beta_i))}{1 + \exp(\vartheta_{t_k} - \beta_i)}. \tag{9}$$

The ratio of the likelihoods is then given by

$$r(\mathbf{y}) = \frac{L(0,0|\mathbf{y})}{L(\hat{\delta},\hat{\gamma}|\mathbf{y})} \tag{10}$$

where numerator and denominator are the maximum of the likelihood under H_0 and the alternative, respectively. The value of the latter is the likelihood taken at the ML estimate, $(\hat{\delta},\hat{\gamma})'$, which is the conditional mode/BLUP.

Figure 2: Illustration of a Smooth Global Trend (Solid Line) and Three Exemplary Deviations (Broken Lines)



The smaller the ratio $r(\mathbf{y})$, the less likely it is that H_0 is true given the observations $\mathbf{y}$. Under H_0 the statistic $W = -2\ \ln(r(\mathbf{y}))$ is asymptotically χ^2_2-distributed (Casella & Berger, 2002). In finite samples, especially when few items are used, the χ^2-approximation can fail. The finite sample performance can be improved by an empirical Bartlett correction: Rather than comparing the test statistic $W = -2\ \log(r(\mathbf{y}))$ to the corresponding quantile of the χ^2_2-distribution, it is first multiplied by an estimate of the factor $2/\ \mathrm{E}(W)$, so that the corrected test statistic estimates (Pawitan, 2001)

$$\mathrm{W}^* = \frac{2\mathrm{W}}{\mathrm{E}(\mathrm{W})}. \tag{11}$$

4.1 Variants of the LRT

There are two potentially useful variants of the LRT: Assume an educator is interested to find out whether (a) a child manages to develop parallel to the mean course (i.e., is the distance to the global mean course stable?), or (b) whether a child under- or overperforms given what growth is expected by the child's initial ability. Mathematically, the first variant (a) is straightforward: The course is parallel if, and only if, $\gamma = 0$. So maximizing $L(\hat{\delta}_{parallel}, 0 \mid \mathbf{y})$ as a function of $\hat{\delta}_{parallel}$ and comparing to the ML-estimate yields the variant of the test statistic. Setting

$$r_{\text{parallel}}(\mathbf{y}) = \frac{L(\hat{\hat{\delta}}_{\text{parallel}}, 0|\mathbf{y})}{L(\hat{\hat{\delta}}, \hat{\gamma}|\mathbf{y})} \tag{12}$$

we define $W_{parallel} = -2 \ln(r_{parallel}(\mathbf{y}))$. Then $W_{parallel}$ is approximately χ^2_1-distributed under H_0: $\gamma = 0$ and one can proceed as before. We mention in passing that a reference value of $\gamma \neq 0$ could also be tested.

For the second variant (b) first note that when the bivariate normal distribution assumption in the SGLDRM holds (with $E[\delta] = E[\gamma] = 0$), the conditional expectation of γ given δ is given by

$$E(\gamma|\delta) = \rho \frac{\sigma_\gamma}{\sigma_\delta} \delta \tag{13}$$

by standard results for the bivariate normal distribution. The parameters in this expression are taken from (estimates) of the covariance matrix in Equation 5. The conditional expectation reflects the best guess of γ given δ, and hence reflects the growth one would expect given the initial performance. The second variant will hence be called the conditional variant, and we set

$$r_{\text{cond}}(\mathbf{y}) = \frac{L(\hat{\hat{\delta}}_{\text{cond}}, E(\gamma|\hat{\hat{\delta}}_{\text{cond}})|\mathbf{y})}{L(\hat{\hat{\delta}}, \hat{\gamma}|\mathbf{y})}, \tag{14}$$

and define $W_{cond} = -2 \ln(r_{cond}(\mathbf{y}))$. As in the parallel case, W_{cond} is approximately χ^2_1-distributed under H_0: $\gamma = E(\gamma \mid \delta)$ and one can proceed as before. However, the null hypotheses for $W_{parallel}$ and W_{cond} now contain more than one point. This means no empirical Bartlett correction can be implemented, since neither $E(W_{parallel} \mid H_0)$ nor $E(W_{cond} \mid H_0)$ are well defined.

4.2 Power of the Person-Level LRT and its Variants

In a simulation, we want to examine the performance of the proposed LRTs for some cases of individual deviation as well as the behavior under the null hypotheses. We focus on the LRT for H_0: $\delta = \gamma = 0$, but provide some insight into the variants.

4.2.1 Power of the LRT and Calibration Error

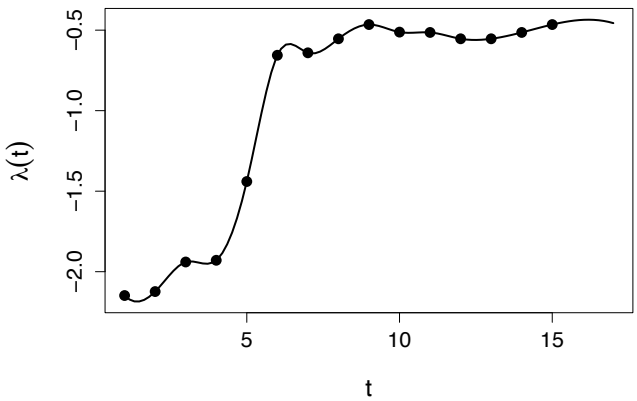
In a first step, assuming a nonlinear mean course of learning and certain values for Σ, the covariance matrix of the individual effects, a calibration sample is generated. To explore the frequency of rejection of the true null hypothesis (type I error),

the calibration data set is generated from Equations 1, 2, 4, and 5 with J subjects, $I = 5$ items, and $T = 10$ measurements per subject. The individual intercept and slope are set to be uncorrelated (thus $\rho = 0$). The different settings for Σ are given in Table 1. To additionally examine the LRT properties without a calibration error, comparisons made in setting N are based on the true parameters rather than estimations made on the calibration data. Since we would like to use the spline functions to their full capacity, we refrain from choosing a polynomial function as global learning course $\lambda(t)$. Instead, we arbitrarily choose 15 numbers and arrange them in ascending and descending sequences. We interpolate this pseudo data with a spline function and use this spline function as $\lambda(t)$. The data points as well as the spline are illustrated in Figure 3.

Table 1: Values of σ_δ and σ_γ for Simulated Data With $\rho = 0$ and Mean Vector $\mathbf{0}$ (see Equations 4 and 5). Values for β_i, $i = 1, \ldots, 5$ are set to $-1, -0.5, 0, 0.5, 1$ for all Calibration Sets (Including Setting N)

Calibration set	σ_γ	σ_δ	J
A	0.2	1	100
B	0.2	1	2500
C	0.75	0	2500
No calibration error (N)	–	–	–

Figure 3: Arbitrary Global Course Used for the Generation of Simulation Data



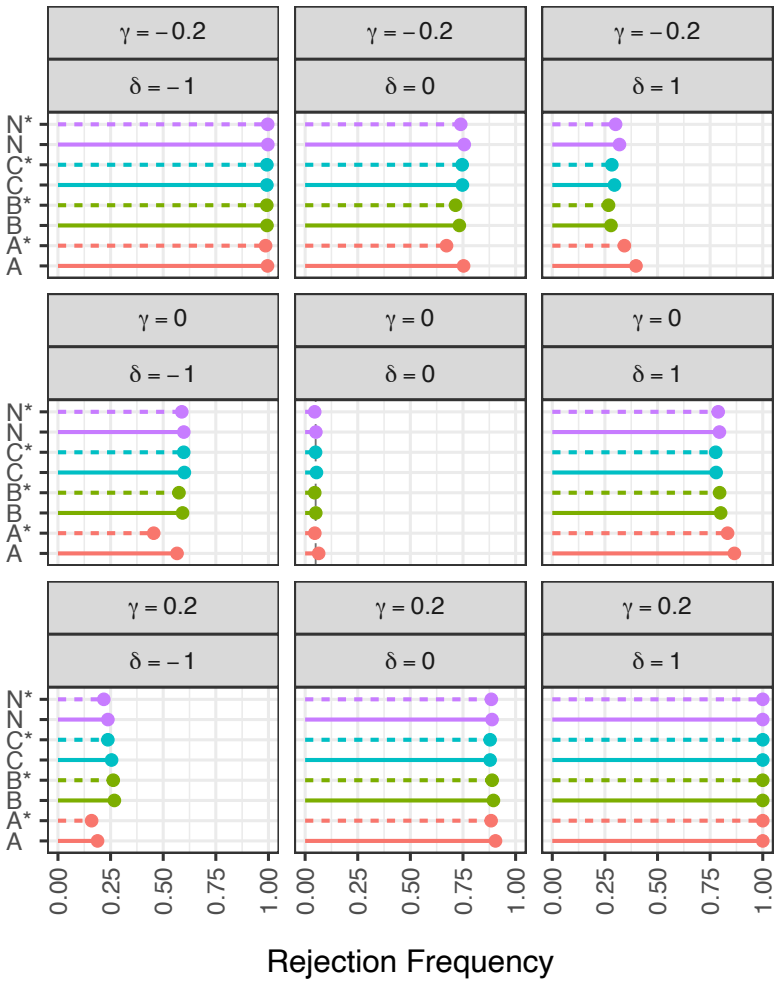
Note. The numbers used for interpolation are −2.15, −2.12, −1.94, −1.93, −1.44, −0.66, −0.64, −0.55, −0.47, −0.51, −0.51, −0.55, −0.55, −0.51, −0.47.

After fitting the model (Equation 6) to the simulated calibration data set an estimate (t) of the smooth course of learning results. The fitted global trends are shown in Figure 5. With only 100 virtual individuals from data set A the overall shape of the development is captured albeit smoother and with less precision than

with higher sample sizes (mean bias in the fixed effects: A: −0.122, B: −0.005, C: −0.006). An increasing variability in individual slopes seems to have no substantial impact on the estimation of λ.

Data points for new individuals are simulated by drawing correlated Bernoulli responses with conditional probabilities shown in Equation 1 for fixed values of γ and δ. To examine hypothesis test behavior under H_0, a first data set is drawn with γ = δ = 0. The LRT statistic, as presented in the previous section, is calculated by using the estimated mean learning curve $\hat{\lambda}(t)$ of the calibration data set and individual ML estimates $\hat{\delta}_j$ and $\hat{\gamma}_j$. This procedure is repeated 5000 times. Figure 4 shows the percentage of test statistics which are greater than the 95% quantile of

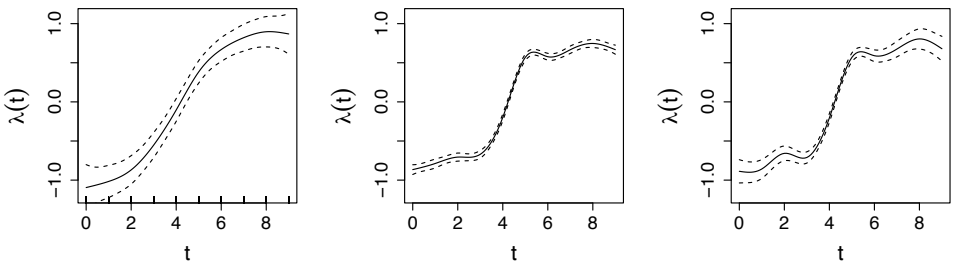
Figure 4: Rejection Frequencies of LRT Test at Significance Level α = 5% for Data Sets A, B, and C and Different Types of Deviation



Note. Bartlett-corrected performances are asterisked.

the χ^2-distribution. Under the null hypothesis (see center panel), it shows a rejection frequency that is slightly higher than 5% for all of the three calibration data sets (6.72%, 5.16%, 5.36%, respectively) as well as the comparison to the true parameters (5.10%). The significance level is exceeded. The Bartlett correction improves the results slightly for all data sets (4.52%, 4.56%, 5.18%, respectively) and setting N (4.54%). Table 2 holds bias, root mean squared error (RMSE), and the coverage of parameter estimations for the four simulation settings.

Figure 5: Average Courses λ(t) Fitted for Data Sets A, B, and C, Respectively



To examine the behavior under H_1, the values for δ and γ are varied systematically: The parameters take three different values respectively, one negative (δ = −1, γ = −0.2), one neutral (δ = γ = 0), and one positive (δ = 1, γ = 0.2). The test seems to find deviations more consistently if the deviations of both parameters are directed the same way. For example, a deviation of a student whose course of learning is both shifted upwards and increases more rapidly than the global course, for example, δ = 1, γ = 0.2, is very likely to be detected. However, individuals with an above-average intercept and a below-average slope are detected only about 30% of the times.

The biases in δ and γ are large for some cases, most pronounced when δ = −1 and γ = −0.2. The deeper reason are floor effects, that is, the test material is way too hard for the simulated persons. One can see the combination of δ = −1 and γ = −0.2 as a misfit of the item set and the person. As seen in Figure 3, the mean trajectories start in the range of −1.2 to −0.9. A person with δ = −1 hence starts at an ability level of −2.2 to −1.9 and hence the chance to solve an item of average difficulty is smaller than 13%. Over the course of the ten simulated measurement occasions, the mean gain is up to 2 units, depending on the mean trajectory in Figure 3. Hence, at Time 10 a γ of −0.2 leads to a latent ability which cancels out the mean gain or is even a net ability decrease over time. As a consequence, the whole trajectory has to be estimated based on potentially many zero scores, so the maximization of the person likelihood in Equation 9 can yield extreme estimates of δ and γ. In other words, the substantial growth we simulate uncovers deficits of the procedure in pathological cases.

Table 2: Rejection Rates of LRTs, Bias, Root Mean Squared Error and Coverage of Asymptotic 95% Confidence Intervals of Parameter Estimates Under Different Settings

Setting/ data set	δ	γ	Rej. rate	Rej. rate Bartlett	Bias δ	Bias γ	RMSE δ	RMSE γ	Cover-age δ	Cover-age γ
N	−1.0	−0.2	0.999	0.998	−1.47	−5.79	18.81	17.46	0.96	1.00
	0.0	−0.2	0.865	0.854	−0.46	−0.10	5.91	2.51	0.97	0.98
	1.0	−0.2	0.315	0.297	−0.06	0.00	0.87	1.21	0.96	0.96
	−1.0	0.0	0.598	0.588	−0.76	0.07	7.19	1.30	0.96	0.97
	0.0	0.0	0.051	0.045	−0.18	0.02	1.05	0.15	0.96	0.96
	1.0	0.0	0.794	0.788	−0.08	0.01	0.76	0.99	0.96	0.95
	−1.0	0.2	0.309	0.290	−0.26	0.03	1.24	1.24	0.96	0.96
	0.0	0.2	0.956	0.954	−0.18	0.03	0.93	0.27	0.96	0.96
	1.0	0.2	1.000	1.000	−0.11	0.03	0.78	0.78	0.97	0.97
A	−1.0	−0.2	0.997	0.994	−3.00	−5.81	29.22	17.86	0.93	0.99
	0.0	−0.2	0.834	0.778	−0.23	−0.13	8.08	2.63	0.93	0.97
	1.0	−0.2	0.376	0.315	0.25	−0.03	0.94	1.24	0.93	0.95
	−1.0	0.0	0.540	0.438	−0.43	−0.01	7.72	1.99	0.93	0.95
	0.0	0.0	0.067	0.045	0.14	−0.01	1.07	0.15	0.93	0.95
	1.0	0.0	0.871	0.839	0.25	−0.02	0.80	1.03	0.92	0.94
	−1.0	0.2	0.313	0.263	−0.00	0.01	1.27	1.22	0.94	0.94
	0.0	0.2	0.971	0.959	0.13	-0.00	0.92	0.24	0.93	0.95
	1.0	0.2	1.000	1.000	0.21	-0.00	0.83	0.82	0.93	0.95
B	−1.0	−0.2	0.997	0.997	−2.68	−5.41	24.15	17.20	0.96	0.99
	0.0	−0.2	0.846	0.833	−0.40	−0.20	4.73	3.42	0.97	0.98
	1.0	−0.2	0.281	0.268	−0.14	0.01	0.87	1.19	0.97	0.96
	−1.0	0.0	0.592	0.575	−0.61	0.03	3.62	1.73	0.97	0.97
	0.0	0.0	0.052	0.046	−0.22	0.03	1.02	0.15	0.97	0.97
	1.0	0.0	0.801	0.795	−0.12	0.02	0.77	0.99	0.96	0.95
	−1.0	0.2	0.340	0.334	−0.32	0.05	1.24	1.26	0.97	0.97
	0.0	0.2	0.962	0.959	−0.19	0.03	0.92	0.28	0.97	0.96
	1.0	0.2	1.000	1.000	−0.18	0.04	0.84	0.77	0.96	0.96
C	−1.0	−0.2	0.998	0.998	−2.60	−5.82	25.43	17.81	0.96	1.00
	0.0	−0.2	0.871	0.869	−0.82	−0.15	10.31	3.41	0.97	0.98
	1.0	−0.2	0.304	0.292	−0.13	0.01	0.88	1.20	0.96	0.96
	−1.0	0.0	0.628	0.627	−0.67	0.01	6.01	2.07	0.96	0.97
	0.0	0.0	0.054	0.052	−0.19	0.02	1.01	0.14	0.97	0.96
	1.0	0.0	0.764	0.763	−0.14	0.02	0.77	0.99	0.96	0.96
	−1.0	0.2	0.302	0.290	−0.29	0.03	1.29	1.25	0.96	0.96
	0.0	0.2	0.949	0.948	−0.21	0.03	0.93	0.27	0.96	0.97
	1.0	0.2	1.000	1.000	−0.17	0.04	0.84	0.78	0.96	0.96

Note. Rej. = rejection; RMSE = root mean squared error.

4.2.2 Behavior of Variants of the LRT

We compare the LRT with its variants in the setting with no calibration error (N). Recall, that the global growth curve covers a range of 2.5 logits and hence reflects extreme growth. We hence also employ the same curve multiplied with a factor of 0.25, reflecting moderate global growth, and use the same parameter settings and procedure otherwise. For W_{cond}, we use $\rho = 0.2$, $\sigma_\delta = 1$, and $\sigma_\gamma = 0.2$ to calculate $E(\gamma \mid \delta)$. For all tests the scenario $\delta = \gamma = 0$ is contained in the null hypothesis, and for $W_{parallel}$, all scenarios with $\gamma = 0$ are consistent with H_0. We determined the proportion of rejected null hypotheses for W, $W_{parallel}$, and W_{cond} in 5000 replications, and the results are presented in Table 3. Since the Bartlett correction is only applicable to W, we omit it in this simulation.

Table 3: Proportion of Rejected LRTs

True parameter			LRT (variant)		
δ	γ	λ	W	$W_{parallel}$	W_{cond}
−1	−0.2	1.00	.994	.996	.999
0	−0.2	1.00	.757	.720	.795
1	−0.2	1.00	.305	**.038**	**.055**
−1	0.0	1.00	.617	**.694**	.702
0	0.0	1.00	.052	.035	.055
1	0.0	1.00	.793	**.825**	.872
−1	0.2	1.00	.237	**.031**	**.063**
0	0.2	1.00	.882	.819	.888
1	0.2	1.00	1.000	1.000	1.000
−1	−0.2	0.25	.999	1.000	1.000
0	−0.2	0.25	.847	.826	.829
1	−0.2	0.25	.391	**.053**	**.056**
−1	0.0	0.25	.791	**.908**	.875
0	0.0	0.25	.047	.048	.054
1	0.0	0.25	.852	**.883**	.910
−1	0.2	0.25	.367	**.048**	**.043**
0	0.2	0.25	.878	.778	.859
1	0.2	0.25	1.000	1.000	1.000

Note. 5000 replications. Cases with excessive rejections of H_0 or low power in boldface.

The upper half of the fourth column of Table 3 is for the W LRT (both, γ and δ, are freely estimated under H_1). This reproduces a portion of Table 2, with minor dis-

crepancies in the third decimal place due to Monte Carlo errors, so we do not repeat the above interpretation. Generally speaking, the variants of the LRT are able to detect violations of the null hypothesis, but there are scenarios which have low power (especially when the global course and the individual course intersect, i.e., $\delta = 1$, $\gamma = -0.2$ or $\delta = -1$, $\gamma = 0.2$). In some scenarios, compatible with the null hypothesis, $W_{parallel}$ will create false positive test results ($\delta = 1$ or $\delta = -1$). When the growth curve is extreme, the recovery of the person parameters is biased, affecting in turn the LRT variants. The effect is less pronounced when the growth is less extreme, but still observable (with 32% and 18% rejections when $\gamma = 0$ and $\delta = 1$ or $\delta = -1$, respectively). We recommend using the variants only when person parameter estimates can be assumed to be unbiased, severely limiting their applicability.

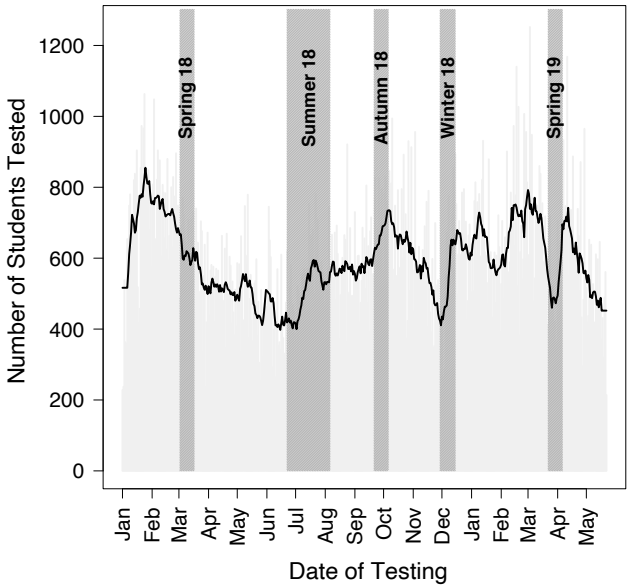
5. Empirical Illustration: Dyscalculia

Data from *Meister Cody von Talasia* (CODY) is reanalyzed. CODY is an online test and training system that was developed to monitor the learning progress of children with mathematical learning difficulties. More specifically, data from a numeracy test previously studied by Schwenk et al. (2017) is reanalyzed. The response accuracy is the target criterion; the two alternatives, response speed and response efficiency (cf. Schwenk et al., 2017), are not considered here. We stress that we are not measuring numeracy in an all-encompassing sense, since response times are not considered here.

Basic arithmetic skills are assessed with addition and subtraction items in a number range between 0 and 20. For each type of calculation task the test taker has 90 seconds to answer as many items correctly as possible. The items are randomly drawn from a pool of $N_{add} = 127$ and $N_{sub} = 143$ calculations and are presented to the user one by one. The tests go along with corresponding training sessions and are integrated into a motivational story. The data on hand was collected between January 2018 and June 2019 and comprises almost 300 000 individual observations made in about 14 500 tests on 3500 children. The usage frequency over the years and potential setbacks due to vacations are pictured in Figure 6. In order to make the SGLDRM applicable to the data, we transformed the time stamps and clustered the items. One should note that there is a variety of possibilities to transform and create the necessary variables and the following approach is taken for demonstration purposes.

Table 4 gives an idea of the longitudinal data set. Two time variables are part of the data set: a *day* variable giving the number of training sessions since the child's first registration and the *date* of the assessment. We combine these two variables to create a quasi-continuous time variable. For each student $t_j = 0$ indicates the time of their registration. We use *day* as a basis and add the time of day as a percentage. Students who are tested on their fifth day at 6 p.m. thus get a *time* value of $t_j = 4.75$.

Figure 6: Frequency of Application Usage Smoothed by Central Moving Average With Window Length 9 Days



Note. Vacation periods for North Rhine-Westphalia are added to indicate nationwide vacation periods.

Table 4: General Structure of Dyscalculia Data

Index	Subject	Day	Calculation	Date	Task type	Correct
1	1	5	4 + 8 = 12	2018-05-13 10:03:08	add	0
2	1	5	9 + 9 = 18	2018-05-13 10:03:19	add	0
3	1	5	11 – 1 = 10	2018-05-13 10:03:30	sub	1
4	2	5	2 + 3 = 5	2018-02-14 15:27:57	add	1
5	2	5	1 + 1 = 2	2018-02-14 15:28:06	add	1
6	2	5	19 – 7 = 12	2018-02-14 15:28:13	sub	0
7	1	10	4 + 8 = 12	2018-05-20 16:05:50	add	1
8	1	10	9 + 9 = 18	2018-05-20 16:06:08	add	0

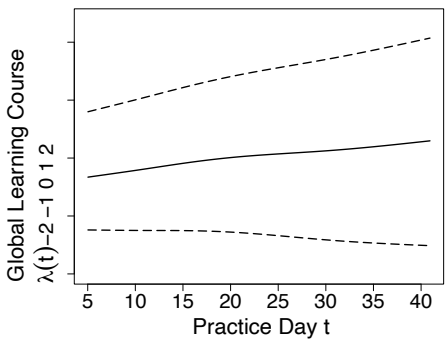
Note. add = addition; sub = subtraction.

There are 270 different calculation tasks in the item pool which would make their estimation computationally intense and consume much of the information that we would rather use to compute the overall course of learning. Instead we categorize the addition and subtraction items into three different groups, respectively: calculations within 10 (*add* ↓ 10, *sub* ↓ 10), between 10 and 20 (*add* ↑ 10, *sub* ↑ 10), and those that require passing the 10 (*add* ↕ 10, *sub* ↕ 10). In that way, there are only

six item parameters to be estimated while we are still able to assess the difficulties of the item clusters.

For model fitting we only use observations from children tested on days 5, 10, 15, ..., as this time lag is recommended by the CODY authors, and who participated in at least six assessments. Apart from that, tests with "perfect" scores are excluded to maintain estimability and interpretation. Thus, we consciously use only 37% of the total observations for parameter estimation, so that the mean curve represents an average child requiring training and adhering to CODY training for at least 30 training days. Figure 7 shows the estimated mean curve that corresponds to $\lambda(t)$ of the SGLDRM. To have some kind of guideline for individual performance, the first and third quantile of the sample parameters are added to the curve. Table 5 shows the estimated item parameters. As anticipated, estimates suggest that addition and subtraction tasks in the number range from 0 to 10 are the easiest. Addition tasks seem to be easier than the according subtraction tasks.

Figure 7: Estimated Global Course of Learning for the Dyscalculia Example



Note. Dashed lines at the first and third sample quartiles of subject parameters δ and γ.

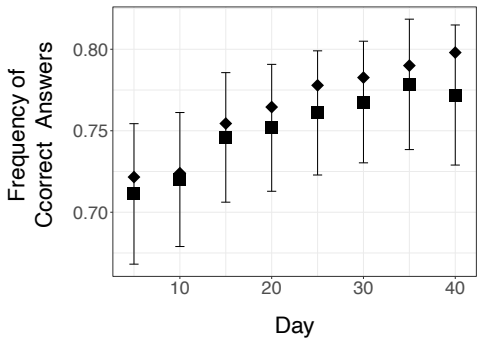
Table 5: Estimations of Item Parameters for the Dyscalculia Example for the SGLDRM

Item parameter	$\beta_{add\updownarrow 10}$	$\beta_{add\uparrow 10}$	$\beta_{add\downarrow 10}$	$\beta_{sub\updownarrow 10}$	$\beta_{sub\uparrow 10}$	$\beta_{sub\downarrow 10}$
estimation	1.25	1.53	2.69	0.61	0.99	2.12

Note. Increasing values indicate decreasing difficulty.

Additionally, we examine overall learning development by simply determining the averages of correct responses per day (see Figure 8). The development is similar to the curve that was estimated for the SGLDRM. Apart from a more defined increase the major difference is a decrease in performance from day 30 to day 40 of the mean when perfect scores are omitted (■), thereby excluding an increasing amount of correct responses as can be seen by comparing the development based on all observations (♦) and those without perfect scores (■). The non-parametric method gives each calculation the same weight in the estimation process, which might lead to discrepancies between the methods especially when only few items are responded to in each test run.

Figure 8: *Mean of Correctly Answered Items per Day of all Given Observations (♦) and Perfect Scores Omitted (■)*



Note. Bars marking ± one standard error of the mean with perfect scores omitted.

We give examples of individual analyses now. Three students are chosen in order to demonstrate different types of developmental patterns. We randomly choose one student each with the characteristics (a) average learner starting at average level, (b) extraordinary learner, and (c) extraordinary initial level. Those three pseudo groups are certainly no educated classifications but are provided for demonstration. Student A seems to have an average performance (see Figure 9) with δ_A = 0.534 and γ_A = −0.017. Their likelihood ratio test statistic (without Bartlett correction) of W_A = 1.095 is below the quantile $\chi^2_{20.95}$ = 5.991, thus the Hypothesis H_0: $\delta_A = \gamma_A = 0$ cannot be rejected with given significance level α = 5%. The non-parametric analysis (see Figure 10) gives the impression of an unsteady performance by Student A oscillating around the average. Both methods show Student B starting at an average level of performance which soon drops rapidly (δ_B = 0.754, γ_B = −0.085). The likelihood ratio test indicates this individual's learning curve to significantly differ from the average course $\lambda(t)$ with W_B = 18.020. The third subject, Student C, also shows a significant finding with a test statistic of W_C = 43.865. However, in contrast to Student B it is due to their particularly good initial level of performance with estimated parameters δ_C = 1.095 and γ_C = 0.000.

Figure 9: Individual Learning Courses of Student A (red), Student B (green), and Student C (blue) as Estimated in the SGLDRM Framework

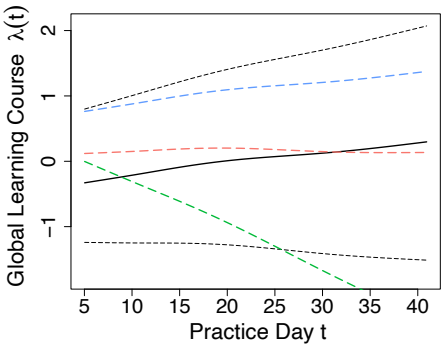
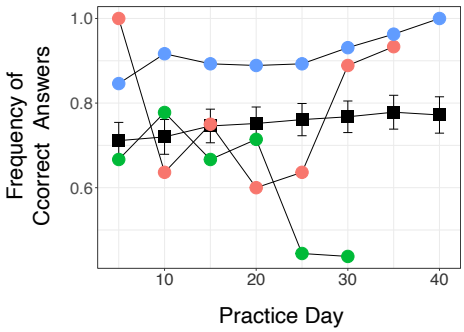


Figure 10: Individual Scores of Student A (red), Student B (green), and Student C (blue) in Comparison to Population Average Scores (black)



It is noteworthy that given analyses do not take any speed component into account. The number of completed items per test run are not used to estimate the students' abilities. This might be a great loss of information as can be seen in this example. While Student C works on 27.83 items on average with the overall mean of given data being 20.16 (*SD* = 5.65) their two fellow students do not do as well with 13.43 (Student A) and 9.33 (Student B) items per test. Although the performance of Student A might be average in regard to calculating accuracy, there might be a lack of speed in the performance. One naive method to deal with this inconvenience is to treat each item of the item pool that was not worked on in a test run as an incorrect answer. In that way it is equally valued to give an incorrect response or to not have the time to work on an item at all. However, this leads to different kinds of issues especially when working with great amounts of items and comparatively low sample sizes. To begin with, the computational costs rise up drastically, as the number of observations increases. Furthermore, estimated item parameters are not interpretable in practice as it is not clear if an incorrect response to an item was made due to high difficulty or if the item simply was not drawn into the test sample.

6. Discussion

The SGLDRM is a model for longitudinal data on the item level. We have demonstrated that a smooth global course of learning (CoL) can be estimated, which can often be interpreted as a reference CoL. In the SGLDRM, deviations from the global CoL are linear, so that the resulting person-level intercept and slope parameters can be estimated from two or more repeated observations. Potential applications include situations with a moderate number of repeated measurements per person. The measurement occasions can be unequally spaced. Also, items might be part of a larger pool of items. Regardless of which items are used, person-level deviations can be tested, allowing to detect deviations from the global course.

6.1 Relationship to Existing Approaches

Andersen's (1985) model expands the Rasch (1960) model by a time-dependent person parameter ϑ_{jt} for person j at measurement occasion t. Similarly, Embretson (1991) assumes a baseline person ability at the first measurement occasion to which the gain in ability is added at each occasion, so that the ability of person j at time t can be written as $\sum_{m=1}^{t} \vartheta_{jm}$. In the linear logistic test model (LLTM; Fischer, 1973) change is modeled in regard to items instead of subjects. It is assumed that the difficulty of an item is composed of different basic (cognitive) operations that are needed in order to solve an item. Change over measurement occasions is analyzed by parameterizing items with a component that indicates learning effects, that is the item difficulty changes due to experiences that were made on items before. The linear logistic model with relaxed assumptions (LLRA; Fischer, 1976) additionally allows for item-specific person abilities.

In a more recent approach, Hecht et al. (2019) use a generalized linear mixed model (GLMM) framework for their continuous-time Rasch model for dichotomous responses similar to the present approach. GLMMs are a special case of GAMMs with the smooth function being a linear function. The authors treat item effects as fixed and person effects as random, however the relation and development of the person ability is motivated and implemented differently. The person ability of person j at (discrete) time t is modeled as an autoregressive process of order 1 (AR[1]), with an underlying continuous time model. This underlying model accounts for global intercept and slope which can be translated to our model by setting the global trend $\lambda(t)$ as a degree 1 polynomial. Hecht et al. (2019) consider a person-specific intercept, however no person-specific slope. This can be interpreted as individual developments running parallel to the average development.

The bivariate normal random effect in the SGLDRM implies that the variance of latent ability at time t is given by an expression quadratic in t. This implies that a norm based purely on the SGLDRM might misestimate latent variance, even if the spline correctly estimates the latent mean curve. If norms are required and sample

size is sufficient, semiparametric continuous norming provides accurate recovery of latent variance in simulations (Lenhard et al., 2019).

In their generalized explanatory longitudinal item response model Cho et al. (2013) include the possibility to subdivide items and subjects into groups and estimate the corresponding averages for discrete time points. The authors use a GLMM framework with a logit link function to fit the model. With only one item group and one subject group the model consists of three subject-based and two item-based components. Person ability is composed of an overall intercept, an average ability for each time point, and a person-specific ability per time point in which the latter is modeled as a random effect. The item component consists of a fixed average and item-specific deviations from this average which are again taken as random effects. Similar to Fischer's models, the assumption of unidimensional items is relaxed. In order to solve an item, multiple cognitive skills are assumed to be required, differing among the items or rather item groups. This approach can be added to the SGLDRM. Fixed item group effects were technically implemented in the context of the dyscalculia example with six item clusters. Additional random item effects would have accounted for item-specific deviations.

6.2 Recommendations, Limitations and Potential Extensions

6.2.1 Item Overlap

One limitation of the fitting process is that item sets administered at different occasions need to overlap. Hence, designs with disjoint tests for each measurement occasion are not suitable for the SGLDRM, at least not without additional (equality) constraints on item parameters. Random sampling of items from a larger pool typically provides enough overlap. One benefit of this model is the possibility to compare student abilities and CoLs even when students worked on disjoint sets of items. An assumption of the SGLDRM is that item difficulty is invariant over time. Situations where difficulty drifts, say due to an intervention or schooling as usual, are not contained in the current approach. We recommend testing the item parameter invariance (e.g., Millsap, 2010).

Note, that the conditional independence assumption means that beyond current ability, no other factors govern test behavior. Especially, repeated measurement with the same items has to proceed with ample time between measurement occasions so that memory effects do not violate conditional independence. Items need to be drawn without replacement at the same measurement occasion to rule out local item dependence.

6.2.2 Time and Other Influences

The SGLDRM can index measurement occasions with fine granularity, that is, time can be continuous in the model. In the sample application to CODY data, a time of zero indicated the beginning of a training, regardless of actual date. As a consequence, two children in the same grade, say one starting the training in August, and one half a year later, start with different amounts of schooling. This implies that person-level random intercept and slope are also influenced by the schooling and should hence be interpreted with some caution. Especially the random intercept is a baseline competence at the start of the training here. In another application, time zero could represent the beginning of the school year. Alternatively, an additional (smooth) term could be added to the model, representing the amount of schooling received, potentially also for the slope. Given enough variation in the start time of the training, the two effects can be disentangled.

Note, that the conditional independence assumption means that beyond current ability, no other factors govern test behavior. Especially, repeated measurement with the same items has to proceed with ample time between measurement occasions so that memory effects do not violate conditional independence. Items need to be drawn without replacement at the same measurement occasion to rule out local item dependence.

In a correctly specified SGLDRM, which includes a correct specification of the spline and the random effects, all model parameters are consistently estimated. The larger the sample, the higher the precision of the estimation of the true global course of learning. When the random effects structure is misspecified, the model parameters still converge with growing sample size, but approach the value minimizing the Kullback-Leibler information (Heagerty & Kurland, 2001). In other words, the random effects specification, that is, the "linear deviation" part of the SGLDRM, needs to be checked. The random effects need to be modified if necessary, for example, by adding a non-smooth ability jump or drop, say after a summer break.

Assuming many measurement occasions per person, the SGLDRM can be extended to allow for more complex deviations from the global course: Introducing latent residuals, that is, $\vartheta_{jt} = \lambda(t) + \delta_j + \gamma_j t + \zeta_{jt}$ for stochastically independent ζ_{jt}, helps to study whether linear deviations are appropriate. The variance of the residuals, as well as their (mean) direction are informative. Quadratic or higher order polynomial terms could be added (or even person-level splines for the deviations). We caution that naive model checking of an SGLDRM's deviance residuals could be misleading, since binary data leads to clustered deviance residuals.

In the context of the dynamic measurement model (DMM) framework (Dumas & McNeish, 2017), trajectories are seen as nested within students who are nested within schools (Dumas et al., 2020). The GAMM framework is flexible enough to incorporate random effects for the school level with a specification parallel to that of Dumas et al. (2020). Similarly, also exponential growth curves which were found to be appropriate for a large data set of math development (Dumas et al.,

2020) are an alternative to linear deviations. Such extensions are beyond the scope of the current work, as they require a study of model identification, the parametrization of the deviations needs to be interpretable to be useful beyond graphical inspection. As dynamic IRT models are actively researched (Dumas et al., 2020), future work should also aim to include local item dependence due to repeated administration of the very same item.

6.2.3 Interpreting Correlations of Time Series

Correlations of time series are often positive if a global trend underlies all series. This has been frequently observed in the domain of macroeconomic time series (e.g., Gilbert & Meijer, 2005). In the context of educational time series, a positive global CoL and a naive calculation of correlations might suggest similarity of CoLs. Especially when (mean) time series are used to compare two or more groups, positive correlations might be spurious results of a global trend and hence potentially misleading.

The person-level deviations from the global CoL in the SGLDRM do not have this problem. They are measures of individual differences in CoL free from the spurious correlations. Once the model is calibrated, the CoL of new persons might be tracked. Then person-level slopes from the SGLDRM might be an indicator of intervention effects, even in single case research designs. Similarly, group comparisons based on group-specific deviations from the global CoL are a potential model extension addressing the issue. For this, a group mean CoL needs to be added to the model equation, represented, say, by a linear function in time.

6.2.4 Null and Perfect Scores

Maximum likelihood (ML) estimation of person ability is often not recommended, since perfect scores or null scores are a problem for the method. Here, however, it is unlikely that a person would have perfect or null scores at all measurement occasions. We recommend to remove persons with all null or all perfect scores from the data set. In both cases, the test at hand is inappropriate to track the CoL.

In case of very short assessments and very few measurement occasions, perfect and null scores might appear though the test is appropriate. The analyst might prefer a Bayesian strategy instead, for example, employ maximum a posteriori (MAP) or expected a posteriori (EAP) estimation of the two deviation factors. Note, that the likelihood ratio test breaks down, as it requires ML estimates. Alternatives for this situation include Bayes factors (Berger & Delampady, 1987).

6.2.5 Person-Level LRT Power

Our simulation results show that statistical power to detect deviations from the global course is low if the individual and global course intersect. Being able to identify students starting above average and falling behind is important in applications, but time spent testing could be less valuable than regular schooling. However, longer baseline assessments are practically feasible in many contexts and reduce uncertainty in the intercept. This in turn improves slope estimation (in analogy to simple linear regression; Atkinson & Donev, 1992, p. 38ff.) and increases power of the LRT.

6.2.6 Interpretation in Application

Due to the lack of empirical values when it comes to putting the SGLDRM into practice there is no ideal guideline for teachers to interpret the outcomes yet. However, well-thought-out guidelines are essential in order for diagnostics to be interpreted correctly by non-experts (cf. Zeuch et al., 2017). There are roughly three ways to present SGLDRM results: The global trend as well as the linear deviation can be presented graphically, facilitating a visual assessment of a student's as well as the general development. Apart from that, the SGLDRM gives three numerical scores per student for interpretation and diagnosis – an intercept, a slope, and whether or not the deviations from the global curve are (statistically) significant. In terms of the decision and judgment inferences of Hopster-den Otter et al. (2019), it would be important to provide guidelines for interpreting each outcome correctly. For better accessibility the parameter values could further be simplified by color schemes, graphical methods, and written explanations. However, one should always keep in mind, that a test performance should be interpreted in relation to other student performances and the teacher's assessments in order to meet the generalization inference (Hopster-den Otter et al., 2019). When implemented and communicated properly the SGDLRM offers valuable information on student development and can support teachers in educational diagnostics.

References

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*(1), 3–16. https://doi.org/10.1007/BF02294143

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: Population parameter estimation. *Journal of Multivariate Analysis*, *95*(1), 1–22. https://doi.org/10.1016/j.jmva.2004.07.005

Atkinson, A. C., & Donev, A. N. (1992). *Optimum experimental designs*. Clarendon.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science, 2*(3), 317–335. https://doi.org/10.1214/ss/1177013238

Binder, H., & Tutz, G. (2008). A comparison of methods for the fitting of generalized additive models. *Statistics and Computing*, *18*(1), 87–99. https://doi.org/10.1007/s11222-007-9040-0

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25. https://doi.org/10.1080/01621459.1993.10594284

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury.

Cho, S.-J., Athay, M., & Preacher, K. J. (2013). Measuring change for a multidimensional test using a generalized explanatory longitudinal item response model. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 353–381. https://doi.org/10.1111/j.2044-8317.2012.02058.x

Cho, S.-J., Brown-Schmidt, S., De Boeck, P., & Naveiras, M. (2022). Space-time modeling of intensive binary time series eye-tracking data using a generalized additive logistic regression model. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000444

Cho, S.-J., Preacher, K. J., Yaremych, H. E., Naveiras, M., Fuchs, D., & Fuchs, L. S. (2022). Modelling multilevel nonlinear treatment-by-covariate interactions in cluster randomized controlled trials using a generalized additive mixed model. *British Journal of Mathematical and Statistical Psychology*. Advance online publication. https://doi.org/10.1111/bmsp.12265

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164–185. https://doi.org/10.1111/jedm.12009

Debeer, D., Janssen, R., Buchholz, J., & Hartig, J. (2014, July 22–25). *Modeling differences in item-position effects in the PISA 2009 reading assessment within and between schools* [Conference presentation]. 79th International Meeting of the Psychometric Society. Madison, WI, United States.

Doebler, A., Doebler, P., & Holling, H. (2013). Optimal and most exact confidence intervals for person parameters in item response theory models. *Psychometrika*, *78*(1), 98–115. https://doi.org/10.1007/s11336-012-9290-4

Dumas, D., McNeish, D., & Greene, J. A. (2020). Dynamic measurement: A theoretical-psychometric paradigm for modern educational psychology. *Educational Psychologist*, *55*(2), 88–105. https://doi.org/10.1080/00461520.2020.1744150

Dumas, D. G., & McNeish, D. M. (2017). Dynamic measurement modeling: Using nonlinear growth models to estimate student learning capacity. *Educational Researcher*, *46*(6), 284–292. https://doi.org/10.3102/0013189X17725747

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495–515. https://doi.org/10.1007/BF02294487

Embretson, S. E. (1997). Structured ability models in tests designed from cognitive theory. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 223–236). Ablex.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta psychologica*, *37*(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Fischer, G. H. (1976). Some probabilistic models for measuring change. In D. de Gruijter & L. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97–110). Wiley.

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika*, *54*(4), 599–624. https://doi.org/10.1007/BF02296399

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal*, *21*(2), 449–460. https://doi.org/10.3102/00028312021002449

Gilbert, P. D., & Meijer, E. (2005). *Time series factor analysis with an application to measuring money* (Research Report 05F10). University of Groningen, Research School SOM.

Green, P., & Silverman, B. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*. Chapman & Hall. https://doi.org/10.1007/978-1-4899-4473-3

Gu, C., & Kim, Y. (2002). Penalized likelihood regression: General formulation and efficient approximation. *Canadian Journal of Statistics*, *30*(4), 619–628. https://doi.org/10.2307/3316100

Heagerty, P. J., & Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, *88*(4), 973–985. https://doi.org/10.1093/biomet/88.4.973

Hecht, M., Hardt, K., Driver, C. C., & Voelkle, M. C. (2019). Bayesian continuous-time Rasch models. *Psychological Methods*, *24*(4), 516–537. https://doi.org/10.1037/met0000205

Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2019). A general framework for the validation of embedded formative assessment. *Journal of Educational Measurement*, *56*(4), 715–732. https://doi.org/10.1111/jedm.12234

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, *40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Klauer, K. C. (1991). Exact and best confidence intervals for the ability parameter of the Rasch model. *Psychometrika*, *56*(3), 535–547. https://doi.org/10.1007/BF02294489

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*(2), 1813–1824. https://doi.org/10.1016/j.compedu.2011.02.003

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4757-4310-4

Kuhn, J.-T., Schwenk, C., Raddatz, J., Dobel, C., & Holling, H. (2018). *CODY-M 2-4: CODY-Mathetest für die 2.-4. Klasse* [CODY-M 2-4: CODY-math-test for 2.-4. Grade]. kaasa health.

Le, L. T. (2007, July 9–13). *Effects of item positions on their difficulty and discrimination: A study in PISA science data across test language and countries* [Conference presentation]. 72nd International Meeting of the Psychometric Society, Tokyo, Japan.

Lenhard, A., Lenhard, W., & Gary, S. (2019). Continuous norming of psychometric tests: A simulation study of parametric and semi-parametric approaches. *PloS ONE*, *14*(9), Article e0222279. https://doi.org/10.1371/journal.pone.0222279

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*(2), 233–245. https://doi.org/10.1007/BF02294018

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, *4*(1), 5–9. https://doi.org/10.1111/j.1750-8606.2009.00109.x

Mühling, A., Gebhardt, M., & Diehl, K. (2017). Formative Diagnostik durch die Onlineplattform LEVUMI [Formative diagnostics through the online platform LEVUMI]. *Informatik Spektrum*, *40*(6), 556–561. https://doi.org/10.1007/s00287-017-1069-7

Nagy, G., Nagengast, B., Becker, M., Rose, N., & Frey, A. (2018). Item position effects in a reading comprehension test: An IRT study of individual differences and individual correlates. *Psychological Test and Assessment Modeling*, *60*(2), 165–187.

Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. Oxford University Press.

Powell, M. J. D. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives* (Numerical Analysis Report, DAMTP 2009/ NA06). Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment*, *84*(3), 228–238. https://doi.org/10.1207/s15327752jpa8403_02

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, *6*(1), 15–32. https://doi.org/10.1214/ss/1177011926

Rost, J., & Spada, H. (1983). Die Quantifizierung von Lerneffekten anhand von Testdaten [Quantifying learning effects on the basis of test data]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *4*(1), 29–49.

Saha, A. (2016). *Bayesian analysis of item response theory and its application to longitudinal education data* (Doctoral dissertation, University of Connecticut).

Schurig, M., Jungjohann, J., & Gebhardt, M. (2019). *Handbuch für Lehrkräfte im Anwendungsbereich Verhalten und Empfinden – Lern-Verlaufs-Monitoring Levumi* [Manual for teachers in the field of application behavior and sensation – learning progress monitoring Levumi]. TU Dortmund University. https://doi.org/10.17877/DE290R-20376

Schwenk, C., Kuhn, J.-T., Gühne, D., Doebler, P., & Holling, H. (2017). Auf Goldmünzenjagd: Psychometrische Kennwerte verschiedener Scoringansätze bei computergestützter Lernverlaufsdiagnostik im Bereich Mathematik [We are going on a gold coin hunt: Psychometric properties of different scorings in computer-based progress monitoring of mathematics ability]. *Empirische Sonderpädagogik, 9*, 123–142.

Souvignier, E., Förster, N., & Salaschek, M. (2014). quop: Ein Ansatz internetbasierter Lernverlaufsdiagnostik mit Testkonzepten für Mathematik und Lesen [quop: An internet-based approach to learning progress monitoring in math and reading]. In M. Hasselhorn, W. Schneider, & U. Trautwein (Eds.), *Lernverlaufsdiagnostik* (pp. 239–256). Hogrefe.

Strathmann, A. M., & Klauer, K. J. (2012). *LVD-M 2-4: Lernverlaufsdiagnostik-Mathematik für zweite bis vierte Klassen* [LVD-M 2-4: Diagnosing the course of learning in mathematics for second to fourth grades]. Hogrefe.

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Springer. https://doi.org/10.1007/0-306-47531-6

Wang, X., Berger, J. O., & Burdick, D. S. (2013). Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics*, *7*(1), 126–153. https://doi.org/10.1214/12-AOAS608

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, *65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Taylor & Francis. https://doi.org/10.1201/9781315370279

Wood, S. N., & Scheipl, F. (2017). *gamm4: Generalized additive mixed models using 'mgcv' and 'lme4'* [R package version 0.2-5]. https://CRAN.R-project.org/package=gamm4

Wu, Q., Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2019). Predictors of individual performance changes related to item positions in PISA assessments. *Large-scale Assessments in Education*, *7*, Article 5. https://doi.org/10.1186/s40536-019-0073-6

Yoshida, T., & Naito, K. (2014). Asymptotics for penalised splines in generalised additive models. *Journal of Nonparametric Statistics*, *26*(2), 269–289. https://doi.org/10.1080/10485252.2014.899360

Zeuch, N., Förster, N., & Souvignier, E. (2017). Assessing teachers' competencies to read and interpret graphs from learning progress assessment: Results from tests and interviews. *Learning Disabilities Research & Practice*, *32*(1), 61–70. https://doi.org/10.1111/ldrp.12126

Appendix

1. Estimation Procedure

When using the statistical software R (R Core Team, 2020) to fit a GAMM, there are several functions to choose from, *mgcv* (Wood, 2017) and *gamm4* (Wood & Scheipl, 2017) being two packages to provide them. As mentioned in the main text thin plate regression splines are used as default by all of the functions. The gamm4 routines use the ML (maximum likelihood) method for parameter estimation in the case of a binomial family and *logit* as link function. A Laplace approximation to the (log-)likelihood is employed (Breslow & Clayton, 1993) and is maximized by bound optimization by quadratic approximation (BOBYQA) by default – a derivative free numerical optimization algorithm that iteratively minimizes a function by making use of quadratic models (Powell, 2009). The gamm routine of the mgcv package finds (approximate) ML estimators iteratively by penalized quasi likelihood (PQL) estimation. PQL estimation leads to asymptotic normality and efficiency of parameters estimates, including the spline (Yoshida & Naito, 2014).

While older simulation studies (Binder & Tutz, 2008) and current complex applications (Cho, Brown-Schmidt et al., 2022; Cho, Preacher et al., 2022) of GAMMs show that current estimation algorithms perform well when ample binary data is available, we have not investigated what the lowest feasible sample size for the SGLDRM is. We expect that it is hard to give general recommendations, as performance will depend on factors like number of items, number of measurement occasions, as well as distributions of latent variables and item difficulties. In case the SGLDRM is to be employed for data of less than 1000 individuals, we recommend targeted parameter recovery simulations to check for bias and RMSE of structural model parameters.

2. Latent Variance

While the model for the latent ability mean is very flexible, the time-specific latent variance is constrained by the model specification. One implication of the linear growth is that the variance of the person parameters will increase with time when the variance of the slopes is substantial: The following formula can be interpreted as indicating that the individual deviations from the global course spread the ability spectrum, though a negative correlation of the slopes and intercepts might delay this effect or reverse it initially when observation time is limited. More specifically, let ϑ_t denote a random variable for the marginal distribution of ability at time t, then

$$\mathrm{Var}(\vartheta_t) = \sigma_\delta^2 + 2t\rho\sigma_\delta\sigma_\gamma + t^2\sigma_\gamma^2. \tag{15}$$

Hence, depending on the entries of Σ, the variance of ϑ_t can increase or decrease with t, but is quadratic in t. Similarly, covariances are highly structured. The following model-implied formula for the SGLDRM's latent covariances can be used for model checking by calculating the covariance for two time points with the help of a different model, say a bivariate Rasch model. For two time points t_1 and t_2 the covariance of the latent variables is given by

$$\mathrm{Cov}(\vartheta_{t_1},\vartheta_{t_2}) = \mathrm{Cov}(\delta_j + \gamma_j t_1, \delta_j + \gamma_j t_2) = \sigma_\delta^2 + (t_1 + t_2)\rho\sigma_\delta\sigma_\gamma + t_1 t_2 \sigma_\gamma^2 \tag{16}$$

which allows for an increase or decrease of the covariance but which is bilinear in t_1 and t_2.

Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01.06
Volume 14 (2022), No. 1, 122–146
CC BY-NC-SA 4.0 Waxmann 2022

Sven Anderson, Daniel Sommerhoff, Michael Schurig, Stefan Ufer, & Markus Gebhardt

Developing Learning Progress Monitoring Tests Using Difficulty-Generating Item Characteristics: An Example for Basic Arithmetic Operations in Primary Schools

Abstract

This study investigates difficulty-generating item characteristics (DGICs) in the context of basic arithmetic operations for numbers up to 100 to illustrate their use in item-generating systems for learning progress monitoring (LPM). The fundament of the item-generating system is based on three theory-based DGICs: arithmetic operation, the necessity of crossing 10, and the number of second-term digits. The Rasch model (RM) and the linear logistic test model (LLTM) were used to estimate and predict the DGICs. The results indicate that under the LLTM approach all of the three hypothesized DGICs were significant predictors of item difficulty. Furthermore, the DGICs explain with 20% a solid part of the variance of the RM's item parameters. The identification and verification of the DGICs under the LLTM approach provide important insights into how to address the challenges in the development of future LPM tests in mathematics.

Sven Anderson, M. A. (corresponding author), ORCID: 0000-0002-2323-8543 · Dr. Michael Schurig, ORCID: 0000-0002-7708-0593, Research in Inclusive Education, Faculty of Rehabilitation Sciences, TU Dortmund University, Emil-Figge-Straße 50, 44227 Dortmund, Germany
email: sven.anderson@tu-dortmund.de
michael.schurig@tu-dortmund.de

Prof. Dr. Daniel Sommerhoff, Department of Mathematics Education, IPN – Leibniz Institute for Science and Mathematics Education, 24098 Kiel, Germany, ORCID: 0000-0002-0559-7120
email: sommerhoff@leibniz-ipn.de

Prof. Dr. Stefan Ufer, Chair of Mathematics Education,
Department of Mathematics, Faculty of Mathematics, Computer Science and Statistics, LMU Munich, Theresienstr. 39, 80333 München, Germany, ORCID: 0000-0002-3187-3459
email: ufer@math.lmu.de

Prof. Dr. Markus Gebhardt, Learning Disability Pedagogy including Inclusive Pedagogy, Faculty of Human Sciences, University of Regensburg, Sedanstraße 1, 93055 Regensburg, Germany, ORCID: 0000-0002-9122-0556
email: markus.gebhardt@paedagogik.uni-regensburg.de

Keywords
learning progress monitoring (LPM), Rasch model (RM), linear logistic test model (LLTM), item-generating rules, elementary arithmetic

Entwicklung eines Tests zur Lernverlaufsdiagnostik mit schwierigkeitsgenerierenden Merkmalen: Ein Beispiel für die Diagnose grundlegender arithmetischer Fertigkeiten in der Grundschule

Zusammenfassung
Diese Studie untersucht den Einfluss schwierigkeitsgenerierender Merkmale für die Gestaltung von Items zur Lernverlaufsdiagnostik arithmetischer Basiskompetenzen im Zahlenraum bis 100. Das System zur Itemkonstruktion basiert dabei auf drei theoriegeleiteten schwierigkeitsgenerierenden Merkmalen: der verwendeten arithmetischen Operation, der Notwendigkeit des Zehnerübergangs, und der Stellenanzahl des zweiten Terms. Zur Schätzung und Vorhersage der Itemparameter wurden das Rasch-Modell (RM) und das linear-logistische Testmodell (LLTM) verwendet. Die Ergebnisse des LLTM-Ansatzes deuten darauf hin, dass alle drei vermuteten schwierigkeitsgenerierenden Merkmale signifikante Prädiktoren für die Itemschwierigkeit sind. Basierend auf den drei schwierigkeitsgenerierenden Merkmalen konnten 20% der Varianz der Itemschwierigkeitsparameter des RM erklärt werden. Diese Studie verdeutlicht, dass die Identifikation und Prüfung schwierigkeitsgenerierender Merkmale wichtige Erkenntnisse liefern, wie Herausforderungen bei der Entwicklung zukünftiger Tests zur Lernverlaufsdiagnostik in Mathematik berücksichtigt werden können.

Schlagworte
Lernverlaufsdiagnostik, Rasch-Modell (RM), Linear-logistisches Testmodell (LLTM), itemgenerierende Regeln, elementare Arithmetik

1. Introduction

Learning progress monitoring (LPM) represents an increasingly popular approach for assessing, monitoring, and visualizing students' individual learning development with short, high-frequency and easy-to-handle tests (e.g., Fuchs et al., 2019). With the results of LPM, teachers obtain insights into students' learning processes, which facilitates the early identification of emerging learning problems. Furthermore, LPM can support teachers in evaluating the success of implemented learning programs (e.g., Deno, 2003). Recent research also concludes that LPM in the digital form, managed as a computer-based or web-based tool, offers a more economical usability in school practice (Mühling et al., 2019; Souvignier, 2018). Previous

research on the systematic use of LPM also showed positive effects on students' learning achievements (e.g., Förster & Souvignier, 2015; Stecker et al., 2005).

Despite the usefulness of LPM, the development of adequate LPM instruments is challenging, as they have to fulfil a number of specific psychometric criteria. In particular, the valid measurement of learning processes requires a large number of parallel tests that are comparable in difficulty and dimensionality (e.g., Wilbert, 2014; Wilbert & Linnemann, 2011). Parallel, but non-identical tests are needed because of memory effects as well as practice effects that may otherwise confound the probability of solving items. Accordingly, homogeneity of test difficulty is a crucial prerequisite for LPM. Learning development can only be reasonably measured if it is ensured that the parallel tests are of equal difficulty. To address this requirement, a large number of test items with the same difficulty is required, which can then be distributed systematically across parallel test versions to avoid memory or practice effects. Knowledge about the characteristics that influence the difficulty of a test item can significantly support this item generation system. Therefore, LPM instruments need not only to fulfil the classical test quality criteria (e.g., validity, reliability), but also one-dimensionality, homogeneous test difficulty, test fairness, and sensitivity to change (Wilbert, 2014). To ensure validity when selecting suitable test items, the construction of LPM instruments relies on two approaches: the robust indicator approach and the curriculum sampling approach (Fuchs, 2004). Finding robust indicators involves choosing tasks that best represent the various subskills of a specific domain or correlate strongly with them. Using the curriculum sampling approach, typical tasks are selected that represent curricular requirements over a school year. For each LPM test, students receive tasks based on the learning goals of an entire school year.

Regardless of the procedure used for developing the LPM, a theory-based systematical item design and an empirical validation of LPM using item response theory (IRT) approaches are often lacking (e.g., Wilbert & Linnemann, 2011). So far, a framework for the systematic item design of LPM measures that explicitly links information about cognitive operations needed to solve an item and their correspondence in item characteristics with LPM development has not yet been established. To generate and select items for LPM, Wilbert (2014) proposes the use of linear extensions of the Rasch model (Rasch, 1980) such as the linear logistic test model (LLTM; Fischer, 1973; Fischer & Molenaar, 1995).

For mathematics, a large part of the research in LPM has focused on computation in primary schools (e.g., Foegen et al., 2007; Hartmann & Müller, 2014; Hosp et al., 2016; Tindal, 2013). In a review of LPM in the field of mathematics computation, Christ et al. (2008) emphasize the need for a framework to achieve higher and more consistent reliability and validity of measurements. Currently, there is still a gap of implementation and evaluation of a theory-based item-generating system for LPM. For the domain of elementary arithmetic, characteristics that influence item parameters are already considered in item selection for parallel LPM tests (e.g., Hartmann & Müller, 2014; Sikora & Voß, 2017). However, there is still a lack of systematic item design based on statistical evaluations. At present, there is

little literature available on the investigation of *difficulty-generating item characteristics* (DGICs) in the area of basic mathematical competencies that can be used for systematic LPM item construction (e.g., Balt et al., 2020). The present study addresses this research gap by focusing on the design and evaluation of an item generation system, applying theory-based modeling of item complexity using LLTM to analyze items exemplarily for an LPM addition and subtraction test for numbers up to 100. The present paper is therefore structured as follows: Firstly, we give a brief introduction of rule-based item design using LLTM. Secondly, based on this, we provide an overview of the characteristics affecting the difficulty of basic arithmetic operations. Building on this, we describe a rule-based item design of arithmetic basic skills for numbers up to 100. Furthermore, we present a study in which this rule-based item design is utilized in order to assess the difficulty of three item characteristics. Finally, we discuss as to what extent rule-based item design using LLTM is useful for the construction and evaluation of LPM instruments.

2. Rule-Based Item Generation With LLTM

Rule-based item design is a method from the research area of automatic item generation (AIG; e.g., Gierl & Haladyna, 2013; Irvine & Kyllonen, 2002) and rests upon the combination of findings from cognitive psychology and psychometric theory. AIG addresses the increased need for large pools of construct valid test items due to the ongoing trend towards computer- or web-based assessments. In rule-based item generation, difficulty characteristics are identified in advance and tested empirically. Knowledge about difficulty characteristics has several advantages: the increase of construct validity, the identification of templates that can be used for time-economical item design, or the opportunity of a content valid interpretation of test results (e.g., Arendasy et al., 2006; Kubinger, 2008). Rule-based item design is also particularly well suited for computer-supported algorithmic item generation (e.g., Geerlings et al., 2011) and of particular interest for the use of LPM in teaching as the tests require many items to measure students' learning development over a longer period (e.g., Wilbert, 2014). In a first step, the cognitive operations required to solve a specific task are identified. Subsequently, item models are generated that reflect the identified cognitive operations. This procedure enables the identification and evaluation of the considered cognitive operations necessary to solve a specific task, which are used to determine DGICs and, ultimately, the predicted probability of solving an item.

The design of items and instruments for LPM is particularly difficult as the use of LPM with multiple parallel tests over a period of time places specific demands on the underlying test theory. For this, IRT models are recommended (e.g., Anderson et al., 2011; Wilbert, 2014; Wilbert & Linnemann, 2011). IRT is an approach that determines the probability of solution for each item in a test, taking into account the ability of test takers based on their response behavior (Reise et al.,

2005). The response behavior (solving or not solving a task) then depends on both the individual characteristics of the person (person parameter) and the difficulty of the task (item parameter). A frequently used IRT model is the Rasch model (RM; Rasch, 1980).

The model equation of the RM for dichotomous data is:

$$P(x_{vi} = 1|\theta_v, \sigma_i) = \frac{\exp(\theta_v - \sigma_i)}{1 + \exp(\theta_v - \sigma_i)}$$

It is assumed that the probability P that person v solves item i depends on the person's ability (θ_v) as well as the item's difficulty (σ_i). Item parameter and person parameter are estimated in the RM for dichotomous data based on manifest response behavior.

An IRT model suitable for analyzing the difficulty of cognitive operations in rule-based items is the LLTM (Fischer, 1973; Fischer & Molenaar, 1995). LLTM allows the estimation of multiple basic parameters (DGICs) that are, based on a theoretical model and prior research, assumed to drive an item's difficulty. Knowledge about DGICs enables the prediction of the difficulty for newly developed items and, by considering different DGICs, items of various difficulty (within a certain range) can be easily generated. The LLTM breaks down the item parameter into a linear combination of specific hypothesized DGICs. Each DGIC involved in an item changes its difficulty. The total difficulty is the weighted sum of the DGICs.

In the LLTM the item's difficulty can be expressed as:

$$\sigma_i = \sum_{j=1}^{m} q_{ij}\eta_j + c$$

Here q_{ij} defines the weight of DGIC j on item i, for example 0 if DGIC j is not present in item j . η_j represents the difficulty parameter corresponding to DGIC j, which is independent of the specific item I. m equals the number of DGICs included in the model. c denotes a normalization constant. The weights of the DGICs are determined on the basis of theoretical hypotheses and prior research, so that the difficulty of the individual DGICs can be evaluated (Poinstingl, 2009). Based on the estimated difficulties of the DGICs, new items with a pre-determined difficulty (based on the LLTM) can be constructed, which is why this information can well be used for automatic item generation (e.g., Embretson & Kingston, 2018).

3. Specification of an Item-Generating System for Basic Arithmetic Operations for Numbers up to 100

3.1 Research on Students' Strategies and Difficulties in Addition and Subtraction

Addition and subtraction are the two algebraic operations introduced first in primary school, often in the first two grades. Although they can be interpreted as very basic, research evidence indicates that multi-digit addition and subtraction, for example 76 + 15 or 97 – 48, is challenging for many children, especially for low-performers and children with learning disabilities (Hickendorff et al., 2019). There are not only performance problems, but also differences in learning development. In their meta-analysis, Reilly et al. (2015) showed that there have been small but stable mean gender differences in favor of boys in mathematics over the past two decades.

Generally, children starting school are already able to solve simple addition and subtraction problems, for example by using counting strategies (e.g., Baroody, 1987), and results indicate that first graders' capabilities to solve such tasks are often underestimated (see further Baroody et al., 2006; Clarke et al., 2006). However, when expanding addition and subtraction to numbers up to 20 and then 100, these initial strategies often cannot be applied anymore or are time-consuming and effortful. Thus, enabling students to solve multi-digit addition and subtraction problems and providing efficient solution strategies for such tasks are among the most important goals in initial primary school mathematics education (e.g., Karp et al., 2011; National Council of Teachers of Mathematics, 2000, 2006). However, research has underlined that corresponding tasks, as for instance 14 + 23 or 90 – 23, vary considerably in their difficulty. This is attributed to students' varying use of different solution strategies, for example, number-based strategies or digit-based strategies (see Hickendorff et al., 2019, for an elaborate framework of correspondent strategies) as well as to different task characteristics (see also Daroczy et al., 2015, for a more general review on factors contributing to word problems' difficulty). For the construction of LPM items, the task characteristics are of particular interest, as they can easily be used for a rule-based item construction. In contrast, it is difficult to create items based on different solution strategies as students decide on their individual solution strategy during the problem-solving process, and even if tasks trigger a certain strategy for some learners, for example 39 + 47, they may not equally trigger this strategy for other learners in the same way. In this regard, research has repeatedly underlined students' heterogeneity, regarding both their individual capabilities for solving addition and subtraction tasks and (obviously related) their use of various solution strategies (see Baroody et al., 2006; Benz, 2005; Cooper et al., 1996; Verschaffel et al., 2007).

There are multiple characteristics of multi-digit addition and subtraction tasks up to 100 that can be interpreted in the sense of DGICs. Often mentioned, funda-

mental DGICs are (a) the *arithmetic operation* itself, that is addition and subtraction (e.g., Beishuizen, 1993; Cooper et al., 1996; Selter, 2001), (b) the necessity of *crossing 10,* that is the necessity to coordinate between ones and 10s as different places in the place-value notation during the operation as the addition/subtraction of the ones is not within the range of 0 to 9 (e.g., Beishuizen et al., 1997; Cooper et al., 1996; Fiori & Zuccheri, 2005), as well as (c) the *number of second-term digits* in the operation, that is either a one- or two-digit addend or subtrahend (e.g., Cooper et al., 1996). Multiple example items are given below.

A study by Benz (2005) revealed great differences in students' solution rates between tasks with different DGICs. First of all, her data revealed significant differences between addition and subtraction tasks, especially towards the end of the second grade and when presented without contextualization (i.e., no word problems). Based on her data, she suspects that differences between the solution rates for tasks with either operation are lower when informal strategies are applied, again highlighting the impact of different solution strategies on item difficulty. Moreover, her data also underlines that tasks with the necessity of crossing 10, for example 27 + 35 that requires the calculation of 7 + 5 = 12 and thus leads to a "crossing of 10" based on the addition of the ones, were less likely to be solved than those without that necessity. Finally, tasks including a second term with two digits were generally less likely to be solved for both arithmetic operations in comparison to tasks with one-digit second terms.

Beyond these three DGICs, research gives (mostly theoretical or qualitative) evidence of multiple other DGICs, implying, for example, that the addition and subtraction of multiples of 10s, for example, 30 + 20 or 50 – 30, is generally easier than tasks as for instance 32 + 24 or 57 – 33, which are comparable regarding the three DGICs pointed out above (i.e., no necessity of crossing 10 and including second terms with two digits) or that the addition resulting in a multiple of 10, for example, 53 + 7 or 56 – 6, is easier than other addition tasks with the necessity of crossing 10 even further, for example, 18 + 3. Moreover, the task 51 – 3 would be classified as being easier than 50 – 30 based on the DGIC that tasks with second terms with only one digit are easier than tasks with two second-term digits but classified as more difficult than 50 – 30 due to the necessity of crossing 10. These examples underline that concentrating on few, rather general DGICs can help to structure the difficulty of tasks and may allow for an easy item classification – and in the context of LPM easily comprehensible feedback for teachers and students. However, to accurately determine the difficulty of items, for example in research contexts, (a) more DGICs would have to be used and (b) in particular not only additively, but also including interactions. This would lead to a higher explained variance and thus a better classification of the items' difficulties, which would be favorable from a research perspective. However, giving teachers feedback on students' skills based on (e.g.) 10 DGICs and higher-level interactions appears unrealistic. Thus, using few central DGICs to explain a relevant portion of variance in order to give teachers a good first indication on how to support students seems indicated.

Overall, although DGICs can be used to approximate the difficulty of tasks and are useful in task construction, empirical evidence of the exact magnitude of the DGICs of addition and subtraction tasks is still lacking, as there are mostly only general tendencies regarding the higher importance of some DGICs over others.

3.2 Item Design of the Current Study

Based on the findings from prior mathematics education research, DGICs of addition and subtraction problems for numbers up to 100 can be identified. For the current study, three important characteristics were used to model the difficulty of the items: the *arithmetic operation* (addition versus subtraction; DGIC 1), the *necessity of crossing 10* (no crossing versus with crossing; DGIC 2), and the *number of second-term digits* (one-digit number versus two-digit numbers; DGIC 3). Therefore, the items used in this study consisted of mixed addition and subtraction problems including addition tasks with two one-digit numbers (e.g., 7 + 4), a one-digit and a two-digit number (e.g., 24 + 5), or the addition of 2 two-digit numbers (e.g., 47 + 26), and subtraction tasks with a one-digit or a two-digit number (e.g., 16 – 7) or 2 two-digit numbers (e.g., 78 – 49). For the tasks, the value of the individual digits varied between 0 and 9. For all problems, a fill-in-the-blank format was used, with the correct answer always being missing. Furthermore, there were tasks with and without crossing the tens barrier as well as tasks to the next 10.

Table 1 exemplifies the design matrix (see also Appendix Table A1) for the computation test with three exemplary items and the three dichotomously scored DGICs arithmetic operation (DGIC 1), necessity of crossing 10 (DGIC 2), and number of second-term digits (DGIC 3). The item number refers to the position of the item in the test. The 0s and 1s are the specified weights of the DGICs for each item. For the DGIC 1 (arithmetic operation), a weight of 0 represents addition, a weight of 1 represents subtraction. A weight of 0 for the DGIC 2 (necessity of crossing 10) means that crossing 10 is not necessary for solving the item, whereas with a weight of 1 it is. The items with a weight of 0 for the DGIC 3 (number of second-term digits) have a one-digit second term, a weight of 1 shows that the second term has two digits. For example, Item 5 (56 + 3 = 59) neither includes a subtraction nor the necessity of crossing 10 nor a two-digit subtrahend. Thus, all weights for this item are 0. In contrast, Item 7 (43 + 9 = 52) includes the crossing of a 10. Consequently, the weight matrix includes a 1 for the crossing of a 10 for this item, otherwise 0s.

Table 1: Exemplary Design Matrix for Items 5, 7, and 8

Item number	DGIC 1	DGIC 2	DGIC 3	Item
5	0	0	0	56 + 3 = 59
7	0	1	0	43 + 9 = 52
8	1	1	1	32 – 17 = 15

Note. DGIC = difficulty-generating item characteristic.

Based to the three DGICs of the design matrix in Table 1, a pool of 80 items was generated. The three DGICs were varied within the item design process so that all possible combinations were adequately represented in the item pool (with three DGICs, eight different combinations were considered; see the design matrix for the first 41 items included in the further analyses in the Appendix Table A1).

4. Research Questions

To test the construct validity of the model, the characteristics of the items in terms of DGICs and the statistical properties of the items were associated. The hypothetically assumed rule-based difficulty-generating item characteristics have to explain the RM's item difficulty parameters. In addition, there is the assumption that differences in gender will become apparent. We followed two central research questions:

Research Question 1: Do the three difficulty-generating item characteristics (DGICs) that were used to create the test items influence their difficulty?

In reference to Question 1, we hypothesized that the three identified DGICs have significant influence on the difficulty of the items as indicated by prior mathematics education research (e.g., Benz, 2005).

Research Question 2: How much variance is explained by the LLTM model with the three DGICs?

With regard to Question 2, we hypothesized that the three identified DGICs have a significant impact in variance explanation of the RM's item difficulty parameters. However, prior research has underlined the existence of multiple other DGICs and the importance of their interrelation for an exact estimation of item difficulty. Moreover, research has underlined that item characteristics are only one aspect influencing item difficulty and that students' solving strategies are also important for item difficulty. This led us to hypothesize that the variance explanation might be somewhere between 10% (i.e., clearly above 0) and 50% (i.e., still leaving much room for other DGICs, strategies, and other variables for variance explanation).

5. Method

5.1 Participants

The sample consisted of N = 591 students (M_{age} = 8.80, SD = 0.76; boys: 52.28%, girls: 47.72%; 9.14% of the students with special educational needs) in Grade 2 (n = 205; 34.69%) and 3 (n = 386; 65.31%) from 13 German primary schools in North Rhine-Westphalia (28 classes). All students who did not reach at least 5 items were excluded from the analysis as reasonable participation in the test could not be guaranteed and students might have been unwilling to participate. Overall, n = 10 participants (1.69%) were excluded.

5.2 Measures

Prior to the study, a first test with fixed order was created based on a pool of 80 items, which included all eight possible combinations of the three DGICs. The fixed order was chosen to safeguard that participants worked on each combination regularly. The test was implemented on the online platform Levumi (www.levumi.de; Gebhardt et al., 2016; Mühling et al., 2019). Between October 2019 and January 2020, the students were tested in their classrooms in groups of 10. To perform the test, each participating student used a tablet device. The test time was 5 minutes. At the beginning, the students received a brief technical instruction, an example item was solved, and the students were given the opportunity to ask the test supervisor questions. Students could start the test themselves by clicking on the start button. After the test time had expired, the test ended automatically. This way the test cannot be considered a power test due to the restricted time but is rather a long speed test.

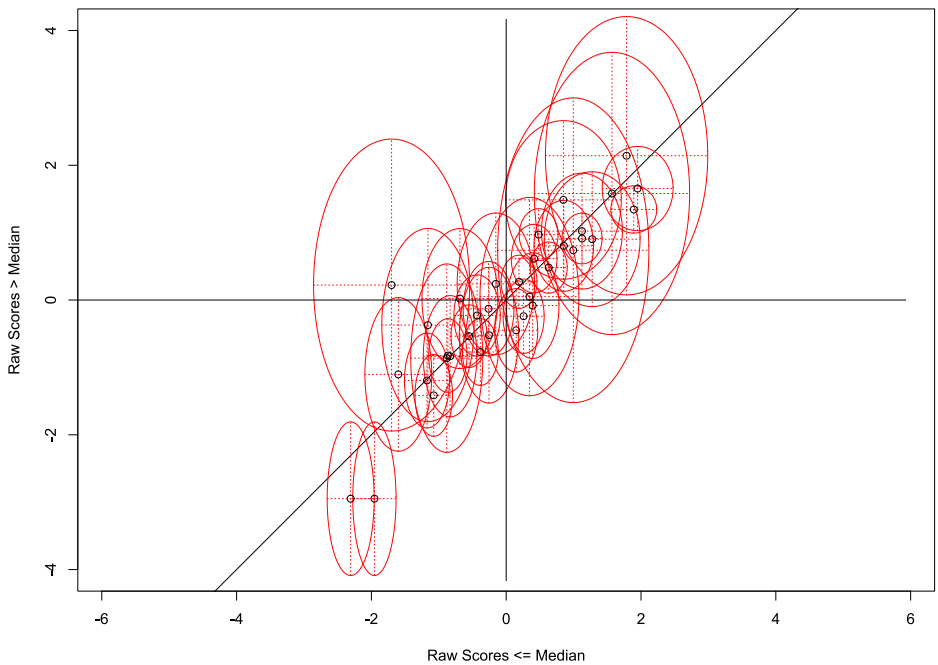
5.3 Statistical Analyses

The participants were not expected to solve nearly as many items as the test contained in the given 5-minute testing time as the high number of items was created to also account for exceptionally good students. Accordingly, most students did not answer all items. The mean number of items answered was 18.48 (SD = 9.59, min = 1, max = 67). For the analyses, only items that were answered by at least 20 students were used to ensure an adequate model estimation. Thus, 41 items were included in the analyses (see the descriptive statistics in the Appendix Table A2). RM is a prerequisite to conduct LLTM for item analysis and model comparisons. The following computations were all done with the package eRM (Mair & Hatzinger, 2007). Thus, a dichotomous RM was fitted. The occurrence of missings is calculated and pooled separately for each subgroup (by NA structure) in eRM

(Mair & Hatzinger, 2007). In our understanding, items that are answered more often should not be weighted more heavily that way. To evaluate the items of the test, the item fit was evaluated by infit and outfit statistics. Items with infit and outfit values below 0.5 and above 1.5 (Wright & Linacre, 1994) were excluded. This resulted in an additional item exclusion of five items, namely Items 32, 34, 36, 38, and 41 (see Appendix Table A2). Item 33 was also excluded due to inappropriate response patterns within subgroups, leaving 581 cases and 35 items. The reliability of the resulting weighted maximum likelihood estimation (WLE) person parameters reached .82.

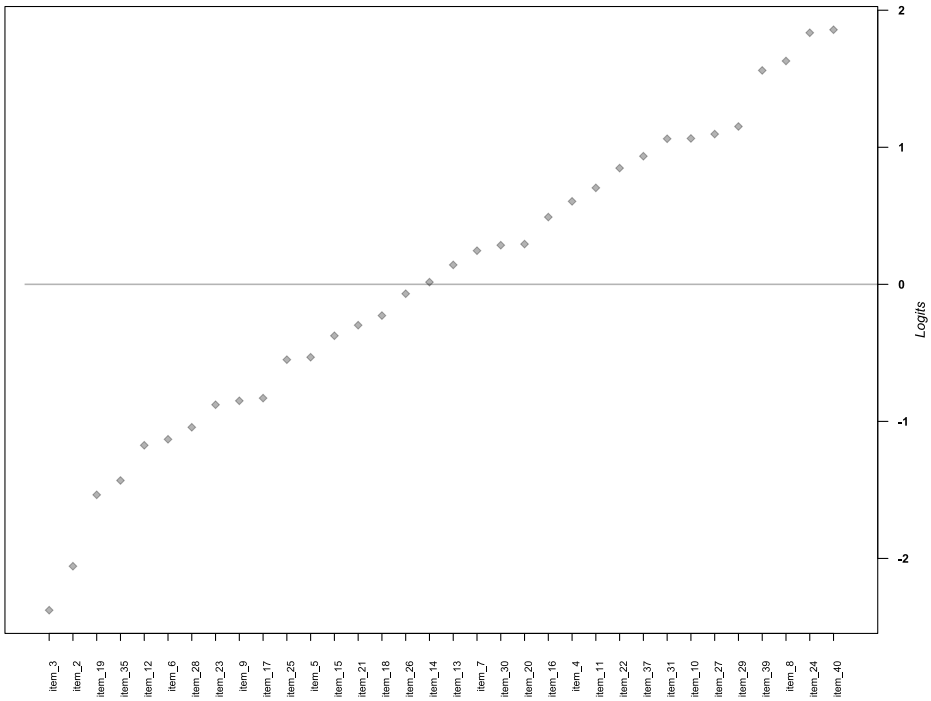
The package *eRM* employs a conditional maximum likelihood approach for parameter estimation. There are different ways of testing the appropriateness of an RM. In this paper, the likelihood ratio test (LRT; Andersen, 1973; Kubinger, 1989), the graphical model test (Rasch, 1980), and a Wald-type test (Glas & Verhelst, 1995) are applied. The LRT is a global test of model fit that checks the assumption of equality of the item parameters between subpopulations. A significant result were differences in the global item parameters between groups. An LRT with a median split resulted in a LRT-value of 32.24 (df = 34, p = .554), an LRT with a mean split in 44.80 (df = 34, p = .102), thus showing no significant differences in item parameters between these subgroups. In the graphical model test, the estimated item parameters of the median-split subgroups are compared against each other (Figure 1). The ellipses indicate the size of the confidence intervals.

Figure 1: Graphical Model Check of Items for Addition and Subtraction of Numbers up to 100



Though the confidence intervals vary in magnitude due to sparse data because of the item location in the test, nearly all item's ellipses are split by the bisecting angle, indicating equal item parameters between the groups. The Wald test specifies the analysis on the item level. Here, all items are analyzed individually. The Wald tests (split criteria: median; mean) showed significant differences in the difficulties of two items in *median*-split subgroups (Item 8 and Item 11) and two items (Item 3 and Item 14) in a *mean*-split subgroup. The items were kept because the infit and outfit values of the item fit statistics showed that they do not have to be deemed detrimental. The final item parameters are given in Figure 2 to illustrate the spread in difficulty. The item numbers in Figure 2 correspond to the sequence in the test. For example, Item 3 is the third item that the students worked on in the test. As additional analyses, differential item functioning (DIF; Holland & Wainer, 1993; see Figure A1 and Figure A2 in the Appendix) was tested for gender and grade. However, no gender effects (see Reilly et al., 2015) could be found in these items. DIF revealed only one item (Item 3) with a relatively high effect of gender, which however was still insignificant. DIF analyses for grade resulted in multiple non-linear significant effects, which, however, were expected based on prior results from mathematics education, which also showed non-linear relationships between second and third grade (e.g., Benz, 2005).

Figure 2: Item Difficulties of Items for Addition and Subtraction of Numbers up to 100



The LLTM was estimated stepwise. In a first step, a model (Model 1) was estimated with DGIC 1 (arithmetic operation). In a second step, another model (Model 2) was estimated that contained DGIC 2 (necessity of crossing 10) in addition to DGIC 1 (arithmetic operation). Finally, DGIC 3 (number of second-term digits) was used to estimate a third model (Model 3) that included all three DGICs identified in Section 3. Overall, the 35 items included in the analyses contained 17 items that showed the characteristic of DGIC 1 (arithmetic operation), 16 of DGIC 2 (necessity of crossing 10), and 17 of DGIC 3 (number of second-term digits). The stepwise estimation with three models was done to (a) illustrate the functionality of LLTMs, and (b) evaluate the additional benefit of adding the DGICs to the model. In the LLTM, the item parameters are regressed on the DGICs item parameters for each item. This contributes to unexplained variance based on the lower degrees of freedom (*df*; Poinstingl, 2009). These item parameters are not random but were fixed before the estimation. To determine the item difficulties, a design matrix with the weights of the three DGICs had to be established. Although the elements of such a design matrix can, in principle, be any positive number, also including fractions (Poinstingl, 2009), we assigned dichotomous values (see Table 1). As this paper sets out to illustrate the applicability of LLTM for LPMs, only additive components were addressed in this study for reasons of simplicity.

6. Results

6.1 Research Question 1: Influence of Identified DGICs on Item Difficulty

To determine the influence of the three identified DGICs and thus address Research Question 1, the DGICs are added stepwise. In a first run, DGIC 1 (arithmetic operation) is implemented as the only parameter (Model 1). In Model 2 and 3, the weight vectors of DGIC 2 (necessity of crossing 10) and DGIC 3 (number of second-term digits) are added to the design matrix. The additive procedure of the establishment of the item difficulties within the concurrent models is exemplified in Table 2.

As the Items 5 and 7 do not include the operation of subtraction, their item parameters are fixed to 0.00 in Model 1. Generally, items in LLTM that do not have a DGIC weight unequal 0 are fixed to difficulty 0.00 as only the presence of DGICs can affect an item difficulty. Here, the easiest combination of characteristics in terms of content is selected as the baseline so that all additions of characteristics that imply difficulty lead to item parameters > 0.

Table 2: Composition of Item Parameters in LLTM Models

Item number	Item	Design matrix assignment	Item difficulty			
			RM	Model 1	Model 2	Model 3
5	56 + 3	0,0,0	−0.53 [−0.77, −0.30]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
7	43 + 9	0,1,0	0.25 [0.02, 0.47]	0.00 [0.00, 0.00]	0.75 [0.65, 0.85]	0.94 [0.83, 1.05]
8	32 − 15	1,1,1	1.63 [1.40, 1.86]	0.52 [0.42, 0.61]	1.46 [1.30, 1.62]	2.54 [2.32, 2.75]

Note. Values in square brackets indicate the 95% confidence interval for each parameter estimation. The parameter estimates are given in logits.

As Item 8 is a subtraction item and thus holds characteristic DGIC 1, the item difficulty is fixed to the global parameter of Model 1, where only this parameter is estimated (see Table 3). In Model 3, when all difficulty parameters are introduced, the item difficulty of Item 5 is still fixed to 0.00, as its design matrix assignment is still 0,0,0. However, the difficulty of Item 8, which combines all three DGICs, is estimated to be 2.54, which reflects an additive combination of all three DGICs (Baghaei & Kubinger, 2015) given in Table 3, 0.75 + 0.94 + 0.84 ≈ 2.54. All items with the same hypothesized DGICs share the same difficulty, as difficulty is only estimated based on the DGICs. In comparison, the RM allows a free estimation of item difficulty for each item, thus allowing a better prediction of empirical data than the LLTM, however leading to less insights on why the items have a certain difficulty. The results of the comparison of the concurrent models (Table 3) show that the more of the three hypothesized DGICs are implemented in the models, the higher the quality of the model with regard to the likelihood, controlled by likelihood ratio tests, and the Akaike Information Criterion (AIC) as well as the Bayesian Information Criterion (BIC).

Table 3 highlights that the introduction of further DGICs causes a more desirable likelihood and better information criteria. Model 2 holds a significantly greater likelihood than Model 1 ($LR_{\text{Model 1 to Model 2}} = 219.19$, $df = 1$, $p < .001$) and Model 3 holds a significantly greater likelihood than Model 2 ($LR_{\text{Model 2 to Model 3}} = 260.5$, $df = 1$, $p < .001$). The RM still holds the most desirable likelihood ($LR_{\text{Model 3 to RM}} = 882.87$, $df = 31$, $p < .001$) and information criteria. This is, however, to be expected as each item's parameter is introduced as an independent random parameter in the model, while the parameters in the LLTMs are treated as fixed based on the linear combination of the DGICs. This means that the parameter estimation for the LLTMs is more parsimonious, which can be illustrated with an example: In this rather easy exemplary calculation, only 35 items are used, resulting in 34 parameters ($i - 1$) to be estimated in the RM. In the LLTMs, only 1 to 3 parameters are calculated (m).

Table 3: Log-Likelihood of the Models and Respective Difficulty Generating Item Characteristics

Model	CLL	Number of parameters	BIC	AIC	Estimate		
					DGIC 1 (η_1)	DGIC 2 (η_2)	DGIC 3 (η_3)
1	−4087.79	1	8181.94	8177.58	0.52 [0.42, 0.61]		
2	−3978.23	2	7969.19	7960.46	0.71 [0.60, 0.81]	0.75 [0.65, 0.85]	
3	−3847.98	3	7715.05	7701.96	0.75 [0.65, 0.86]	0.94 [0.83, 1.05]	0.84 [0.74, 0.95]
RM	−3406.54	34	7158.68	7023.38	–	–	–

Note. CLL = conditional log-likelihood; BIC = Bayesian Information Criterion; AIC = Akaike Information Criterion; DGIC = difficulty-generating item characteristic. Values in square brackets indicate the 95% confidence interval for each parameter estimation. The parameter estimates are given in logits.

Table 3 shows that all item parameters hold significant explanatory power, even when controlled for one another (identifiable for Model 3 by the confidence intervals of the DGICs in the square brackets). For DGIC 1 (arithmetic operation), subtraction is about three quarters of a logit more difficult than addition in Model 3 and for DGIC 2 (necessity of crossing 10), the crossing of a 10 is almost a full logit more difficult than not crossing a 10. For DGIC 3 (number of second-term digits), if there are two digits in the second term, it is about four fifth of a logit more difficult than if there is only a single digit in the second term.

6.2 Research Question 2: Impact of the Three DGICs in Variance Explanation of the RM's Item Difficulty Parameters

To answer Research Question 2 regarding the variance explanation of the three identified DGICs, correlations are calculated across the models. The appropriateness of the item parameters themselves can be assessed by the correlations of the item parameters between the models, especially when compared to the RM. The correlations are given in Table 4.

The item parameters of Model 1 and the item parameters of the RM reach an insignificant correlation of $r = .11$, which, however, is expected, since a model using only 1 *df* (thus allowing only two different values, 0.00 and 0.52, as item difficulty) is compared to a model with 34 *df*s. Yet, the correlation reaches a highly significant value of $r = .41$ for the item parameters of Model 3 and the RM. This can be interpreted as 20% explained variance in the item difficulties due to the introduced parameters. It underlines that the introduction of only three parameters (a) explains 20% of the item parameters in the Rasch model with 34 parameters, (b) can be deemed successful, and (c) confirms their importance for the general item difficul-

ty. However, 20% also shows that there is room for more DGICs, interactions, as well as other variables to account for unexplained variance, which would be valuable for further item generation.

Table 4: Means, Standard Deviations, and Correlations of the Item Parameters with Confidence Intervals

Model	*M*	*SD*	Model 1	Model 2	Model 3
Model 1	−0.25	0.26			
Model 2	−0.69	0.46	.59** [.35, .76]		
Model 3	−1.17	0.65	.44** [.15, .66]	.76** [.60, .87]	
RM	−0.00	1.08	.11 [−.21, .40]	.23 [−.08, .50]	.41** [.12, .64]

Note. Values in square brackets indicate the 95% confidence interval for each correlation.
*$p < .05$. **$p < .01$.

7. Discussion

In the current study, we developed an item-generating system for basic arithmetic operations for numbers up to 100 based on three DGICs. As expected, all three DGICs (arithmetic operation, necessity of crossing 10, number of second-term digits) significantly contributed to the prediction of item difficulty parameters in the LLTM. In addition, the three DGICs contribute substantially to the variance explanation of the RM's item difficulty parameters (Research Question 2). Previous assumptions about the difficulty characteristics of multi-digit addition and subtraction tasks up to 100 were thus confirmed. Gender-specific differences could not be identified. While we deliberately chose only three DGICs in the context of our research, as this relatively easy model allows easy-to-comprehend feedback for teachers, it would be highly valuable from a research perspective to consider additional DGICs and their interactions in order to achieve a higher variance explanation, thus allowing a better determination of item difficulty. Still, based on different solution strategies (e.g., counting strategies; see Hickendorff et al., 2019), the DGICs will anyhow only explain a certain share of the variance. Investigating students' currently used strategies has the potential to generate further information about the way the items are solved. This information can be used to derive additional variance explanation on the individual level. According to Wilbert (2014), the identification of DGICs is of interest in several aspects:

First, the information on the DGICs can contribute to the validation and further development of psychological theories. DGIC analyses can help to refine the initial theoretical assumptions about DGICs and can lead to a further development of appropriate theoretical models. In this context, findings about the relative diffi-

culties of DGICs can provide a basis for interpreting the importance of DGICs. Furthermore, more DIF analyses could show whether DGICs apply equally to different samples, making it possible to analyze which groups have particular difficulties in not only performing tasks generally, but also regarding specific hurdles in terms of DGICs. Especially in the field of inclusive education, this could be interesting for students with and without special educational needs or between groups of different special educational needs.

Second, knowledge about the DGICs is particularly useful for the construction of parallel test versions that can be used to monitor learning progress. For LPM, which requires frequent measurements over a long period of time, many different test items with known and comparable difficulty are needed. For this purpose, linear extensions of the RM such as LLTM are useful to identify DGICs, which can then be used as templates to generate items of the required difficulty without additional effort. With regard to previous research on LPM instruments in the field of computation, Christ et al. (2008) conclude that the use of more systematic sampling and item construction would improve the quality of LPM. To date, there has been little literature on the investigation on DGICs in the field of basic mathematical skills and for developing LPM (e.g., Balt et al., 2020; Ehlert et al., 2013). However, for the theory-based construction of test items for LPM, it is necessary to obtain valid information about which characteristics influence the difficulty of an item. This will support the development of knowledge about robust indicators. These enable the construction of LPM test procedures that can be used across classes and independently of graded curricula. Such curriculum-independent LPM tests are then also suitable for use in inclusive education, where often not all of the students are taught according to the same curriculum (Gebhardt et al., 2016).

Third, the identification of DGICs is also useful for the fine-grained analysis of students' learning development as it is possible to clearly highlight important aspects for further planning of individualized interventions. This allows the provision of formative feedback that has a concrete impact on the planning of future teacher interventions, rather than giving teachers only mean values that only allow for very general pedagogical conclusions. Previous research has shown that many teachers struggle with interpreting LPM results (e.g., Espin et al., 2017; Stecker, 2017). By identifying a small number of DGICs that serve as a framework for item construction, it is possible to make it easier for teachers to interpret results and thus establish a basis for designing appropriate interventions. This type of qualitative feedback can provide teachers with concise information about the domains in which a student is still struggling. For the domain of multi-digit addition and subtraction tasks up to 100 that we addressed in this study, more specific feedback could consist of informing the teacher that students have already confidently mastered addition tasks with two-digit summands but have not yet mastered crossing 10. With computer-based and web-based LPM tools, it is possible to quickly provide such qualitative feedback.

8. Conclusion

As noted at the beginning of this paper, the formulation and verification of a set of DGICs based on educational research results using LLTM enables the investigation of construct validity and the examination of the influence of the underlying DGICs on the difficulty of items. The results are therefore relevant for the establishment and further development of psychological theories, for test development, and educational practice. The explanation of roughly 20% of the variance in 34 item parameters by three DGICs alone is a solid result. Yet, of course, this also shows that more DGICs are needed to achieve stronger variance explanation of item difficulty. So far, we do not control for a certain type of strategy to solve the items but with the DGICs, we imply a particular type of strategy (calculating using structures of the decimal system). Differential profiles of item difficulties might exist when controlling for strategies.

In the present study, it is not possible to examine additional DGICs due to the systematically created item pool. In addition, it has not yet been possible to control adequately for a greater range of background characteristics (e.g., dyscalculia) that might have a potential influence on item difficulty. One important future question is the influence of sequence effects due to the limited processing time of LPM. The DGICs robustness against time was only tested in a limited fashion, too. By now, a fixed item order was evaluated. In the future, this has to be extended for randomized item orders. In future studies, the processing time of the items or the test has to be considered as an item characteristic or function of a difficulty generating characteristic. Rasch Poisson counts models seem to be a promising alternative (Baghaei & Doebler, 2019). Such models can also take into account the nested structure of the data which has not been considered in this study. The nested structure may also have an effect on the standard error (*SE*) of the item parameters which has not been accounted for now. Cognitive diagnostic modeling is another alternative for detailed modeling when deterministic information of the component matrices is available (Ravand & Robitzsch, 2015). In addition, the question should be investigated, whether the parameters causing difficulties primarily affect students in the lower grades and students with special educational needs.

In summary, the study emphasizes the usefulness of the LLTM for the identification and evaluation of difficulty-generating characteristics for an item-generating system for basic arithmetic operations. With the identified and verified three DGICs, it is possible to easily develop items of varying difficulty for parallel test versions, a basic prerequisite for the effective use of LTMs. The three DGICs provide a basis for computer-based LPM and thus enable an easy to use, practical, and easy to interpret application.

Acknowledgements

The current research is part of the project Dortmund Profile for Inclusion-Oriented Learning and Teacher Training – DoProfiL. DoProfiL is part of the *Qualitätsoffensive Lehrerbildung*, a joint initiative of the Federal Government and the Länder, which aims to improve the quality of teacher training. The program is funded by the Federal Ministry of Education and Research [Bundesministerium für Bildung und Forschung; Förderkennzeichen 01JA1930]. The authors are responsible for the content of this publication.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch Model. *Psychometrika*, *38*(1), 123–140. https://doi.org/10.1007/BF02291180

Anderson, D., Lai, C.-F., Alonzo, J., & Tindal, G. (2011). Examining a grade-level math CBM designed for persistently low-performing students. *Educational Assessment*, *16*(1), 15–34. https://doi.org/10.1080/10627197.2011.551084

Arendasy, M. E., Sommer, M., Gittler, G., & Hergovich, A. (2006). Automatic generation of quantitative reasoning items – A pilot study. *Journal of Individual Differences*, *27*(1), 2–14. https://doi.org/10.1027/1614-0001.27.1.2

Baghaei, P., & Doebler, P. (2019). Introduction to the Rasch Poisson counts model: An R tutorial. *Psychological Reports*, *122*(5), 1967–1994. https://doi.org/10.1177/0033294118797577

Baghaei, P., & Kubinger, K. D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research, and Evaluation, 20*, Article 1. https://doi.org/10.7275/8f33-hz58

Balt, M., Fritz, A., & Ehlert, A. (2020). Insights into first grade students' development of conceptual numerical understanding as drawn from progression-based assessments. *Frontiers in Education*, *5*, Article 80. https://doi.org/10.3389/feduc.2020.00080

Baroody, A. J. (1987). The development of counting strategies for single-digit addition. *Journal for Research in Mathematics Education*, *18*(2), 141–157. https://doi.org/10.2307/749248

Baroody, A. J., Iung Lai, M., & Mix, K. S. (2006). The development of young children's early number and operation sense and its implications for early childhood education. In B. Spodek & O. N. Saracho (Eds.), *Handbook of research on the education of young children* (2nd ed., pp. 187–221). Lawrence Erlbaum.

Beishuizen, M. (1993). Mental strategies and materials or models for addition and subtraction up to 100 in Dutch second grades. *Journal for Research in Mathematics Education*, *24*(4), 294–323. https://doi.org/10.2307/749464

Beishuizen, M., Van Putten, C. M., & Van Mulken, F. (1997). Mental arithmetic and strategy use with indirect number problems up to one hundred. *Learning and Instruction*, *7*(1), 87–106. https://doi.org/10.1016/S0959-4752(96)00012-6

Benz, C. (2005). *Erfolgsquoten, Rechenmethoden, Lösungswege und Fehler von Schülerinnen und Schülern bei Aufgaben zur Addition und Subtraktion im Zahlenraum bis 100* [Students' success rates, calculation methods, solutions and mistakes in addition and subtraction tasks in the range up to 100]. Franzbecker.

Christ, T. J., Scullin, S., Tolbize, A., & Jiban, C. L. (2008). Implications of recent research: Curriculum-based measurement of math computation. *Assessment for Effective Intervention*, *33*(4), 198–205. https://doi.org/10.1177/1534508407313480

Clarke, B., Cheeseman, J., & Clarke, D. (2006). The mathematical knowledge and understanding young children bring to school. *Mathematics Education Research Journal*, *18*(1), 78–102. https://doi.org/10.1007/BF03217430

Cooper, T. J., Heirdsfield, A., & Irons, C. J. (1996). Children's mental strategies for addition and subtraction word problems. In J. T. Mulligan & M. C. Mitchelmore (Eds.), *Children's number learning* (pp. 147–162). Australian Association of Mathematics Teachers and Mathematics Education Research Group of Australasia.

Daroczy, G., Wolska, M., Meurers, W. D., & Nuerk, H.-C. (2015). Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology*, *6*, Article 348. https://doi.org/10.3389/fpsyg.2015.00348

Deno, S. L. (2003). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention*, *28*(3–4), 3–12. https://doi.org/10.1177/073724770302800302

Ehlert, A., Fritz, A., Arndt, D., & Leutner, D. (2013). Arithmetische Basiskompetenzen von Schülerinnen und Schülern in den Klassen 5 bis 7 der Sekundarstufe [Basic arithmetic competencies of secondary school students from Grades 5 to 7]. *Journal für Mathematik-Didaktik*, *34*(2), 237–263. https://doi.org/10.1007/s13138-013-0055-0

Embretson, S. E., & Kingston, N. M. (2018). Automatic item generation: a more efficient process for developing mathematics achievement items? *Journal of Educational Measurement*, *55*(1), 112–131. https://doi.org/10.1111/jedm.12166

Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., & Rooij, M. de (2017). Data-based decision-making: Developing a method for capturing teachers' understanding of CBM graphs. *Learning Disabilities Research & Practice*, *32*(1), 8–21. https://doi.org/10.1111/ldrp.12123

Fiori, C., & Zuccheri, L. (2005). An experimental research on error patterns in written subtraction. *Educational Studies in Mathematics*, *60*(3), 323–331. https://doi.org/10.1007/s10649-005-7530-6

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models*. Springer. https://doi.org/10.1007/978-1-4612-4230-7

Foegen, A., Jiban, C., & Deno, S. L. (2007). Progress monitoring measures in mathematics. *The Journal of Special Education*, *41*(2), 121–139. https://doi.org/10.1177/00224669070410020101

Förster, N., & Souvignier, E. (2015). Effects of providing teachers with information about their students' reading progress. *School Psychology Review*, *44*(1), 60–75. https://doi.org/10.17105/SPR44-1.60-75

Fuchs, L. S. (2004). The past, present and future of curriculum-based measurement research. *School Psychology Review*, *33*(2), 188–192. https://doi.org/10.1080/02796015.2004.12086241

Fuchs, L. S., Fuchs, D., Seethaler, P. M., & Zhu, N. (2019). Three frameworks for assessing responsiveness to instruction as a means of identifying mathematical learning disabilities. In A. Fritz, V. G. Haase, & P. Räsänen (Eds.), *International handbook of mathematical learning difficulties* (pp. 669–681). Springer. https://doi.org/10.1007/978-3-319-97148-3_39

Gebhardt, M., Diehl, K., & Mühling, A. (2016). Online-Lernverlaufsmessung für alle SchülerInnen in inklusiven Klassen. www.LEVUMI.de. [Online learning progress monitoring for all students in inclusive classes. www.LEVUMI.de]. *Zeitschrift für Heilpädagogik, 67*(10), 444–454.

Geerlings, H., Glas, C. A. W., & Van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, *76*(2), 337–359. https://doi.org/10.1007/S11336-011-9204-X

Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge. https://doi.org/10.4324/9780203803912

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 69–95). Springer. https://doi.org/10.1007/978-1-4612-4230-7_5

Hartmann, E., & Müller, C. M. (2014). *Lernfortschrittsdiagnostik Grundrechenarten. 120 Drei-Minuten-Tests für den inklusiven Mathematikunterricht – ZR bis 100: 1.-4. Klasse* [Learning progress monitoring: Basic arithmetic operations.120 three-minute-tests for inclusive mathematics teaching]. Persen.

Hickendorff, M., Torbeyns, J., & Verschaffel, L. (2019). Multi-digit addition, subtraction, multiplication, and division strategies. In A. Fritz, V. G. Haase, & P. Räsänen (Eds.), *International handbook of mathematical learning difficulties* (pp. 543–560). Springer. https://doi.org/10.1007/978-3-319-97148-3_32

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement* (2nd ed.). The Guilford Press.

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development*. Lawrence Erlbaum.

Karp, K., Caldwell, J., Zbiek, R. M., & Bay-Williams, J. (2011). *Developing essential understanding of addition and subtraction for teaching mathematics in Pre-K–Grade 2*. National Council of Teachers of Mathematics.

Kubinger, K. D. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie [Critical evaluation of latent trait theory]. In K. D. Kubinger (Ed.), *Moderne Testtheorie – Ein Abriß samt neuesten Beiträgen* (2nd ed., pp. 19–83). Psychologie Verlags Union.

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item-generating rules to measuring item administration effects. *Psychology Science*, *50*(3), 311–327.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1–20. https://doi.org/10.18637/jss.v020.i09

Mühling, A., Jungjohann, J., & Gebhardt, M. (2019). Progress monitoring in primary education using Levumi: A case study. In H. Lane, S. Zvacek, & J. Uhomoibhi (Eds.), *CSEDU 2019. Proceedings of the 11th International Conference on Computer Supported Education* (pp. 137–144). SCITEPRESS – Science and Technology Publications.

Müller, M. (2020). Item fit statistics for Rasch analysis: Can we trust them? *Journal of Statistical Distributions and Applications*, *7*, Article 5. https://doi.org/10.1186/s40488-020-00108-7

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*.

Poinstingl, H. (2009). The linear logistic test model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychology Science Quarterly*, *51*(2), 123–134.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research, and Evaluation, 20*(11), 1–12. https://doi.org/10.7275/5g6f-ak15

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement. A meta-analysis of National Assessment of Educational

Progress assessments. *Journal of Educational Psychology*, *107*(3), 645–662. https://doi.org/10.1037/edu0000012

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory. *Current Directions in Psychological Science*, *14*(2), 95–101. https://doi.org/10.1111/j.0963-7214.2005.00342.x

Selter, C. (2001). Addition and subtraction of three-digit numbers: German elementary children's success, methods and strategies. *Educational Studies in Mathematics*, *47*(2), 145–173. https://doi.org/10.1023/A:1014521221809

Sikora, S., & Voß, S. (2017). Konzeption und Güte curriculumbasierter Messverfahren zur Erfassung der arithmetischen Leistungsentwicklung in den Klassenstufen 3 und 4 [Conception and quality of curriculum-based measurements for the computation performance of primary school students in grade 3 and 4]. *Empirische Sonderpädagogik*, *9*(3), 236–257. https://doi.org/10.25656/01:15163

Souvignier, E. (2018). Computerbasierte Lernverlaufsdiagnostik [Computer-based learning progress assessment]. *Lernen und Lernstörungen*, *7*(4), 219–223. https://doi.org/10.1024/2235-0977/a000240

Stecker, P. M. (2017). Reflections on teachers' data-based decision making. *Learning Disabilities Research & Practice*, *32*(1), 71–72. https://doi.org/10.1111/ldrp.12128

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. https://doi.org/10.1002/pits.20113

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education,* Article 958530. https://doi.org/10.1155/2013/958530

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Verschaffel, L., Greer, B., & DeCorte, E. (2007). Whole number concepts and operations. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 557–628). Information Age Publishing.

Wilbert, J. (2014). Instrumente zur Lernverlaufsdiagnostik: Gütekriterien und Auswertungsherausforderungen [Tools for learning progress monitoring: Quality criteria and challenges with regard to interpretation]. In M. Hasselhorn, W. Schneider, & U. Trautwein (Eds.), *Tests und Trends: Vol. 12: Lernverlaufsdiagnostik* (pp. 281–308). Hogrefe.

Wilbert, J., & Linnemann, M. (2011). Kriterien zur Analyse eines Tests zur Lernverlaufsdiagnostik [Criteria for analyzing a test measuring learning progress]. *Empirische Sonderpädagogik*, *3*(3), 225–242.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370–371.

Appendix

Table A1: Design Matrix and Item Descriptives

Item number	DGIC 1	DGIC 2	DGIC 3	Item	*n*	*M*	*SD*
1	1	0	0	78 – 6 = 72	581	0.69	0.46
2	0	0	1	30 + 20 = 50	581	0.88	0.32
3	0	1	0	7 + 4 = 11	581	0.91	0.29
4	1	1	0	93 – 7 = 86	581	0.52	0.50
5	0	0	0	56 + 3 = 59	581	0.70	0.46
6	1	0	0	47 – 7 = 40	581	0.79	0.41
7	0	1	0	43 + 9 = 52	571	0.59	0.49
8	1	1	1	32 – 17 = 15	552	0.36	0.48
9	0	1	0	24 + 6 = 30	529	0.77	0.42
10	0	1	1	47 + 26 = 73	511	0.47	0.50
11	1	0	1	76 – 23 = 53	470	0.54	0.50
12	1	0	1	70 – 30 = 40	428	0.83	0.38
13	0	0	1	42 + 24 = 66	403	0.67	0.47
14	0	1	1	37 + 43 = 80	365	0.70	0.46
15	1	1	0	13 – 6 = 7	341	0.76	0.43
16	1	0	1	83 – 23 = 60	313	0.65	0.48
17	0	1	0	8 + 4 = 12	287	0.83	0.38
18	1	0	0	67 – 4 = 63	274	0.76	0.43
19	0	0	1	40 + 10 = 50	250	0.90	0.31
20	1	1	0	23 – 6 = 17	236	0.70	0.46
21	0	0	0	24 + 5 = 29	208	0.78	0.42
22	1	0	1	48 – 26 = 22	188	0.62	0.49
23	1	0	0	97 – 7 = 90	164	0.84	0.37
24	1	1	1	78 – 49 = 29	155	0.47	0.50
25	0	1	0	78 + 2 = 80	133	0.78	0.41
26	0	1	0	73 + 8 = 81	121	0.73	0.44
27	0	1	1	43 + 38 = 81	112	0.58	0.50
28	1	0	1	80 – 70 = 10	95	0.81	0.39
29	0	0	1	32 + 17 = 49	88	0.56	0.50
30	1	0	1	67 – 47 = 20	77	0.68	0.47
31	0	1	1	49 + 31 = 80	60	0.58	0.50
32	1	1	0	16 – 7 = 9	51	0.73	0.45
33	0	1	0	9 + 8 = 17	47	0.75	0.44
34	0	0	1	70 + 20 = 90	41	0.81	0.40
35	1	0	0	38 – 2 = 36	41	0.85	0.36
36	1	1	0	76 – 8 = 68	38	0.74	0.45
37	1	1	1	48 – 29 = 19	34	0.68	0.48
38	1	0	0	28 – 8 = 20	31	0.84	0.37
39	0	0	0	81 + 7 = 88	29	0.66	0.48
40	0	1	0	47 + 9 = 56	25	0.64	0.49
41	1	0	1	50 – 30 = 20	21	0.95	0.22

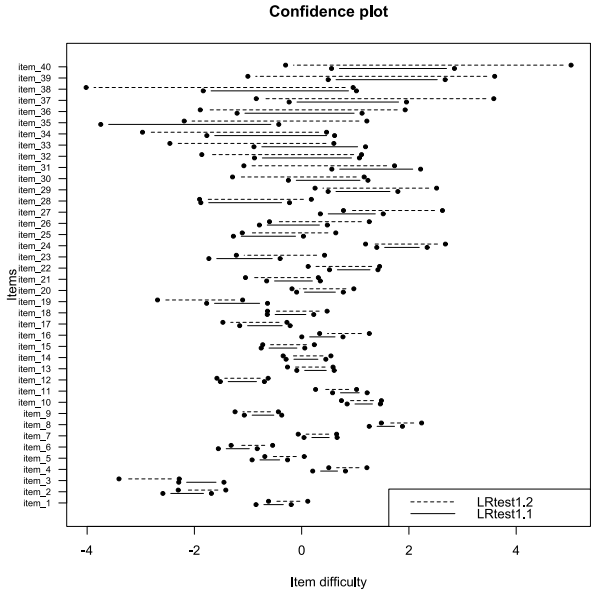
Note. DGIC = difficulty-generating item characteristic.

Table A2: Item Fit Statistics in the RM on Item Level

	Unconditional outfit	Unconditional infit	Conditional outfit		Conditional infit	
Item	MSQ	MSQ	MSQ	*SE*	MSQ	*SE*
1	0.80	0.87	0.74	0.09	0.85	0.05
2	0.95	0.90	0.81	0.19	0.95	0.09
3	1.29	1.08	0.88	0.22	0.98	0.11
4	1.04	1.03	0.89	0.06	0.93	0.04
5	1.12	1.13	0.92	0.09	1.00	0.05
6	0.80	0.89	0.69	0.12	0.85	0.06
7	0.98	1.03	0.91	0.06	0.93	0.04
8	0.83	0.91	0.79	0.07	0.86	0.04
9	0.89	1.02	0.80	0.11	0.91	0.05
10	0.91	0.93	0.89	0.06	0.91	0.04
11	1.06	1.08	0.84	0.06	0.89	0.04
12	1.12	1.04	0.85	0.12	0.90	0.06
13	0.83	0.91	0.76	0.07	0.82	0.04
14	0.61	0.74	0.59	0.07	0.68	0.04
15	0.93	1.00	0.76	0.09	0.88	0.05
16	0.95	0.92	0.71	0.06	0.77	0.04
17	0.94	1.17	0.85	0.13	0.98	0.06
18	0.94	0.85	0.751	0.07	0.81	0.04
19	1.15	1.17	0.71	0.16	0.89	0.07
20	0.78	0.86	0.81	0.07	0.88	0.04
21	0.85	1.05	0.79	0.08	0.89	0.04
22	1.09	1.01	0.92	0.06	0.91	0.04
23	0.87	0.79	0.71	0.11	0.86	0.05
24	0.81	0.93	0.91	0.06	0.95	0.04
25	1.21	1.16	0.85	0.07	0.91	0.04
26	1.13	0.92	0.90	0.09	0.91	0.05
27	0.87	0.91	0.80	0.06	0.87	0.04
28	0.82	0.97	0.97	0.10	1.01	0.91
29	0.74	0.90	1.07	0.06	1.11	0.00
30	0.89	0.84	0.93	0.06	0.96	0.34
31	1.35	1.10	1.20	0.06	1.18	0.04
32	1.84	1.20	0.93	0.07	0.95	0.17
33	1.19	1.18	1.38	0.08	1.32	0.04
34	2.04	1.07	1.47	0.08	1.34	0.04
35	0.66	0.94	0.80	0.07	0.90	0.04
36	0.43	0.71	1.26	0.06	1.15	0.04
37	0.88	1.13	1.39	0.08	1.28	0.04
38	4.71	1.34	2.08	0.06	1.60	0.04
39	0.96	0.90	2.17	0.06	1.66	0.04
40	1.47	1.08	1.56	0.06	1.45	0.04
41	0.05	0.27				

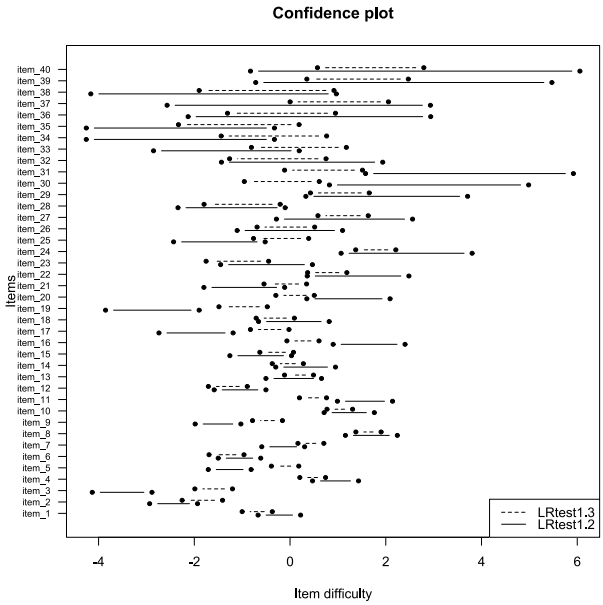
Note. MSQ = mean square residual; the unconditional fit statistics based upon the estimation of the item and person parameters; conditional fit statistics only rely on the item parameters (e.g., Müller, 2020). To estimate the conditional item fit statistics data without missingness is necessary. Therefore, single imputation was done with R package mice (Van Buuren & Groothuis-Oudshoorn, 2011).

Figure A1: Plot of the Confidence Intervals for the Item Parameters with Split Criterion Gender



Note. The solid line indicates the group of girls.

Figure A2: Plot of the Confidence Intervals for the Item Parameters with Split Criterion Grade



Note. The solid line indicates the group of second graders.

Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01.07
Volume 14 (2022), No. 1, 147–173
CC BY-NC-SA 4.0 Waxmann 2022

Holger Gärtner*, Felicitas Thiel*, Christopher Kellermann & Max Nachbauer

Ein theoriebasierter Schülerfragebogen für Unterrichtsevaluation

Zusammenfassung

Der vorliegende Beitrag beschreibt die theoriegeleitete Entwicklung und psychometrische Überprüfung eines Schülerfragebogens zur Erfassung von Unterrichtsqualität, der vorrangig als formatives Evaluationsinstrument in der schulischen Praxis eingesetzt werden soll. Grundlage des Schülerfragebogens stellt ein theoretisch begründetes Modell von Unterrichtsqualität dar, welches die drei Anforderungsbereiche Unterstützung des Wissenserwerbs, Motivierung und Klassenmanagement sowie die querliegende Dimension Individualisierung umfasst. Der Fragebogen wurde an einer Stichprobe von Schüler:innen an Berufsschulen eingesetzt. Am ersten Messzeitpunkt wurde eine vorläufige Version des Schülerfragebogens verwendet. Basierend auf den Ergebnissen wurde der Schülerfragebogen optimiert und an einem zweiten Messzeitpunkt überprüft. Konfirmatorische Faktorenanalysen belegen eine gute Passung zwischen empirischen Daten und theoretischem Modell. Die Skalen sind empirisch klar voneinander trennbar, mit der Ausnahme einer relativ hohen Faktorkorrelation auf Klassenebene. Auch die Reliabilitäten der Skalen erweisen sich mit einer Ausnahme als gut bis sehr gut. Die Ergebnisse sprechen dafür, dass Lehrkräfte mithilfe des Schülerfragebogens konkrete Ansatzpunkte für die Weiterentwicklung ihres Unterrichts ermitteln können.

Schlagworte

Unterrichtsevaluation, Feedback, Unterrichtsqualität, Schülerfragebogen

---

* geteilte Erstautor:innenschaft

Prof. Dr. Holger Gärtner (Korrespondenzautor), ORCID: 0000-0001-8814-9341, Freie Universität Berlin, Arbeitsbereich Unterrichts- und Schulevaluation, Habelschwerdter Allee 45, 14195 Berlin
E-Mail: holger.gaertner@fu-berlin.de

Prof. Dr. Felicitas Thiel · M.A. Christopher Kellermann · Dipl.-Päd. Max Nachbauer, Freie Universität Berlin, Arbeitsbereich Schulpädagogik/Schulentwicklungsforschung, Habelschwerdter Allee 45, 14195 Berlin
E-Mail: felicitas.thiel@fu-berlin.de
christopher.kellermann@fu-berlin.de
max.nachbauer@fu-berlin.de

Development of a Theory-Driven Student Questionnaire for Teacher Evaluation

Abstract

This article describes the theory-driven development and psychometric evaluation of a student questionnaire for the assessment of teaching quality. The student questionnaire is supposed to be used in the context of formative evaluation within schools. It is based on a model of teaching quality that comprises the domains supporting knowledge acquisition, motivating, and classroom management as well as the lateral dimension individualization. The questionnaire was used on a sample of students in vocational schools. At the first measurement time point, a preliminary version of the student questionnaire was applied. Based on the results, the questionnaire was revised and tested at a second measurement time point. Confirmatory factor analyses yielded a good fit between empirical data and theoretical model. All scales are empirically distinguishable from each other with the exception of one relatively high factor correlation on the class level. Further, reliabilities of the scales are good to very good with one exception. The results suggest that teachers can use the student questionnaire for identifying leverage points for improving their teaching.

Keywords

teacher evaluation, feedback, teaching quality, student questionnaire

1. Einleitung

Der im vorliegenden Beitrag vorgestellte Schülerfragebogen zur Messung von fachübergreifenden Aspekten der Unterrichtsqualität wurde im Rahmen des von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts „Unterrichtsbeurteilung und -feedback durch Schulleitungen“ entwickelt. Für eine Interventionsstudie wurde hier zunächst ein Ratingbogen für Schulleitungen, die Unterricht in ganz unterschiedlichen Fächern (auch fachfremd) beurteilen müssen und auf dieser Grundlage ein Feedback geben, entwickelt und validiert (Gärtner, Thiel, Nachbauer & Kellermann, 2021). Der komplementär zu diesem Ratingfragebogen entwickelte Schülerfragebogen wird im vorliegenden Beitrag vorgestellt. Ein auf das Ratinginstrument abgestimmter Schülerfragebogen erlaubt nicht nur eine Evaluation des Unterrichts in kürzeren Abständen zwischen den regelmäßigen, aber häufig nur maximal einmal jährlich stattfindenden Unterrichtsbeurteilungen und -feedbacks durch die Schulleitung, er eröffnet auch eine weitere Perspektive auf den Unterricht (Clausen, 2002; Fauth, Decristan, Rieser, Klieme & Büttner, 2014; Gruehn, 2000). Eine kombinierte Einschätzung der Unterrichtsqualität aus unterschiedlichen Perspektiven ist deshalb für die Unterrichtsentwicklung von hohem Wert.

Im Folgenden wird zunächst die Konzeptualisierung von Unterrichtsqualität in einem Modell, das sowohl dem Ratingbogen als auch dem Schülerfragebogen zugrunde liegt, dargestellt, mit dem Ziel, die theoretischen Dimensionen und Unterrichtsmerkmale zu identifizieren, die wesentliche Ansatzpunkte für die Unterrichtsentwicklung darstellen. Die theoretische Konstruktion geht von drei basalen Anforderungen des Unterrichtens aus: (a) der Unterstützung der schülerseitigen Verarbeitung von Informationen zu bedeutungsvollem Wissen, (b) der Motivierung zur Aneignung eines Lernangebots sowie zur Aufrechterhaltung von Anstrengung auch im Falle des Misserfolgs, (c) der Gestaltung und Steuerung der Interaktion in der Schulklasse mit dem Ziel der Maximierung der aktiven Lernzeit. Für alle drei Dimensionen sowie für das Unterrichtsmerkmal Individualisierung werden auf der Grundlage des nachfolgend dargestellten theoretischen Modells Subskalen ausdifferenziert. Es wird davon ausgegangen, dass die entsprechenden Einzelmerkmale eine konkrete Rückmeldung zu unterschiedlichen Ansatzpunkten der Unterrichtsentwicklung liefern.

2. Unterrichtsevaluation und Konzeptualisierung von Unterrichtsqualität

Im Zuge der Implementation umfassender Qualitätssicherungs- und -entwicklungsverfahren in den letzten 20 Jahren wurden in fast allen deutschen Bundesländern Verfahren der externen und internen Evaluation implementiert (Thiel et al., 2019). Der Evaluation des Unterrichts kommt hier eine besondere Bedeutung zu. Allerdings kommen bislang nur in wenigen Ländern validierte Evaluationsinstrumente zum Einsatz. Das gilt für Ratingbögen von Schulinspektionen in ähnlicher Weise wie für Schülerfragebögen, die im Rahmen der internen Evaluation eingesetzt werden (Helmke et al., 2018; Thiel et al., 2019). Auf der anderen Seite wurden – ausgehend von den TIMSS-Videostudien (Klieme & Rakoczy, 2003) und dem in den USA entwickelten Instrument Classroom Assessment Scoring System (CLASS; Pianta, La Paro & Hamre, 2008) – in der Unterrichtsforschung in den letzten Jahren mehrere Instrumente zur Unterrichtsbeurteilung entwickelt, die drei Basisdimensionen der Unterrichtsqualität erfassen (Praetorius, Klieme, Herbert & Pinger, 2018). Die Forschung zu den drei Basisdimensionen prägt die gegenwärtige Diskussion um die Operationalisierung von Unterrichtsqualität. Obwohl sich die Operationalisierung dieser Basisdimensionen in den unterschiedlichen Instrumenten teilweise deutlich unterscheidet, haben sich als gemeinsame Bezeichnungen der Dimensionen inzwischen die drei Begriffe effiziente Klassenführung, kognitive Aktivierung und schülerorientiertes Unterrichtsklima (Klieme & Rakoczy, 2003) bzw. classroom management, cognitive activation und student support durchgesetzt (Praetorius et al., 2018). Die Identifikation der Dimensionen erfolgte zunächst im Rahmen empirischer Analysen, später wurden die drei Dimensionen mit Bezug auf konstruktivistische Lerntheorien, Motivationstheorien, Unterrichtsklimaforschung

und Ansätze zum Klassenmanagement theoretisch unterlegt. Durch die Zusammenführung mehrerer Unterrichtsmerkmale in wenige, inhaltlich relevante Dimensionen genügen die Instrumente forschungsrelevanten Anforderungen wie Sparsamkeit, Vermeidung von Multikollinearität oder hoher Erklärkraft für den Lernerfolg von Schüler:innen in hohem Maß (Praetorius et al., 2018). Ursprünglich für das Fach Mathematik entwickelt, wurde infolge zahlreicher Studien in anderen Fächern der Anspruch auf einen generischen Charakter der drei Basisdimensionen erhoben. Von Vertreter:innen der Fachdidaktik wird allerdings immer wieder der Bedarf an einer fachdidaktischen Spezifikation, insbesondere der Items zur kognitiven Aktivierung, artikuliert. So wird gegenwärtig z. B. eine fachdidaktische Weiterentwicklung des Modells der Basisdimensionen in der Mathematikdidaktik diskutiert (Prediger & Neugebauer, 2021), eine ähnliche Diskussion findet in den Didaktiken der Naturwissenschaft statt (Steffensky & Neuhaus, 2018).

Unabhängig von einer fachlichen Spezifizierung stellt die Beurteilung der Dimension kognitive Aktivierung hohe Ansprüche an die Beurteilenden, werden hier doch auf der Grundlage einer konstruktivistischen Lerntheorie hoch-inferente Indikatoren zur Unterrichtsqualität operationalisiert. Außerdem scheint es erforderlich zu sein, Unterricht über einen längeren Zeitraum zu beobachten, um zu einer zuverlässigen Einschätzung der Qualität kognitiver Aktivierung zu kommen. Praetorius et al. (2014) konnten in einer Generalisierbarkeitsstudie zeigen, dass für eine zuverlässige Einschätzung dieser Qualitätsdimension wenigstens neun Unterrichtsstunden erforderlich sind. Die Items der Dimension kognitive Aktivierung erfassen höherwertige Verstehensprozesse, die für den Erwerb von elaborierten Kompetenzen im Fachunterricht unverzichtbar sind (Praetorius et al., 2018). Dagegen werden empirisch gut abgesicherte Merkmale guten Unterrichts wie Klarheit und Strukturiertheit (Helmke et al., 2018), die für die Unterstützung kognitiver Prozesse ebenfalls relevant sind, sowohl in den Modellannahmen als auch in den gegenwärtigen Instrumenten oft nicht operationalisiert.

Vor dem Hintergrund der fachdidaktischen Diskussion auf der einen Seite und der Schwierigkeit der Operationalisierung der Dimension kognitive Aktivierung auf der anderen Seite ist die Frage zu stellen, ob für Evaluationsinstrumente, die eine Beurteilung von Unterrichtsqualität sowie ein Unterrichtsfeedback aus fachübergreifender Perspektive (Ratingbogen für Schulleitungen) oder Schülerperspektive (Schülerfragebogen) unterstützen und deren primärer Zweck die Unterrichtsentwicklung ist, nicht ein Ansatz fruchtbar sein könnte, der sich stärker an allgemeinen kognitionspsychologischen und motivationspsychologischen Ansätzen orientiert. Ein Vorteil dieses Ansatzes ist, dass mit dem Instrument zugleich ein leicht verständliches theoretisches Modell vorliegt, das keine fachdidaktischen Kenntnisse voraussetzt. Ein solches Modell ist für die Gestaltung der Qualifizierung von Beurteilenden aus der Praxis, hier Schulleitungen, wie für die Führung eines bedeutungsvollen Feedbackgesprächs gleichermaßen hilfreich (Heneman & Milanowski, 2003). An diesen Überlegungen ansetzend, haben wir eine an allgemeinen kognitionspsychologischen, motivationspsychologischen sowie mikrosoziologischen Ansätzen orientierte Operationalisierung von drei Basisdimensionen vorgenommen

(Ophardt & Thiel, 2013). Die hier vorgeschlagene Operationalisierung der drei Basisdimensionen unterscheidet sich in vier Punkten von den vorliegenden Fragebögen, die im Rahmen unterschiedlicher Forschungsprojekte entwickelt worden sind:

1. Anstatt kognitiver Aktivierung wird der gesamte Prozess der Unterstützung des Wissenserwerbs auf der Grundlage eines Lehr-Lern-Prozessmodells (Klauer & Leutner, 2012) operationalisiert; das heißt, es werden auch basale Merkmale der Unterrichtsqualität wie Klarheit und Strukturiertheit oder Üben einbezogen.
2. An die Stelle der Dimension Schülerorientierung/konstruktive Unterstützung, die sowohl motivationspsychologische als auch kognitionspsychologische Konstrukte beinhaltet, tritt die Dimension Motivierung. Auf der Grundlage des Erwartung-mal-Wert-Modells (Eccles & Wigfield, 2002; Rheinberg, 2008) werden unterschiedliche Ansatzpunkte für die Motivierung der Schüler:innen identifiziert.
3. Die Dimension Klassenmanagement umfasst alle Aspekte, die sich auf die Gestaltung einer lerndienlichen Interaktion in der Schulklasse beziehen; also auch das Klassen-/Unterrichtsklima, das anders als Motivierungsstrategien auf die grundlegende Beziehung zwischen Lehrkraft und Schüler:innen abhebt, die wiederum die Grundlage für die Gestaltung und Steuerung der sozialen Interaktionen in der Schulklasse darstellt und damit theoretisch dem Klassenmanagement zuzuordnen ist.
4. Es werden Merkmale der Individualisierung des Unterrichts für die kognitiven und motivationalen Aspekte des Lernens operationalisiert, weil diese Merkmale für die Unterrichtsentwicklung in inklusiven Schulen eine große Rolle spielen.

Für alle drei Dimensionen sowie für das Unterrichtsmerkmal Individualisierung werden auf der Grundlage des nachfolgend dargestellten theoretischen Modells Subskalen ausdifferenziert. Die entsprechenden Einzelmerkmale erlauben eine konkrete Rückmeldung zu unterschiedlichen Ansatzpunkten der Unterrichtsentwicklung. Die theoretische Konstruktion geht von drei basalen Anforderungen des Unterrichtens aus: (a) der Unterstützung der schülerseitigen Verarbeitung von Informationen zu bedeutungsvollem Wissen, (b) der Motivierung zur Aneignung eines Lernangebots sowie zur Aufrechterhaltung von Anstrengung auch im Falle des Misserfolgs, (c) der Gestaltung und Steuerung der Interaktion in der Schulklasse mit dem Ziel der Maximierung der aktiven Lernzeit.

2.1 Unterstützung des Wissenserwerbs

Die erste Dimension, die die kognitiven Funktionen des Lernens betrifft, wird als Unterstützung des Wissenserwerbs bezeichnet. Ihre Konzeption beruht auf dem Lehr-Lern-Prozessmodell von Klauer und Leutner (2012), das wiederum unterschiedliche Ansätze der Informationsverarbeitung und des Gedächtnisses integriert. Das Modell unterscheidet vier kognitive Komponenten des Lernens: (a) Informierung, (b) Informationsverarbeitung, (c) Speicherung und Abruf von In-

formationen und (d) Transfer. Informierung wird in unserem Modell als Subskala *Erklärung und Präsentation* operationalisiert. Die Herausforderung bei der Einführung neuer Konzepte oder der Darstellung neuen Wissens besteht darin, dass im sensorischen Register viele Reize aufgenommen werden, aber die Kapazität des Arbeitsgedächtnisses für die Weiterverarbeitung der Reize sehr gering ist. Informationen sollten daher so aufbereitet werden, dass die extrinsische kognitive Belastung des Arbeitsgedächtnisses, im Unterschied zur lernbezogenen Belastung (germane load), möglichst gering ist. Das bedeutet: Die Gestaltung der Lernmaterialien muss so erfolgen, dass keine unnötige Beanspruchung des Arbeitsgedächtnisses der Lernenden entsteht (Klauer & Leutner, 2012; Sweller, 1988). Eine gute Informationsdarbietung ist entsprechend durch eine strukturierte Präsentation sowie verständliche Erklärungen und Veranschaulichungen durch Beispiele gekennzeichnet. Im Anschluss an die Informationsaufnahme erfolgt die Informationsverarbeitung. Lehrkräfte sollten in dieser Phase die Verstehensprozesse der Schüler:innen unterstützen. Diesbezüglich unterscheiden Klauer und Leutner (2012) zwei Aspekte: die Elaboration von Informationen und ihre anschließende Reduktion. Elaborieren umfasst Prozesse wie Klärung von Zusammenhängen, die Kontextualisierung von Informationen oder die Entfaltung unterschiedlicher Detailaspekte. Angeregt werden können Elaborationsprozesse durch Vergleiche und kognitiv aktivierende Fragen, die auf Voraussetzungen, Bedingungen und Folgen abzielen. Reduktive Prozesse haben die Funktion, Informationen zu komprimieren. Lehrkräfte können diese anregen, indem sie behandelte Inhalte noch einmal zusammenfassen. Neben elaborativen und reduktiven Strategien ist der Einsatz von metakognitiven Strategien hilfreich für die Schüler:innen. Metakognitive Strategien umfassen die Planung, Überwachung und Anpassung des Lernprozesses (Friedrich & Mandl, 1992). Lehrkräfte können den Einsatz metakognitiver Strategien beispielsweise dadurch unterstützen, dass sie die Schüler:innen dazu auffordern, ihre Vorgehensweisen zu planen und ihre Lösungen selbst zu bewerten. Die langfristige Verfügbarkeit von Wissen ist von einer nachhaltigen Speicherung im Langzeitgedächtnis durch ausreichende Übung abhängig. Neben einer ausreichenden Häufigkeit von Übungsphasen ist es relevant, wie die Übungen konkret ausgestaltet werden. Wichtig ist, dass besonders schwierige Aspekte intensiv geübt werden und dass die Resultate des Übens überprüft und gegebenenfalls korrigiert werden (Anderson, 1982). Die vierte kognitive Funktion des Lernens ist der Transfer. Hier geht es darum, erworbenes Wissen so zu flexibilisieren, dass es auf neue Probleme übertragen werden kann bzw. darum, gelernte Prinzipien oder Strategien in anderen Zusammenhängen anzuwenden. Die Schüler:innen müssen dazu die Tiefenstruktur eines Sachverhalts erfassen (Klauer & Leutner, 2012). Der Transfer kann angeregt werden, indem die Schüler:innen ihr theoretisches Wissen auf praktische Probleme anwenden oder indem die Schüler:innen zu einem Thema mehrere Aufgaben bearbeiten, die den Kontext variieren.

2.2 Motivierung

Neben den kognitiven Funktionen sind die motivationalen Funktionen für das Lernen zentral. Motivation sorgt dafür, dass Lernaktivitäten begonnen und aufrechterhalten werden. Grundlage für die Konzeptualisierung der Unterrichtsmerkmale zur Erfassung der Motivierungsqualität ist das Erwartung-mal-Wert-Modell (Eccles & Wigfield, 2002; Rheinberg, 2008). Vier Aspekte der Motivierung, die für den Unterricht eine zentrale Bedeutung haben, lassen sich hieraus ableiten: Durch (a) das Wecken von Interesse, (b) die Förderung von Lernfreude und (c) die Verdeutlichung des Nutzens können Anreize für Lernaktivitäten geschaffen werden, während (d) die Stärkung der Selbstwirksamkeitserwartung eine günstige Einschätzung der eigenen Erfolgswahrscheinlichkeit beim Lernen fördert (Eccles & Wigfield, 2002).

Das Wecken von Interesse ist ein wichtiger Ansatzpunkt für die Motivierung der Schüler:innen. Interesse bezieht sich auf die jeweiligen Lerninhalte und stellt damit einen gegenstandsbezogenen Anreiz dar. Es setzt sich aus wertbezogenen Überzeugungen und Gefühlsüberzeugungen zusammen (Schiefele, 2009). Lehrkräfte können Interesse wecken, indem sie die Schüler:innen mit neuen Einsichten überraschen oder indem sie im Unterricht mit Enthusiasmus auftreten und so selbst Interesse an den Lerninhalten demonstrieren. Im Gegensatz zum Interesse als gegenstandsbezogenem Anreiz bezieht sich Lernfreude auf die Art der Lernaktivitäten und wird daher auch als tätigkeitsbezogener Anreiz bezeichnet (Rheinberg, 2008). Die Lehrkräfte sollten demnach darauf achten, dass die Schüler:innen im Unterricht Tätigkeiten nachgehen können, bei denen sie Freude und Spaß erleben. Hierzu zählen beispielsweise explorierende, gestalterische oder spielerische Aktivitäten, die die Lernenden entsprechend ihrer Fähigkeiten herausfordern und ein Aufgehen im Tun ermöglichen (Csikszentmihalyi & Rathunde, 1993). Ein weiterer Anreiz bezieht sich darauf, welchen Nutzen die Lerninhalte für die Schüler:innen haben (Hulleman & Harackiewicz, 2009). Für die entsprechende Motivierungsstrategie ist es entscheidend, dass Lehrkräfte verdeutlichen, warum bestimmte Lerninhalte für die Schüler:innen relevant sind und dass sie Beispiele geben, in welchen Lebenssituationen bestimmte Kenntnisse oder Kompetenzen benötigt werden. Anreize alleine motivieren allerdings nicht zum Lernen. Die Schüler:innen müssen gleichzeitig auch die Erwartung haben, das Lernziel mit ihren eigenen Kompetenzen erreichen zu können. Aus diesem Grund stellt die Förderung von Selbstwirksamkeitserwartungen (Bandura, 1976) einen wichtigen Ansatzpunkt für Motivierung dar. Relevant ist hierbei, dass die Lehrkraft den Schüler:innen zeigt, dass sie ihnen etwas zutraut und dass sie die Schüler:innen spezifisch und aufgabenbezogen lobt (Brophy, 1981). Weiterhin sollte die Lehrkraft Herausforderungen formulieren, deren Bewältigung möglich ist. Bei Schwierigkeiten sollte die Lehrkraft die Schüler:innen ermutigen, aber die Lösung nicht vorwegnehmen, sodass die Schüler:innen Erfolgserfahrungen machen können (Bandura, 1993).

2.3 Klassenmanagement

Schulklassen sind dynamische Interaktionssysteme, in denen Lernen nur dann gelingt, wenn ein gewisses Maß an sozialer Ordnung sichergestellt ist (Doyle, 2006; Ophardt & Thiel, 2013; Piwowar, 2013). Diese sozialen Faktoren des Lernens werden im Konstrukt des Klassenmanagements konzeptualisiert. Im Vordergrund steht die Maximierung der aktiven Lernzeit für alle Schüler:innen der Klasse. Fünf zentrale Funktionen können im Hinblick auf das Klassenmanagement der Lehrkraft unterschieden werden (Thiel, Richter & Ophardt, 2012): (a) Unterrichtsklima, (b) Monitoring, (c) Steuerung des Unterrichtsflusses, (d) Gruppenaktivierung und (e) der Umgang mit Störungen.

Für ein lernförderliches Unterrichtsklima muss zunächst eine Interaktionsordnung durch die Einführung von Regeln etabliert werden. Die Beachtung von Regeln durch die Schüler:innen in der Form von gegenseitigem Respekt, Rücksichtnahme und die Bereitschaft zur Unterstützung findet ebenso Ausdruck im Unterrichtsklima wie entsprechende Verhaltensweisen der Lehrkraft, in denen das Arbeitsbündnis mit der Klasse zum Ausdruck kommt. Ein effektives Klassenmanagement setzt voraus, dass die Lehrkraft das Geschehen in der ganzen Klasse wahrnimmt (Kounin, 2006). Dies betrifft das Monitoring. Damit ist gemeint, dass die Lehrkraft über die gesamte Unterrichtsstunde hinweg die Aktivitäten aller Schüler:innen im Blick behält und bemerkt, wenn sich einzelne Schüler:innen mit unterrichtsfremden Dingen beschäftigen. Für die Maximierung aktiver Lernzeit ist weiterhin eine effektive Steuerung des Unterrichtsflusses wichtig, einerseits, weil bei unnötigen Pausen Lernzeit verloren geht, andererseits, weil das Handlungsprogramm durch Leerläufe oder Nachfragen unterbrochen wird und dann mühevoll wiederaufgebaut werden muss (Thiel et al., 2012). Für die Aufrechterhaltung des Unterrichtsflusses sind nahtlose Übergänge zwischen Unterrichtsphasen und die Vermeidung von inhaltlichen Abschweifungen wichtig. Auf die Bedeutung der Gruppenaktivierung hat bereits Kounin (2006) mit seiner Videostudie aufmerksam gemacht. Gruppenaktivierung meint, dass die Lehrkraft alle Schüler:innen aktiv in den Unterricht einbindet und sicherstellt, dass kein Leerlauf für einzelne Schüler:innen entsteht. Störungen des Unterrichts entstehen, weil die Schüler:innen die Aufmerksamkeit nicht über längere Zeit aufrechterhalten können, weil sie von ihren Mitschüler:innen abgelenkt werden, diese beindrucken wollen oder weil sie sich ungerecht behandelt fühlen (Thiel, 2016). Wenn Störungen auftreten, sollte die Lehrkraft sich klar verhalten und das Störverhalten möglichst unmittelbar beenden. Ein angemessener Umgang mit Störungen berücksichtigt die Schwere des Störverhaltens. Bei kleineren Störungen sind beiläufige nonverbale Signale häufig ausreichend. Bei schwereren Störungen können dagegen verbale Ermahnungen und gegebenenfalls auch die Ankündigung von Sanktionen notwendig sein (Thiel, 2016). Für jede Störungsintervention gilt, dass sie das Arbeitsbündnis zwischen Lehrkraft und Schüler:innen nicht beschädigen sollte. Zurechtweisungen müssen aus diesem Grund konstruktiv erfolgen.

2.4 Individualisierung

Da sich Schüler:innen hinsichtlich ihrer Lernvoraussetzungen unterscheiden, stehen Lehrkräfte schließlich vor der Herausforderung der Individualisierung des Unterrichts (Klieme & Warwas, 2011). Individualisierung wird im vorliegenden Modell als querliegende Dimension verstanden, das heißt, Individualisierung stellt keinen eigenständigen Anforderungsbereich dar, sondern bezieht sich inhaltlich auf die zuvor dargestellten Bereiche. Individualisierende Maßnahmen sind primär in den Bereichen Unterstützung des Wissenserwerbs und Motivierung erforderlich, da die Schüler:innen sowohl unterschiedliche kognitive (z. B. Vorwissen) als auch unterschiedliche motivationale Lernvoraussetzungen (z. B. Interessen) aufweisen. Im Bereich des Klassenmanagements ist Individualisierung weniger relevant, da sich Klassenmanagement per definitionem auf die gesamte Klasse bezieht. Dementsprechend werden die zwei Aspekte individualisierter Wissenserwerb und individualisierte Motivierung unterschieden.

Auf der Grundlage der dargestellten Subdimensionen des theoretischen Modells wurden Skalen sowohl für den Unterrichtsbeurteilungsbogen für Schulleitungen (Gärtner et al., 2021) als auch für den in diesem Beitrag vorgestellten Schülerfragebogen operationalisiert.

3. Entwicklung eines fachübergreifenden Schülerfragebogens

Mehrere Studien haben gezeigt, dass Schüler:innen in der Lage sind, eine bedeutungsvolle, valide und reliable Einschätzung zur Qualität von Unterricht zu geben (Gärtner, 2010; Gärtner & Brunner, 2018; Göllner et al., 2016; Gruehn, 2000; Kunter et al., 2005; Wagner, Göllner, Helmke, Trautwein & Lüdtke, 2013). Schülerbefragungen zum Unterricht haben im Vergleich zu Unterrichtsratings durch trainierte Beobachter:innen eine Reihe von Vorteilen (vgl. auch Clausen & Göbel, 2020): Erstens legen Schüler:innen ihrer Urteilsbildung eine größere Zahl von Ereignissen bzw. einen längeren Beobachtungszeitraum zugrunde als externe Beobachter:innen, zweitens haben sie viele Möglichkeiten, unterschiedliche Lehrkräfte zu vergleichen und damit eine soziale Bezugsnorm zu entwickeln, drittens ist die subjektive Perspektive der Lernenden auf die Unterstützung des Lernprozesses oder die Beziehung zur Lehrkraft von besonderem Interesse und viertens kann ein Vergleich der Perspektive der Schüler:innen mit der Perspektive der Lehrkraft zur Überprüfung der eigenen Wahrnehmung führen (Helmke et al., 2018).

Die Entwicklung (und der Einsatz) von Schülerfragebögen ist allerdings mit besonderen Herausforderungen verbunden. Dass die Schüler:innen- und die Lehrkraftperspektive in vielen Studien vergleichsweise weit auseinanderliegen (Clausen, 2002; Fauth, et al., 2014), kann durchaus auch als Validitätsproblem interpretiert werden. In jüngster Zeit ist die bereits von Clausen (2002) angeregte Diskussion

um den Einfluss eines „affektiven Gesamteindrucks“ auf die Wahrnehmung einzelner Merkmale wieder aufgenommen und in der wahrnehmungspsychologischen Perspektive des *Halo-Effekts* untersucht worden (Röhl & Rollett, 2020; Wagner, 2008). Röhl und Rollett (2020) konnten einen *Communion-Bias*, also die dominierende Bedeutung der subjektiven Einschätzung der Beziehung zur Lehrkraft für die Beurteilung der Unterrichtsqualität durch Schüler:innen nachweisen. Ein anderer Untersuchungsstrang beschäftigt sich damit, ob Schüler:innen die Bedeutung einzelner Items überhaupt verstehen. So untersucht Lenske (2016) das Verständnis der Items bei Grundschüler:innen und kommt zu dem Befund, dass die Bedeutung einiger Items von den Schüler:innen nicht korrekt erschlossen wird. Dies gilt insbesondere für die hoch-inferenten Items zur kognitiven Aktivierung. In einer aktuellen Studie von Lenske und Praetorius (2020) wurde der Befund der unzureichenden oder falschen semantischen Decodierung der Items zur kognitiven Aktivierung an einer Stichprobe von Sekundarschüler:innen bestätigt. Die Befunde stehen im Einklang mit der – im Vergleich mit der Dimension des Klassenmanagements – schwachen Korrelation von Schülerurteilen zur kognitiven Aktivierung und zu den Leistungen (Hanisch, 2018). Neben Fragen der Urteilsverzerrung durch einen Sympathie-Bias und den Untersuchungen zum Verständnis von Items ist die Frage der Trennschärfe zwischen den einzelnen Dimensionen relevant. Hier zeigt eine Untersuchung von Wagner (2008), dass Schüler:innen die Differenzierung zwischen unterschiedlichen theoretischen Dimensionen häufig schwerfällt.

Die Vorteile einer Nutzung von Schüler:inneneinschätzungen der Unterrichtsqualität für Unterrichtsentwicklung sind, so lassen sich die vorliegenden Befunde interpretieren, an bestimmte Voraussetzungen der Fragebogenkonstruktion und Itemformulierung gebunden:

- Items sollten durchgängig auf einen Bezug auf Schüler:innenverhalten verzichten, um eine mögliche selbstdienliche Urteilsverzerrung zu vermeiden.
- Items sollten möglichst konkrete Verhaltensweisen der Lehrkraft beschreiben, deren Einschätzung für die Schüler:innen nicht an das Verständnis von didaktischen Kernkonzepten (z. B. Fehlerfreundlichkeit) geknüpft ist.
- Items sollten grundsätzlich einfach formuliert sein.
- Die theoretische Trennschärfe der unterschiedlichen Dimensionen sollte auch für Schüler:innen prinzipiell erkennbar sein.

4. Ziele der Studie

Auf der Grundlage des dargestellten Modells von Unterrichtsqualität wurde, als Ergänzung zum bereits etablierten Beobachtungsbogen, ein Schülerfragebogen entwickelt. Das Ziel der vorliegenden Studie ist die psychometrische Überprüfung dieses Fragebogens anhand unterschiedlicher Gütekriterien. Ausgangspunkt der Prüfung ist die Fragestellung, ob Lehrkräfte mittels des vorliegenden Instrumentes reliable und valide Rückmeldungen über verschiedene Aspekte des Unterrichts erhalten

können. Zur Prüfung gehören dementsprechend verschiedene Maße der Reliabilität der entwickelten Skalen sowie Maße der Konstruktvalidität. Der letztgenannte Aspekt bezieht sich auf die Frage, ob die einzelnen Skalen eindimensionale Konstrukte abbilden und ob sich die Konstrukte empirisch voneinander trennen lassen.

5. Methode

5.1 Entwicklung des Schülerfragebogens

Ausgehend von dem beschriebenen Modell der Unterrichtsqualität wurden Items für die zuvor beschriebenen Aspekte von Unterrichtsqualität entwickelt. Hierbei wurde, wo dies sinnvoll erschien, auf bereits vorliegende Skalen zurückgegriffen (Baumert et al., 2008; Baumert, Gruehn, Heyn, Köller & Schnabel, 1997; Gärtner, 2010; Helmke, 2017; Jäger & Helmke, 2008; Nationales Bildungspanel, 2013; Rakoczy, Buff & Lipowsky, 2005; Seidel, Prenzel, Duit & Lehrke, 2003; Thiel, Ophardt & Piwowar, 2013). Die Itemformulierungen des Schülerfragebogens orientieren sich an den in Kapitel 2 auf der Grundlage des Forschungsstands herausgearbeiteten Kriterien:

- Um soziale Erwünschtheit zu minimieren, wurden alle Items auf das Lehrkraftverhalten bezogen. Mit der Ausnahme des Unterrichtsklimas zwischen den Schüler:innen enthält der Fragebogen keine Items, die eine Selbstbewertung der Schüler:innen nahelegen (z. B. Clausen, 2002).
- Auch zur Erfassung der Unterstützung kognitiver Aktivitäten (z. B. Vergleiche) werden konkrete Aktivitäten der Lehrkraft abgefragt (z. B. „Unsere Lehrkraft lässt unterschiedliche Lösungswege von Aufgaben vergleichen und bewerten."). Es wird vermieden, dass Schüler:innen ihre eigenen kognitiven Prozesse einschätzen müssen. Dadurch wird das Verständnis der Items erleichtert (z. B. Lenske & Praetorius, 2020).
- Alle Items sind sprachlich einfach formuliert. Auf komplexe Formulierungen wurde verzichtet (z. B. Lenske, 2016).
- Die drei Dimensionen sind trennscharf entsprechend der theoretisch hergeleiteten Unterrichtsmerkmale (siehe Kapitel 1) operationalisiert, um für die Schüler:innen die Abgrenzung von Fragen zur Unterstützung des Verstehensprozesses, zur Motivierung und zum Klassenmanagement nachvollziehbar zu machen (z. B. Wagner, 2008).

Als Format wurden 4-stufige Likert-Skalen gewählt. Beispielitems finden sich in Tabelle 1.

Tabelle 1: Unterrichtsmerkmale nach Anforderungsbereichen, Beispielitems und Reliabilitäten zu Messzeitpunkt 1 und 2

		Messzeitpunkt 1				Messzeitpunkt 2			
Unterrichtsmerkmal	Beispielitem	Items	Ω	ICC1	ICC2	Items	Ω	ICC1	ICC2
Unterstützung des Wissenserwerbs									
Erklärung und Präsentation	Unsere Lehrkraft gibt anschauliche Beispiele, die uns das Verstehen der Lerninhalte erleichtern.	3	.75	.29	.87	4	.84	.32	.85
Unterstützung des Verstehens	Unsere Lehrkraft lässt uns unsere Aufgabenlösungen selbst korrigieren.	4	.70	.11	.67	5	.76	.12	.63
Speicherung durch Üben	Wir erhalten gezielt Übungsaufgaben für besonders schwierige Themen.	5	.73	.20	.80	5	.80	.24	.79
Anwendung und Transfer	Unsere Lehrkraft stellt uns häufig Aufgaben, bei denen wir unser Wissen auf neue Situationen übertragen müssen.	5	.75	.21	.81	5	.80	.17	.72
Motivierung									
Weckung von Interesse	Auch bei bekannten Themen überrascht uns unsere Lehrkraft oft mit neuem Wissen.	6	.90	.28	.86	6	.91	.31	.85
Förderung von Freude	Die Bearbeitung der Aufgaben, die unsere Lehrkraft stellt, macht mir Spaß.	4	.89	.29	.87	4	.88	.19	.74
Verdeutlichung des Nutzens	Unsere Lehrkraft erklärt uns, warum es für das spätere Leben wichtig ist, das Unterrichtsthema zu lernen.	4	.82	.28	.86	4	.87	.24	.79
Förderung von Selbstwirksamkeit	Unsere Lehrkraft muntert uns auf, damit wir auch bei schwierigen Aufgaben nicht den Mut verlieren.	4	.88	.24	.84	4	.88	.28	.83
Klassenmanagement									
Lerndienliches Unterrichtsklima SuS	Im Unterricht gehen die Schülerinnen und Schüler respektvoll miteinander um.	2	.87	.32	.88	4	.91	.23	.79
Lerndienliches Unterrichtsklima LK	Unsere Lehrkraft ist immer freundlich zu den Schülerinnen und Schülern.					4	.88	.26	.81
Monitoring	Unsere Lehrkraft bekommt mit, was die Schülerinnen und Schüler im Unterricht machen.	4	.87	.16	.76	4	.90	.22	.77
Steuerung des Unterrichtsflusses	Unsere Lehrkraft schweift häufig vom Thema ab.[1]	4	.78	.22	.82	4	.77	.07	.49
Gruppenaktivierung	Im Unterricht unserer Lehrkraft beteiligen sich meist alle Schülerinnen und Schüler.	2	.75	.21	.81	4	.88	.16	.70
Umgang mit Störungen	Unsere Lehrkraft reagiert angemessen auf Störungen (also nicht zu streng und nicht zu mild).	4	.83	.21	.81	4	.84	.25	.80
Individualisierung	Unsere Lehrkraft gibt uns je nach Leistungsstärke unterschiedliche Aufgaben.	3	.75	.20	.80	6	.89	.11	.60

Anmerkungen. ICC = Intraklassenkorrelation; SuS = Schüler:innen; LK = Lehrkraft. [1] = Das Item wurde invertiert, sodass hohe Werte für eine gute Steuerung des Unterrichtsflusses stehen.

5.2 Stichprobe

Eingesetzt wurde der Schülerfragebogen im Rahmen des DFG-Projektes „Unterrichtsbeurteilungen und -feedback durch Schulleitungen“. Hierbei handelt es sich um ein Forschungsprojekt zur Überprüfung der Wirksamkeit von Unterrichtsfeedback. Entsprechend erfolgte die Rekrutierung der Schülerstichprobe über die am Projekt beteiligten Schulen und Lehrkräfte. Die vorliegende Studie verwendet Daten von zwei Messzeitpunkten. Messzeitpunkt 1 diente der Pilotierung der ersten Version des Fragebogens sowie der weiteren Optimierung. Entsprechend wurden für Messzeitpunkt 1 mehr Items entwickelt als für die Endversion intendiert. Auf der Grundlage der psychometrischen Analysen wurde eine überarbeitete Version des Fragebogens erstellt und für Messzeitpunkt 2 verwendet.

Die Erhebungen fanden im Oktober 2017 (Messzeitpunkt 1) und im Juni 2019 (Messzeitpunkt 2) an beruflichen Schulen in Berlin statt. Zu Messzeitpunkt 1 beteiligten sich 6 berufliche Schulen mit insgesamt 81 Klassen und 1299 Schüler:innen an der Ausgangsbefragung zur Wahrnehmung des Unterrichts (durchschnittlich 16 Schüler:innen pro Klasse). Die Befragung erfolgte in unterschiedlichen Jahrgangsstufen (13.8 % Jahrgangsstufe 10, 32.3 % Jahrgangsstufe 11, 13.2 % Jahrgangsstufe 12, 40.6 % Missing) und unterschiedlichen Fächern (33.3 % MINT[1]-Fächer, 13.7 % sprachliche Fächer, 8.2 % sozialwissenschaftliche Fächer, 4.1 % sonstige Fächer, 40.6 % Missing). Zu Messzeitpunkt 2 beteiligten sich an den 6 beruflichen Schulen noch 24 Klassen und 294 Schüler:innen an der Befragung. Wiederum wurden durch die Befragung unterschiedliche Jahrgangsstufen (6 % Jahrgangsstufe 11, 54.8 % Jahrgangsstufe 12, 14.4 % Jahrgangsstufe 13, 24.7 % Missing) und unterschiedliche Fächer (56.9 % MINT-Fächer, 8.7 % sprachliche Fächer, 7.7 % sozialwissenschaftliche Fächer, 2 % sonstige Fächer, 24.7 % Missing) abgedeckt. Alle Befragungen wurden vor Ort von geschulten Testleiter:innen mit Papierfragebogen durchgeführt.

5.3 Analysen

Die aus dem theoretischen Modell abgeleiteten Skalen wurden zunächst einer Itemanalyse unterzogen, um besonders leichte bzw. schwierige sowie wenig trennscharfe Items zu identifizieren. Innerhalb der Itemanalyse wurden explorative Faktorenanalysen (EFA) durchgeführt, da es sich um ein neues Instrument handelt, über dessen empirische Faktorenstruktur bisher noch nichts bekannt ist. Zur Optimierung der einzelnen Skalen wurden Items mit geringen Faktorladungen bzw. Parallelladungen ausgeschlossen. Die explorativ ermittelten Skalen wurden anschließend mithilfe von konfirmatorischen Faktorenanalysen (CFA) im Rahmen des Gesamtmodells überprüft. Da ein Gesamtmodell mit allen Unterrichtsmerkmalen auf zwei Ebenen aufgrund zu vieler zu schätzender Parameter zu komplex für die

1 MINT = Mathematik, Informatik, Naturwissenschaften und Technik.

vorliegende Stichprobe war, wurde zunächst eine CFA auf Individualebene mit allen Unterrichtsmerkmalen und anschließend getrennte Multilevel-CFAs für die drei Bereiche Unterstützung des Wissenserwerbs, Motivierung und Klassenmanagement durchgeführt. Das Unterrichtsmerkmal Individualisierung bleibt als querliegende Anforderung alleinstehend. Die Güte der Modellpassung wird anhand der folgenden Fit-Indizes geprüft: CFI, RMSEA, χ^2/df (Hu & Bentler, 1999).

Weiterhin wurden Reliabilitätsmaße berechnet, einmal als interne Konsistenz in der Form des Koeffizienten Ω (McDonald, 1999) sowie für die jeweils aggregierten Skalenmittelwerte in Form der Intraklassenkorrelation (ICC2; Gärtner, 2010; Lüdtke, Trautwein, Kunter & Baumert, 2006). Diese Reliabilität gilt als Voraussetzung für die Berechnung von Klassenmittelwerten für die Unterrichtsevaluation. Ergänzt werden diese Kennwerte um die Angabe der ICC1, welche ausweist, ob sich die gemessenen Aspekte der Unterrichtsqualität zwischen den einzelnen Klassen unterscheiden (Lüdtke et al., 2006). Die Analysen zur faktoriellen Validität und internen Konsistenz erfolgten mit Mplus 8 (Muthén & Muthén, 1998–2017). Fehlende Werte wurden durch die Full-Information-Maximum-Likelihood-Schätzmethode behandelt, sodass jeweils alle Fälle eingeschlossen werden können (Lüdtke, Robitzsch, Trautwein & Köller, 2007). Der Anteil fehlender Werte lag an Messzeitpunkt 1 je nach Item zwischen 0.9 % und 7.1 %, an Messzeitpunkt 2 zwischen 0.0 % und 1.7 %.

6. Ergebnisse

6.1 Deskriptive Statistiken

Die Items des Schülerfragebogens wurden zunächst deskriptiv anhand der Mittelwerte und Standardabweichungen analysiert. An beiden Messzeitpunkten zeigte sich, dass mit Ausnahme der Items zur Individualisierung alle Itemmittelwerte über dem theoretischen Mittelwert liegen. Demnach bewerten die Schüler:innen bei der großen Mehrheit der Merkmale den Unterricht als gut.

6.2 Psychometrische Eigenschaften der ursprünglichen Version des Schülerfragebogens zu Messzeitpunkt 1

Von den ursprünglich 98 Items bezogen auf 15 Skalen wurden im Rahmen der EFA 44 Items ausgeschlossen, sodass sich die nachfolgenden Analysen zunächst auf 14 Skalen bestehend aus 54 Items (2 bis 6 Items pro Skala) beziehen. Gründe für den Ausschluss von Items waren geringe Faktorladungen und Parallelladungen auf zwei unterschiedlichen Unterrichtsmerkmalen.

Aufgrund der Ergebnisse der EFA wurden zwei inhaltliche Änderungen an der Gesamtkonstruktion des Fragebogens vorgenommen. Erstens umfasste die

Skala zum Unterrichtsklima zwei Faktoren, von denen einer das Verhalten der Schüler:innen untereinander beschreibt und der andere das Verhalten der Lehrkraft. Der Lehrkraft-Faktor konnte aufgrund der geringen Itemanzahl zu Messzeitpunkt 1 zunächst nicht abgebildet werden, dies wurde bei der Überarbeitung des Schülerfragebogens aufgegriffen. Zweitens konnte die zunächst angenommene Unterscheidung von zwei unterschiedlichen Skalen zur Erfassung von Individualisierung, einmal im Bereich Unterstützung des Wissenserwerbs und einmal im Bereich Motivierung, empirisch nicht bestätigt werden. Daher wurden die Items zu einer bereichsübergreifenden Individualisierungs-Skala zusammengefasst. Diese zwei Änderungen sind im Einklang mit empirischen Analysen zum Beobachtungsbogen, bei denen sich die gleiche Faktorenstruktur zeigte (Gärtner et al., 2021).

Die Annahme von 14 inhaltlich separaten Aspekten von Unterrichtsqualität konnte durch die CFA auf Individualebene bestätigt werden. Die Fit-Indizes sprechen für eine gute Modellpassung: CFI = .95, RMSEA = .03, χ^2/df = 1.99. Alle Items laden substanziell auf den intendierten Skalen ($.52 < \lambda < .86$; s. Tabelle A1). Die Faktorkorrelationen variieren zwischen $r = -.25$ und $r = .89$ und zeigen, dass sich die theoretischen Konstrukte empirisch voneinander trennen lassen (s. Tabelle A2).

Die Ergebnisse der separaten Multilevel-CFAs zeigen, dass die theoretisch angenommene Faktorenstruktur von wenigen Ausnahmen abgesehen auch auf der Klassenebene Gültigkeit hat. Sowohl im Bereich Unterstützung des Wissenserwerbs als auch im Bereich Klassenmanagement sind gute Modellpassungen und theoriekonforme Faktorladungen und -korrelationen auf Klassenebene gegeben (Unterstützung des Wissenserwerbs: CFI = .97, RMSEA = .02, χ^2/df = 1.69, $.45 < \lambda < 1.00$, $.45 < r < .84$; Klassenmanagement: CFI = .96, RMSEA = .04, χ^2/df = 2.78, $.77 < \lambda < 1.00$, $.39 < r < .86$). Im Bereich Motivierung liegen ebenfalls eine gute Modellpassung (CFI = .96, RMSEA = .04, χ^2/df = 2.77) und substanzielle Faktorladungen ($.84 < \lambda < 1.00$) vor, allerdings weisen die Faktoren Weckung von Interesse und Förderung von Freude eine sehr hohe Korrelation von $r = .97$ auf (restliche Faktorkorrelationen $.82 < r < .90$; s. Tabellen A1 und A3). Faktorkorrelationen über $r > .90$ legen nahe, dass zwei Skalen im Wesentlichen das gleiche Konstrukt erfassen (Kline, 2011). Somit können die Schülerwahrnehmung der beiden Unterrichtsmerkmale Weckung von Interesse und Förderung von Freude auf der Klassenebene empirisch nicht klar getrennt werden. Angesichts der inhaltlichen Nähe der beiden Konstrukte ist dieser Befund theoretisch plausibel.

Schließlich wurden die Reliabilitäten der Skalen überprüft. Sowohl die internen Konsistenzen der Skalen ($.70 < \Omega < .90$) als auch die Reliabilitäten der aggregierten Klassenwerte ($.67 < ICC2 < .88$) sind als gut zu bewerten. Zudem unterscheiden sich die aggregierten Klassenwerte substanziell zwischen den Klassen. Mehrheitlich werden zwischen 20 % und 30 % der Varianz der Schülerurteile durch die Klassenzugehörigkeit bestimmt ($.11 < ICC1 < .32$; s. Tabelle 1).

6.3 Psychometrische Eigenschaften der überarbeiteten Version des Schülerfragebogens zu Messzeitpunkt 2

Zur Optimierung des Fragebogens wurden neue Items generiert, um Skalen mit wenigen Items inhaltlich zu erweitern. Insgesamt wurde der Fragebogen um 13 neue Items ergänzt, sodass alle Skalen nun vier bis sechs Items umfassen (siehe Tabelle 1). Von besonderem Interesse war hierbei, dass der auf die Lehrkraft bezogene Aspekt des Unterrichtsklimas mit der überarbeiteten Version angemessen erfasst werden kann. Der Fragebogen wurde zunächst mit einer CFA überprüft. Aufgrund der deutlich kleineren Stichprobe wurde hierbei nur eine CFA auf Individualebene berechnet. Die Ergebnisse zeigen eine zufriedenstellende Passung zwischen der theoretisch angenommenen Faktorenstruktur und den empirischen Daten (CFI = .91, RMSEA = .04, χ^2/df = 1.51). Alle Items laden substanziell auf den zugewiesenen Konstrukten ($.53 < \lambda < .90$; s. Tabelle A2). Somit erweisen sich die neu formulierten Items als faktoriell valide. Erfreulicherweise zeigte sich, dass das Unterrichtsklima in Bezug auf das Verhalten der Lehrkraft nun basierend auf einer höheren Anzahl an Items empirisch abgebildet werden kann. Die Faktorkorrelationen fallen teilweise etwas höher aus und bewegen sich jetzt bei zwei Korrelationen an der Obergrenze einer empirischen Trennbarkeit (s. Tabelle A2). Die Korrelation zwischen Weckung von Interesse und Förderung von Freude beträgt r = .90, die Korrelation zwischen Erklärung und Präsentation und Speicherung durch Üben beträgt r = .89 (restliche Korrelationen $.17 < r < .81$). Schließlich wurden die Reliabilitäten der Skalen des überarbeiteten Fragebogens analysiert (s. Tabelle 1). Die internen Konsistenzen der Einzelskalen fallen weiterhin gut aus ($.76 < \Omega < .91$). Die ICC2 fällt gut bis akzeptabel aus ($.60 < \text{ICC2} < .85$), mit Ausnahme der Skala zur Steuerung des Unterrichtsflusses (ICC2 = .49). Dies ist insofern überraschend, als dass diese Skala nicht verändert wurde. Die ICC1 verweist mehrheitlich auf bedeutende Varianzanteile zwischen den Klassen ($.07 < \text{ICC1} < .32$). Insgesamt gesehen sprechen die empirische Abbildung aller Unterrichtsmerkmale, die valide Faktorenstruktur und die mit einer Ausnahme zufriedenstellenden Reliabilitäten dafür, dass die Überarbeitung des Fragebogens zu einer Verbesserung des Instruments geführt hat.

7. Diskussion

Ausgangspunkt dieses Beitrags ist ein Ansatz für Unterrichtsevaluation, welcher auf einem theoretisch begründeten Modell von Unterrichtsqualität basiert. Dieses Modell ist insbesondere für Unterrichtsevaluation und -feedback geeignet, weil es eine klare theoretische Konzeptualisierung der Unterrichtsmerkmale liefert, den Fokus auf veränderbare Aspekte des Unterrichts legt, an konkreten Verhaltensweisen der Lehrkraft im Unterricht ansetzt und mit empirischen Befunden der Lehr-Lernforschung vereinbar ist. Auf der Grundlage dieses Modells wurde in früheren

Arbeiten ein beobachtungsbasiertes Feedbackverfahren entwickelt und erfolgreich empirisch überprüft. Die vorliegende Studie erweitert den Ansatz nun um einen Schülerfragebogen. Dieser kann entweder als Ergänzung oder als Alternative zum Beobachtungsbogen verwendet werden.

Vor dem Hintergrund des Forschungsstands zu Schülerurteilen zum Unterricht wurde bei der Itemformulierung darauf geachtet, dass keine Selbstbewertung der Schüler:innen nahegelegt wird, dass Prozesse der Unterstützung komplexer kognitiver Aktivitäten als konkrete Aktivitäten der Lehrkraft abgefragt werden und die Items sprachlich leicht verständlich sind.

Die Ergebnisse der durchgeführten Analysen zeigen insgesamt gute Kennwerte, sodass den Lehrkräften mithilfe des Schülerfragebogens ein belastbares Klassenprofil zur Wahrnehmung des Unterrichts rückgemeldet werden kann. Dieses Profil unterscheidet inhaltlich eng gefasste Unterrichtsmerkmale als Stellschrauben für die Entwicklung der Unterrichtsqualität. Die konfirmatorische Prüfung konnte mehrheitlich eine gute Passung der empirischen Daten an das theoretische Modell bestätigen. Ausnahmen betreffen die beiden begründbaren Änderungen hinsichtlich des Unterrichtsklimas (als Trennung von Lehrkraft- und Schüler:innenverhalten) und der Individualisierung (als Zusammenfassung von Aspekten der individualisierten Unterstützung des Wissenserwerbs und der Motivierung). Die CFAs belegen die empirische Trennbarkeit der Unterrichtsmerkmale. Die einzige Ausnahme bildet hierbei die hohe Korrelation zwischen Weckung von Interesse und Förderung von Freude auf der Klassenebene. Die hohe Korrelation verweist darauf, dass gegenstandsbezogene Motivation (Interesse) und tätigkeitsbezogene Motivation (Spaß am Lernen) im Unterricht häufig nahe beieinanderliegen oder – darauf verweist auch die Theorie der organismischen Integration der Motivation von Deci und Ryan (1985) – ineinander übergehen können.

Auch die Reliabilitäten der Skalen fallen gut aus, wiederum mit einer Ausnahme (Steuerung des Unterrichtsflusses). Zusammenfassend ist festzuhalten, dass sowohl faktorielle Validität als auch Reliabilität jeweils bei fast allen Skalen gewährleistet sind. Somit genügt eine Unterrichtsevaluation basierend auf dem Schülerfragebogen wichtigen psychometrischen Anforderungen.

Wir betrachten die Befunde als Bestätigung der eingangs formulierten Annahme, dass eine fächerübergreifende Modellierung von Unterrichtsqualität auf Basis klassischer Bezugstheorien trennscharf operationalisiert werden kann. Die Operationalisierung von Items auf der Grundlage der zentralen theoretischen Ansätze der Informationsverarbeitung, der Motivation und der sozialen Interaktion scheint hinsichtlich der empirischen Trennbarkeit von Anforderungsbereichen des Unterrichtens ein vielversprechender Ansatz zu sein. Die wahrgenommene Nützlichkeit der Operationalisierung für die Ableitung von konkreten Ansatzpunkten der Unterrichtsentwicklung aus Sicht von Lehrkräften muss weiterführend untersucht werden.

Einschränkend ist festzuhalten, dass die vorliegenden Analysen auf einer Stichprobe im berufsbildenden Bereich beruhen. Die Stichprobe ist somit hinsichtlich der beteiligten Fächer sehr heterogen, was im vorliegenden Fall die Überprüfung

der Messinvarianz über verschiedene Fächer hinweg nicht ermöglicht hat. Zukünftige Forschung mit dem Schülerfragebogen sollte somit folgende Fragen in den Blick nehmen:

- Replikation der Ergebnisse an einer Stichprobe aus allgemeinbildenden Schulen,
- Prüfung der Messinvarianz über Fächer, Jahrgangsstufen und Zeit,
- Validierung anhand von Außenkriterien, wie z. B. der Leistungs- oder Interessenentwicklung,
- Prüfung des Einsatzes des Schülerfragebogens in der Primarstufe (Lenske, 2016),
- Perspektivenvergleich zwischen Beobachtung und Schülerwahrnehmung.

Neben weiteren psychometrischen Analysen sollte außerdem explizit die Nützlichkeit von Unterrichtsevaluationen mithilfe des Schülerfragebogens überprüft werden. Nützlichkeit bezieht sich darauf, ob ein Unterrichtsfeedback basierend auf dem Schülerfragebogen zu einer Verbesserung der Unterrichtsgestaltung führt (Gärtner & Vogt, 2013). Im Rahmen der Studie zum beobachtungsbasierten Unterrichtsfeedback durch Schulleitungen hat sich der Vorteil der gut verständlichen Kategorisierung der drei Basisdimensionen sowie der Kommunikation der Einzelskalenergebnisse als Stellschrauben für die Unterrichtsentwicklung bestätigt (Kellermann et al., 2022). Entsprechend gehen wir davon aus, dass auch ein alleiniger Einsatz des Schülerfragebogens mit entsprechenden Erläuterungen des theoretischen Modells an die Lehrkräfte zu einer Verbesserung der Unterrichtsgestaltung führen kann (Gärtner & Vogt, 2013).

Literatur

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4), 369–406. https://doi.org/10.1037/0033-295X.89.4.369

Bandura, A. (1976). *Lernen am Modell.* Klett.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*(2), 117–148. https://doi.org/10.1207/s15326985ep2802_3

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., … Tsai, Y.-M. (2008). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente.* Max-Planck-Institut für Bildungsforschung.

Baumert, J., Gruehn, S., Heyn, S., Köller, O. & Schnabel, K.-U. (1997). *Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU). Dokumentation – Band 1. Skalen Längsschnitt I, Wellen 1–4.* Max-Planck-Institut für Bildungsforschung.

Brophy, J. (1981). Teacher praise: A functional analysis. *Review of Educational Research, 51*(1), 5–32. https://doi.org/10.3102/00346543051001005

Clausen, M. (2002). *Unterrichtsqualität: eine Frage der Perspektive?* Waxmann.

Clausen, M. & Göbel, K. (Hrsg.). (2020). Unterrichtsrückmeldungen durch Schüler*innen [Themenheft]. *Empirische Pädagogik, 34*(1).

Csikszentmihalyi, M. & Rathunde, K. (1993). The measurement of flow in everyday life: Toward a theory of emergent motivation. In R. A. Dienstbier & J. E. Jacobs

(Hrsg.), *Nebraska symposium on motivation 1992. Developmental perspectives on motivation* (S. 57–97). University of Nebraska Press.

Deci, E. L. & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behaviour*. Plenum. https://doi.org/10.1007/978-1-4899-2271-7

Doyle, W. (2006). Ecological approaches to classroom management. In C. M. Evertson & C. S. Weinstein (Hrsg.), *Handbook of classroom management. Research, practice, and contemporary issues* (S. 97–125). Erlbaum.

Eccles, J. S. & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Fauth, B., Decristan, J., Rieser, S., Klieme, E. & Büttner, G. (2014). Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg. *Zeitschrift für Pädagogische Psychologie*, *28*(3), 127–137. https://doi.org/10.1024/1010-0652/a000129

Friedrich, H. F. & Mandl, H. (1992). Lern- und Denkstrategien – ein Problemaufriß. In H. F. Friedrich & H. Mandl (Hrsg.), *Lern- und Denkstrategien. Analyse und Intervention* (S. 3–54). Hogrefe.

Gärtner, H. (2010). Wie Schülerinnen und Schüler ihre Lernumwelt wahrnehmen: Ein Vergleich verschiedener Maße zur Übereinstimmung von Schülerwahrnehmungen. *Zeitschrift für Pädagogische Psychologie*, *24*(2), 111–122. https://doi.org/10.1024/1010-0652/a000009

Gärtner, H. & Brunner, M. (2018). Once good teaching, always good teaching? The differential stability of student perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability, 30*(2), 159–182. https://doi.org/10.1007/s11092-018-9277-5

Gärtner, H., Thiel, F., Nachbauer, M. & Kellermann, C. (2021). Unterrichtsentwicklung durch Unterrichtsfeedback – Erste Entwicklungsschritte eines Beobachtungsbogens für Schulleitungen. *Zeitschrift für Bildungsforschung, 11*(3), 529–547. https://doi.org/10.1007/s35834-021-00321-7

Gärtner, H. & Vogt, A. (2013). Wie Lehrkräfte Ergebnisse eines Schülerfeedbacks verarbeiten und nutzen. *Unterrichtswissenschaft*, *41*(3), 252–267.

Göllner, R., Wagner, W., Klieme, E., Lüdtke, O., Nagengast, B. & Trautwein, U. (2016). Erfassung der Unterrichtsqualität mithilfe von Schülerurteilen: Chancen, Grenzen und Forschungsperspektiven. In Bundesministerium für Bildung und Forschung (Hrsg.), *Bildungsforschung Bd. 44: Forschungsvorhaben in Ankopplung an Large-Scale-Assessments* (S. 63–82). BMBF.

Gruehn, S. (2000). *Unterricht und schulisches Lernen. Schüler als Quellen der Unterrichtsbeschreibung*. Waxmann.

Hanisch, A.-K. (2018). *Kognitive Aktivierung im Rechtschreibunterricht: Eine Interventionsstudie in der Grundschule*. Waxmann.

Helmke, A. (2017). *Schülerfragebogen zum Unterricht*. Verfügbar unter www.unterrichtsdiagnostik.de/downloads/fragebogen/

Helmke, A., Helmke, T., Lenske, G., Pham, G., Praetorius, A.-K., Schrader, F.-W. & Ade-Thurow, M. (2018). *Unterrichtsdiagnostik mit EMU. Evidenzbasierte Methoden der Unterrichtsdiagnostik Version 7.0*. Verfügbar unter www.unterrichtsdiagnostik.de/media/files/Broschuere_Version_7.0.pdf

Heneman, H. G. & Milanowski, A. T. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education*, *17*(3), 171–195. https://doi.org/10.1023/B:PEEV.0000032427.99952.02

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hulleman, C. S. & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science, 326*(5958), 1410–1412. https://doi.org/10.1126/science.1177067

Jäger, R. & Helmke, A. (2008). *Projekt MARKUS Skalendokumentation*. Verfügbar unter www.iqb.hu-berlin.de/fdz/studies/MARKUS/MARKUSSkalendoku.pdf

Kellermann, C., Nachbauer, M., Gärtner, H. & Thiel, F. (2022). The effect of school leaders' feedback on teaching quality perceived by students – Results of an intervention study. *School Effectiveness and School Improvement. Advance online publication*. https://doi.org/10.1080/09243453.2022.2075018

Klauer, K. J. & Leutner, D. (2012). *Lehren und lernen: Einführung in die Instruktionspsychologie*. Beltz.

Klieme, E. & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 333–359). Springer. https://doi.org/10.1007/978-3-322-97590-4_12

Klieme, E. & Warwas, J. (2011). Konzepte der individuellen Förderung. *Zeitschrift für Pädagogik*, *57*(6), 805–818.

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. Guilford.

Kounin, J. S. (2006). *Techniken der Klassenführung*. Waxmann.

Kunter, M., Brunner, M., Baumert, J., Klusmann, U., Krauss, S., Blum, W., … Neubrand, M. (2005). Der Mathematikunterricht der PISA-Schülerinnen und -Schüler. *Zeitschrift für Erziehungswissenschaft*, *8*(4), 502–520. https://doi.org/10.1007/s11618-005-0156-8

Lenske, G. (2016). *Schülerfeedback in der Grundschule*. Waxmann.

Lenske, G. & Praetorius, A.-K. (2020). Schülerfeedback – was steckt hinter dem Kreuz auf dem Fragebogen? *Empirische Pädagogik*, *34*(1), 11–29.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. Probleme und Lösungen. *Psychologische Rundschau, 58*(2), 103–117. https://doi.org/10.1026/0033-3042.58.2.103

Lüdtke, O., Trautwein, U., Kunter, M. & Baumert, J. (2006). Analyse von Lernumwelten. Ansätze zur Bestimmung der Reliabilität und Übereinstimmung von Schülerwahrnehmungen. *Zeitschrift für Pädagogische Psychologie*, *20*(1/2), 85–96. https://doi.org/10.1024/1010-0652.20.12.85

McDonald, R. P. (1999). *Test theory. A unified treatment*. Erlbaum.

MET Project. (2012). *Gathering feedback for teaching. Combining high-quality observations with student surveys and achievement gains*. Bill & Melinda Gates Foundation.

Muthén, L. K. & Muthén, B. O. (1998–2017). *Mplus user's guide. Eighth edition*. Muthén & Muthén.

Nationales Bildungspanel. (2013). *Startkohorte 3: Klasse 5 (SC3). Welle 2. Erhebungsinstrumente (Feldversion)*. Verfügbar unter www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC3/5-0-0/SC3_Q_w2.pdf

Ophardt, D. & Thiel, F. (2013). *Klassenmanagement. Ein Handbuch für Studium und Praxis*. Kohlhammer.

Pianta, R. C., La Paro, K. & Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS)*. Paul H. Brookes.

Piwowar, V. (2013). Multidimensionale Erfassung von Kompetenzen im Klassenmanagement: Konstruktion und Validierung eines Beobachter- und eines Schülerfragebogens für die Sekundarstufe. *Zeitschrift für Pädagogische Psychologie, 27*(4), 215–228. https://doi.org/10.1024/1010-0652/a000108

Praetorius, A.-K., Klieme, E., Herbert, B. & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three basic dimensions. *ZDM* –

Mathematics Education, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K. & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002

Prediger, S. & Neugebauer, P. (2021). Capturing teaching practices in language-responsive mathematics classrooms. Extending the TRU framework "teaching for robust understanding" to L-TRU. *ZDM – Mathematics Education, 53*(2), 289–304. https://doi.org/10.1007/s11858-020-01187-1

Rakoczy, K., Buff, A. & Lipowsky, F. (2005). Teil 1: Befragungsinstrumente. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis"*. GFPF.

Rheinberg, F. (2008). *Motivation*. Kohlhammer.

Röhl, S. & Rollett, W. (2020). Alles nur die Nettigkeit der Lehrkraft? Die Communion der Lehrkräfte als Erklärung für den Halo-Bias in Schülerbefragungen zur Unterrichtsqualität. *Empirische Pädagogik*, *34*(1), 30–45.

Schiefele, U. (2009). Situational and individual interest. In K. R. Wenzel & A. Wigfield (Hrsg.), *Educational psychology handbook series. Handbook of motivation at school* (S. 197–222). Routledge/Taylor & Francis Group.

Seidel, T., Prenzel, M., Duit, R. & Lehrke, M. (2003). *Technischer Bericht zur Videostudie „Lehr-Lern-Prozesse im Physikunterricht"*. Leibniz-Institut für die Pädagogik der Naturwissenschaften.

Steffensky M. & Neuhaus B. J. (2018). Unterrichtsqualität im naturwissenschaftlichen Unterricht. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Theorien in der naturwissenschaftsdidaktischen Forschung* (S. 299–313). Springer. https://doi.org/10.1007/978-3-662-56320-5_18

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Thiel, F. (2016). *Interaktion im Unterricht*. Barbara Budrich. https://doi.org/10.36198/9783838545714

Thiel, F., Ophardt, D. & Piwowar, V. (2013). *Kompetenzen des Klassenmanagements. Schülerfragebogen*. Verfügbar unter www.ewi-psy.fu-berlin.de/einrichtungen/arbeitsbereiche/schulentwicklungsforschung/forschung/abgelaufeneprojekte/kodek/kodek/Buch_Dokumente/KODEK-S_Lehrerin.pdf

Thiel, F., Richter, S. G. & Ophardt, D. (2012). Steuerung von Übergängen im Unterricht. Eine Experten-Novizen-Studie zum Klassenmanagement. *Zeitschrift für Erziehungswissenschaft*, *15*(4), 727–752. https://doi.org/10.1007/s11618-012-0325-5

Thiel, F., Tarkian, J., Lankes, E.-M., Maritzen, N., Ricke-Baulecke, T. & Kroupa, A. (Hrsg.). (2019). *Datenbasierte Qualitätssicherung und -entwicklung in Schulen. Eine Bestandsaufnahme in den Ländern der Bundesrepublik Deutschland*. Springer. https://doi.org/10.1007/978-3-658-23240-5

Wagner, W. (2008). *Methodenprobleme bei der Analyse der Unterrichtswahrnehmung aus Schülersicht – am Beispiel der Studie DESI der Kultusministerkonferenz*. Universität Koblenz-Landau.

Wagner, W., Göllner, R., Helmke, A., Trautwein, U. & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction*, *28*, 1–11. https://doi.org/10.1016/j.learninstruc.2013.03.003

Appendix

Tabelle A1: Itemstatistiken (Mittelwert, Standardabweichung und Faktorladung λ aus CFA und CFA_{2level}) zu MZP 1 und MZP 2

	MZP 1						MZP 2			
Item	*N*	*M*	*SD*	λ_{CFA}	$\lambda_{CFA\text{-}level_1}$	$\lambda_{CFA\text{-}level_2}$	*N*	*M*	*SD*	λ_{CFA}
Unterstützung des Wissenserwerbs – Erklärung und Präsentation										
Unsere Lehrkraft präsentiert neues Wissen klar und strukturiert.	1287	3.21	0.70	.73	.64	.99	293	3.27	0.67	.79
Unsere Lehrkraft kann schwierige Themen gut erklären.	1271	3.12	0.74	.75	.69	.98	294	3.24	0.72	.82
Unsere Lehrkraft bringt anschauliche Beispiele, die uns das Verstehen der Lerninhalte erleichtern.	1269	3.21	0.75	.66	.59	.88	294	3.36	0.64	.67
Wenn unsere Lehrkraft uns etwas erklärt, wissen wir immer ganz genau, was sie meint.							292	2.98	0.67	.73
Unterstützung des Wissenserwerbs – Unterstützung des Verstehens										
Unsere Lehrkraft lässt uns unsere Aufgabenlösungen selbst korrigieren.	1245	2.64	0.90	.52	.47	.79	293	3.02	0.75	.65
Unsere Lehrkraft stellt Aufgaben, bei denen mehrere Lösungswege möglich sind.	1248	2.79	0.84	.66	.63	.86	294	3.14	0.65	.65
Unsere Lehrkraft lässt unterschiedliche Lösungswege von Aufgaben vergleichen und bewerten.	1254	2.88	0.82	.71	.68	1.00	292	3.14	0.74	.64
Bei unserer Lehrkraft kann ich zum Lösen schwieriger Aufgaben meine eigenen Strategien einsetzen.	1242	2.98	0.80	.56	.59	.45	292	3.16	0.75	.67
Unsere Lehrkraft stellt Aufgaben, für deren Lösung man sich Zeit zum Nachdenken nehmen muss.							293	2.81	0.88	.53
Unterstützung des Wissenserwerbs – Speicherung durch Üben										
Wir erhalten gezielt Übungsaufgaben für besonders schwierige Themen.	1255	2.89	0.81	.58	.53	.90	294	3.09	0.80	.65
Im Unterricht gibt es regelmäßig Phasen, in denen wir Übungsaufgaben bearbeiten.	1264	3.19	0.76	.54	.51	.86	294	3.25	0.79	.71
Zu Übungsaufgaben oder Hausaufgaben erhalten wir immer eine Rückmeldung, was an unseren Lösungen richtig oder falsch ist.	1260	3.07	0.86	.53	.49	.84	293	3.11	0.85	.64
Unter den Übungsaufgaben, die uns unsere Lehrkraft gibt, sind oft Aufgaben, bei denen man sieht, ob man etwas wirklich verstanden hat.	1254	2.98	0.78	.68	.63	1.00	293	3.15	0.70	.70
Die Übungsaufgaben im Unterricht unserer Lehrkraft sind ähnlich, aber doch immer wieder anders, so dass man genau aufpassen muss.	1248	2.96	0.77	.62	.55	.98	290	3.06	0.71	.61
Unterstützung des Wissenserwerbs – Anwendung und Transfer										
Bei unserer Lehrkraft nutzen wir unser Wissen zur Bearbeitung von Alltagsproblemen.	1270	2.63	0.87	.61	.57	.67	293	2.85	0.86	.70
Unsere Lehrkraft gibt uns Problemaufgaben, in denen wir unser Wissen anwenden müssen.	1261	2.94	0.79	.57	.57	.90	294	3.22	0.69	.66

Tabelle A1 wird fortgesetzt

Fortsetzung Tabelle A1

Item	MZP 1						MZP 2			
	N	*M*	*SD*	λ_{CFA}	$\lambda_{CFA\text{-}level_1}$	$\lambda_{CFA\text{-}level_2}$	*N*	*M*	*SD*	λ_{CFA}
Wir machen im Unterricht kleine Projekte, bei denen wir das, was wir gelernt haben, anwenden können.	1256	2.74	0.95	.56	.51	.67	294	2.89	0.87	.55
Unsere Lehrkraft stellt uns häufig Aufgaben, bei denen wir unser Wissen auf neue Situationen übertragen müssen.	1258	2.83	0.79	.70	.67	.99	294	3.09	0.72	.81
Wenn wir im Unterricht unserer Lehrkraft üben, wenden wir das Gelernte oft auf andere Dinge an.	1244	2.83	0.77	.62	.54	.93	293	2.99	0.75	.67
Motivierung – Weckung von Interesse										
Unsere Lehrkraft kann Schülerinnen und Schüler manchmal richtig begeistern.	1263	2.91	0.86	.76	.70	.94	293	3.02	0.82	.79
Unsere Lehrkraft kann auch trockene Themen wirklich interessant machen.	1267	2.80	0.85	.80	.73	1.00	293	2.97	0.81	.79
Auch bei bekannten Themen überrascht uns unsere Lehrkraft oft mit neuem Wissen.	1256	2.89	0.82	.72	.65	.95	292	3.05	0.77	.78
Unsere Lehrkraft steckt uns mit ihrer Begeisterung an.	1257	2.74	0.87	.79	.73	.98	293	2.92	0.87	.85
Unsere Lehrkraft gestaltet den Unterricht oft sehr spannend.	1263	2.76	0.88	.82	.77	.98	292	2.90	0.79	.80
Die von unserer Lehrkraft behandelten Inhalte sind richtig spannend.	1266	2.64	0.88	.79	.74	.93	292	2.74	0.81	.78
Motivierung – Förderung von Freude										
Die Bearbeitung der Aufgaben, die unsere Lehrkraft stellt, macht mir Spaß.	1261	2.76	0.85	.83	.78	.98	292	2.80	0.79	.83
Bei unserer Lehrkraft ist man oft so mit Aufgaben beschäftigt, dass man gar nicht merkt, wie die Zeit vergeht.	1257	2.82	0.91	.72	.64	.96	291	2.75	0.85	.73
Im Unterricht unserer Lehrkraft bin ich mit Freude dabei.	1261	2.94	0.84	.85	.81	1.00	289	2.98	0.80	.83
Der Unterricht unserer Lehrkraft macht mir Spaß.	1259	2.98	0.87	.86	.82	.98	291	3.01	0.79	.86
Motivierung – Verdeutlichung des Nutzens										
Unsere Lehrkraft sagt uns, wofür die Lerninhalte nützlich sind.	1264	3.15	0.77	.73	.69	.93	293	3.35	0.71	.80
Unsere Lehrkraft informiert uns darüber, dass die Lerninhalte wichtig für andere Themen sind.	1261	3.12	0.77	.69	.62	.94	293	3.38	0.66	.74
Unsere Lehrkraft erklärt uns, warum es für das spätere Leben wichtig ist, das Unterrichtsthema zu lernen.	1255	2.96	0.91	.75	.69	.88	292	3.19	0.83	.81
Unsere Lehrkraft macht deutlich, dass das Unterrichtsthema auch im Alltag wichtig ist.	1252	2.86	0.91	.73	.67	.84	291	3.09	0.83	.77
Motivierung – Förderung von Selbstwirksamkeit										
Unsere Lehrkraft muntert uns auf, damit wir auch bei schwierigen Aufgaben nicht den Mut verlieren.	1251	2.90	0.84	.78	.71	.97	291	3.01	0.81	.78
Unsere Lehrkraft lobt unsere Fortschritte.	1252	2.99	0.83	.74	.70	.97	291	3.19	0.84	.80

Tabelle A1 wird fortgesetzt

Fortsetzung Tabelle A1

Item	MZP 1						MZP 2			
	N	*M*	*SD*	λ_{CFA}	$\lambda_{CFA\text{-}level_1}$	$\lambda_{CFA\text{-}level_2}$	*N*	*M*	*SD*	λ_{CFA}
Unsere Lehrkraft zeigt uns, welche Fortschritte wir durch Anstrengung erreichen können.	1241	2.85	0.84	.73	.69	.94	292	3.07	0.84	.82
Unsere Lehrkraft traut uns etwas zu.	1254	3.17	0.74	.67	.64	.90	291	3.34	0.74	.78
Klassenmanagement – Lerndienliches Unterrichtsklima SuS										
Im Unterricht gehen die Schülerinnen und Schüler respektvoll miteinander um.	1260	3.14	0.85	.74	.81	.98	291	3.16	0.80	.83
Im Klassenzimmer nehmen alle aufeinander Rücksicht.	1251	3.00	0.92	.81	.86	1.00	291	3.00	0.86	.90
Im Unterricht beachten die Schülerinnen und Schüler die Regeln für einen guten Umgang miteinander.							292	3.08	0.77	.90
Wenn es im Unterricht zu Konflikten kommt, versuchen die Schülerinnen und Schüler, gemeinsam eine Lösung zu finden.							291	3.03	0.85	.75
Klassenmanagement – Lerndienliches Unterrichtsklima LK[1]										
Unsere Lehrkraft hat eine gute Beziehung zu den Schülerinnen und Schülern.							293	3.35	0.67	.72
Unsere Lehrkraft verhält sich uns gegenüber fair.							293	3.42	0.70	.83
Wir vertrauen unserer Lehrkraft.							291	3.40	0.74	.86
Unsere Lehrkraft ist immer freundlich zu den Schülerinnen und Schülern.							290	3.43	0.65	.81
Klassenmanagement – Monitoring										
Unsere Lehrkraft bekommt mit, was die Schülerinnen und Schüler im Klassenzimmer machen.	1254	3.20	0.79	.74	.72	.97	293	3.27	0.77	.86
Unsere Lehrkraft merkt, was beim Rest der Klasse vorgeht, auch wenn sie mit einzelnen Schülerinnen und Schülern beschäftigt ist.	1248	3.04	0.79	.81	.77	.99	293	3.17	0.76	.85
Unsere Lehrkraft merkt sofort, wenn Schülerinnen und Schüler beginnen, sich mit etwas anderem zu beschäftigen.	1245	2.96	0.84	.79	.77	.97	293	3.16	0.77	.78
Unsere Lehrkraft hat immer die ganze Klasse im Blick.	1248	3.07	0.82	.81	.78	1.00	291	3.15	0.79	.86
Klassenmanagement – Steuerung des Unterrichtsflusses										
Bei unserer Lehrkraft kommt es im Unterricht häufig zu Verzögerungen und Wartezeiten. (umgepolt)	1251	2.82	0.88	.69	.67	.82	292	2.14	0.80	.71
Unsere Lehrkraft schweift häufig vom Thema ab. (umgepolt)	1241	2.98	0.89	.68	.66	.77	291	1.93	0.81	.61
Bei unserer Lehrkraft nehmen Übergänge von einem Unterrichtsabschnitt zum nächsten (z. B. vom Klassengespräch zur Stillarbeit) oft viel Zeit in Anspruch. (umgepolt)	1240	2.69	0.85	.60	.58	.89	293	1.93	0.75	.62
Bei unserer Lehrkraft dauert es zu Beginn der Stunde sehr lange, bis die Schülerinnen und Schüler ruhig werden und zu arbeiten beginnen. (umgepolt)	1260	3.02	0.94	.75	.67	.99	291	1.70	0.79	.73

Tabelle A1 wird fortgesetzt

Fortsetzung Tabelle A1

Item	MZP 1						MZP 2			
	N	*M*	*SD*	λ_{CFA}	$\lambda_{CFA\text{-}level_1}$	$\lambda_{CFA\text{-}level_2}$	*N*	*M*	*SD*	λ_{CFA}
Klassenmanagement – Gruppenaktivierung										
Im Unterricht unserer Lehrkraft beteiligen sich meist alle Schülerinnen und Schüler.	1253	2.74	0.77	.74	.68	.96	293	2.75	0.73	.81
Bei unserer Lehrkraft nimmt meist die gesamte Klasse aufmerksam am Unterrichtsgeschehen teil.	1248	2.76	0.81	.81	.78	1.00	294	2.97	0.73	.85
Unsere Lehrkraft schafft es, dass sich meist alle Schülerinnen und Schüler mit dem Unterrichtsthema beschäftigen.							290	2.85	0.76	.86
Bei unserer Lehrkraft werden alle Schülerinnen und Schüler in den Unterricht mit einbezogen.							292	3.10	0.72	.67
Klassenmanagement – Umgang mit Störungen										
Bei unserer Lehrkraft können wir im Unterricht ungestört arbeiten.	1258	3.06	0.80	.67	.61	.95	294	3.10	0.74	.60
Unsere Lehrkraft reagiert angemessen auf Störungen (also nicht zu streng und nicht zu mild).	1242	3.11	0.85	.65	.61	.77	291	3.24	0.74	.67
Unserer Lehrkraft gelingt es, Störungen schnell zu beenden.	1250	3.15	0.81	.84	.81	.99	293	3.29	0.73	.88
Unsere Lehrkraft reagiert so auf Störungen, dass die Schülerinnen und Schüler aufhören zu stören.	1250	3.11	0.81	.81	.78	.97	292	3.28	0.73	.87
Individualisierung										
Unsere Lehrkraft gibt uns je nach Leistungsstärke unterschiedliche Aufgaben.	1253	1.96	0.99	.69			294	2.21	0.92	.76
Unsere Lehrkraft gibt uns je nach Interesse unterschiedliche Aufgaben.	1244	2.07	0.94	.83			293	2.41	0.96	.69
Unsere Lehrkraft gibt uns Inhalte und Themen zur Auswahl.	1231	2.20	0.95	.70			294	2.57	0.94	.72
Unsere Lehrkraft gibt guten Schülerinnen und Schülern Extraaufgaben, die schwieriger sind.							293	2.22	0.94	.84
Wenn Schülerinnen und Schüler etwas nicht verstehen, erhalten sie zusätzliche Aufgaben, um noch einmal zu üben.							292	2.43	0.94	.78
Unsere Lehrkraft stellt verschiedene Aufgaben, so dass alle Schülerinnen und Schüler Spaß am Lernen haben.							294	2.68	0.89	.74

Anmerkungen. MZP = Messzeitpunkt; CFA = konfirmatorische Faktorenanalyse; λ_{CFA} = Faktorladungen aus Gesamtmodellprüfung auf Schülerebene; $\lambda_{CFA\text{-}level_1}$ = Faktorladungen auf Schülerebene aus Modellprüfung pro Anforderungsdimension mit 2 Ebenen; $\lambda_{CFA\text{-}level_2}$ = Faktorladungen auf Klassenebene aus Modellprüfung pro Anforderungsdimension mit 2 Ebenen.

Tabelle A2: Faktorinterkorrelationen zu MZP 1 (unterhalb der Diagonale) und MZP 2 (über der Diagonale) aus CFA auf Schülerebene

Unterrichtsmerkmal	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
Unterstützung des Wissenserwerbs															
F1 Erklärung und Präsentation		.71	.74	.67	.79	.74	.64	.81	.39	.71	.64	.51	.62	.63	.39
F2 Unterstützung des Verstehens	.49		.89	.79	.62	.58	.64	.68	.24	.64	.56	.37	.46	.49	.48
F3 Speicherung durch Üben	.72	.65		.80	.63	.61	.75	.76	.23	.72	.56	.45	.55	.61	.53
F4 Anwendung und Transfer	.63	.59	.74		.68	.68	.74	.65	.33	.60	.50	.38	.51	.51	.49
Motivierung															
F5 Weckung von Interesse	.76	.54	.65	.70		.90	.64	.80	.33	.65	.59	.40	.70	.59	.53
F6 Förderung von Freude	.71	.46	.59	.59	.89		.60	.73	.40	.61	.52	.44	.64	.54	.46
F7 Verdeutlichung des Nutzens	.60	.44	.63	.71	.69	.64		.72	.30	.63	.54	.25	.49	.47	.45
F8 Förderung von Selbstwirksamkeit	.74	.55	.67	.61	.79	.75	.68		.29	.80	.67	.36	.63	.63	.55
Klassenmanagement															
F9 Lerndienliches Unterrichtsklima SuS	.30	.19	.24	.22	.28	.28	.19	.29		.32	.39	.43	.45	.45	.25
F10 Lerndienliches Unterrichtsklima LK											.64	.35	.46	.58	.35
F11 Monitoring	.61	.39	.53	.49	.61	.58	.56	.61	.33			.39	.53	.69	.30
F12 Steuerung des Unterrichtsflusses	.30	.02	.17	.02	.10	.12	-.02	.12	.23		.23		.42	.54	.17
F13 Gruppenaktivierung	.58	.36	.46	.48	.63	.64	.51	.62	.50		.57	.17		.62	.46
F14 Umgang mit Störungen	.62	.36	.54	.45	.53	.52	.44	.59	.49		.64	.36	.68		.35
F15 Individualisierung	.33	.48	.40	.54	.54	.48	.50	.45	.17		.31	-.25	.46	.24	

Anmerkungen. MZP = Messzeitpunkt; CFA = konfirmatorische Faktorenanalyse; SuS = Schüler:innen; LK = Lehrkraft.

Tabelle A3: Faktorinterkorrelationen zu MZP 1 auf Klassenebene

Unterrichtsmerkmal	F1	F2	F3	F4	F5	F6	F7	F8	F9	F11	F12	F13	F14
Unterstützung des Wissenserwerbs													
F1 Erklärung und Präsentation													
F2 Unterstützung des Verstehens	.59												
F3 Speicherung durch Üben	.84	.69											
F4 Anwendung und Transfer	.85	.45	.80										
Motivierung													
F5 Weckung von Interesse													
F6 Förderung von Freude					.97								
F7 Verdeutlichung des Nutzens					.90	.90							
F8 Förderung von Selbstwirksamkeit					.90	.87	.82						
Klassenmanagement													
F9 Lerndienliches Unterrichtsklima SuS													
F11 Monitoring									.40				
F12 Steuerung des Unterrichtsflusses									.61	.39			
F13 Gruppenaktivierung									.72	.80	.50		
F14 Umgang mit Störungen									.70	.83	.75	.86	

Anmerkungen. MZP = Messzeitpunkt; SuS = Schüler:innen.

Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01.08
Volume 14 (2022), No. 1, 174–185
CC BY-NC-SA 4.0 Waxmann 2022

Marion Händel, Svenja Bedenlier, Michaela Gläser-Zikuda,
Rudolf Kammerl, Bärbel Kopp & Albert Ziegler

Do Higher Education Students Have the Means to Learn During the COVID-19 Pandemic? Student Demands for Distance Learning in a Suddenly Digital Landscape

Abstract

Across the globe, 2020 terms began under conditions incited by the corona pandemic. Within a relatively short amount of time, universities started to develop and implement online courses for distance learning. The current study is about an online survey at a German university investigating the unique circumstances under which students began the digital 2020 summer term. Of approximate-

---

PD Dr. Marion Händel (corresponding author), ORCID: 0000-0002-3069-5582 · Prof. Dr. Albert Ziegler, ORCID: 0000-0002-2122-0673, Universität Erlangen-Nürnberg, Lehrstuhl für Pädagogische Psychologie und Exzellenzforschung, Regensburger Str. 160, 90478 Nürnberg, Germany
email: marion.haendel@uni-a.de
albert.ziegler@fau.de

Prof. Dr. Svenja Bedenlier, ORCID: 0000-0002-7940-5232, Universität Erlangen-Nürnberg, ILI – Institut für Lern-Innovation, Dr.-Mack-Straße 77, 90762 Fürth, Germany
email: svenja.bedenlier@ili.fau.de

Prof. Dr. Michaela Gläser-Zikuda, ORCID: 0000-0002-3071-2995, Universität Erlangen-Nürnberg, Lehrstuhl für Schulpädagogik mit dem Schwerpunkt empirische Unterrichtsforschung, Regensburger Str. 160, 90478 Nürnberg, Germany
email: michaela.glaeser-zikuda@fau.de

Prof. Dr. Rudolf Kammerl, ORCID: 0000-0001-5500-0487, Universität Erlangen-Nürnberg, Lehrstuhl für Pädagogik mit dem Schwerpunkt Medienpädagogik, Regensburger Straße 160, 90478 Nürnberg, Germany
email: rudolf.kammerl@fau.de

Prof. Dr. Bärbel Kopp, Universität Erlangen-Nürnberg, Lehrstuhl für Grundschulpädagogik und -didaktik mit dem Schwerpunkt Lehren und Lernen, Regensburger Straße 160, 90478 Nürnberg, Germany
email: baerbel.kopp@fau.de

Marion Händel is now at the University of Augsburg.

ly 38 500 students, N = 5563 students from across all institutional faculties took part in the survey. Results indicate how well students are equipped with devices for digital learning, what kind of experiences they have already made with online learning, and how competent they reported feeling regarding digital learning. The study provides important insights into digital higher education during the exceptional pandemic situation. The results are intended to feed into student counselling systems via support by way of access to devices or courses regarding digital skills, or through counselling for students with special social burdens.

Keywords
higher education, e-learning, digital readiness, COVID-19 pandemic

Verfügen Studierende über die Mittel und Wege, um während der Coronaviruspandemie zu lernen? Bedarfe für Online-Lehre in einer plötzlich digitalisierten Landschaft

Zusammenfassung
Weltweit begann das Jahr 2020 unter dem Einfluss der Coronapandemie. Innerhalb sehr kurzer Zeit begannen Universitäten, Online-Kurse für den Fernunterricht zu entwickeln und implementieren. Die vorliegende Studie untersuchte anhand einer Online-Umfrage an einer deutschen Universität die besonderen Umstände, unter denen Studierende das digitale Sommersemester 2020 begannen. Von etwa 38 500 Studierenden nahmen N = 5563 Studierende aller Fakultäten an der Umfrage teil. Die Ergebnisse zeigen, wie gut die Studierenden mit Geräten für das digitale Lernen ausgestattet sind, welche Erfahrungen sie zuvor bereits mit dem Online-Lernen gemacht hatten und wie kompetent sie sich in Bezug auf das digitale Lernen fühlen. Die Studie liefert wichtige Einblicke in die digitale Hochschulbildung während der außergewöhnlichen Pandemiesituation. Die Ergebnisse können in studentische Beratungssysteme einfließen, sei es durch Unterstützung beim Zugang zu Geräten, durch Kurse zu digitalen Schlüsselkompetenzen oder durch Beratung von Studierenden mit besonderen sozialen Belastungen.

Schlagworte
Hochschulbildung, E-Learning, Digital Readiness, Coronapandemie

1. Introduction

The corona pandemic led to an ad hoc changeover at universities across the globe from predominantly face-to-face teaching to exclusively digitally supported teaching formats (“emergency remote teaching”; Bond et al., 2021; Hodges et al., 2020). Studies on remote learning provide important insights into student equipment and

engagement. However, those studies do not transfer to the current situation of distance learning in a suddenly digital landscape. Hence, the current study investigates student technical equipment, experiences, digital skills, and further possible individual and contextual challenges at a German university.

Studies on technical equipment and media use by students suggest that higher education students have a range of digital devices at their disposal (e.g., Dolch & Zawacki-Richter, 2018; Sailer et al., 2018) – although inter-individual differences are to be expected. For example, the media-use study by Zawacki-Richter et al. (2016) revealed that female students use social networks more often during their studies than did male students. Differences are furthermore to be expected between traditional and non-traditional students (e.g., those with children; see Paulsen & McCormick, 2020) as well as due to study subject. For example, the study by Weber et al. (2019) found information-seeking behavior relative to study subject: students of the humanities or social sciences more often consult libraries but less often use online resources for information seeking.

In emergency remote teaching, students' digital readiness might contribute to how students cope with the change to online formats. Inter alia, skills regarding the use of digital tools as well as information seeking or information sharing behavior seem to be important (Hong & Kim, 2018; Rubach & Lazarides, 2019; Weber et al., 2019). For example, students with high self-confidence in the use of digital technologies seem to have significantly fewer barriers in social interaction or motivation (Muilenburg & Berge, 2005).

Accordingly, the success of the switch to digital teaching in the pandemic situation may depend on access to digital devices, previous experience in media use, as well as students' digital skills. In contrast to students purposefully enrolling in a course of study at a distance teaching university (cf. Stöter et al., 2014), students during the corona pandemic did not voluntarily choose to participate in a digital semester themselves but had to adopt a quick and pragmatic approach in the wake of the pandemic (Kerres, 2020). Due to contact restrictions, students were more dependent than usual on their personal equipment because access to computer rooms could no longer be provided at universities. Furthermore, students were less able to receive technical assistance through personal contact. Finally, subgroups of students with children might have faced further challenges due to a lack of childcare (daycare and school closures).

2. Aims of the Study

The aim of the study is to assess how students are equipped for the first pandemic online semester by way of equipment with digital devices, e-learning experiences, and digital skills. The study aims to identify groups of students with digital learning accessibility difficulties. In line with previous study results, gender- and subject-specific differences can be assumed. Furthermore, differences among study

experience and domestic condition are considered. We investigate two research questions (RQ):

RQ 1: How well are higher education students prepared for a complete shift to digital distance learning?

RQ2: What are the specific preconditions of higher education students with respect to their expected degree, study subject, gender, or domestic conditions?

3. Method

The study reports the first measurement of an online survey study during the summer semester 2020 (Bedenlier et al., 2021; Händel et al., 2020), to which all students at a full university were invited. The current report refers to data of the week before the start of the official lecture period, where all students enrolled at the university were invited via email to take part in a survey on the general conditions of digital teaching.

3.1 Instruments

Students were asked about their workplace equipment (quiet workplace, internet access) as well as PC devices (desktop, tablet, notebook, min = 0, max = 3). Furthermore, experiences with nine e-learning offerings (see Table 2; Froebus & Bender, 2019) at the university were recorded: For each of the nine university e-learning offerings, students indicated whether they had used them before (min = 0, max = 9).

Finally, two four-item subscales (Digital Tool Application – DTA, and Information-Sharing Behavior – ISB) from the Digital Readiness for Academic Engagement Questionnaire (DRAE; Hong & Kim, 2018) were used to assess students' self-assessed digital skills. The items had to be answered on a 6-point Likert scale (1 = *not at all correct* to 6 = *completely correct*); DTA: $\alpha = .76$, for example, "I can use software or apps on a computer or mobile device"; and ISB: $\alpha = .83$, for example, "I can interact with fellow students using real-time communication media, e.g., video conferencing tools or messenger services."

3.2 Sample

In total, 5563 students (which corresponds to 15% of all students) took part in the survey; 62% of the respondents were female, 38% male, 0.3% non-binary; 14% did not provide any information on gender. The participating students were on average 23.2 years old (*SD* = 4.5) and were enrolled in their 4.5 semester (*SD* = 2.8). Less than 5% of the respondents had at least one child to look after in the house-

hold, which corresponds to the distribution of students in Germany (Middendorff et al., 2017). Table 1 shows the relative frequencies in the surveyed sample. Students across all five faculties participated in the survey and the participation distribution corresponds well with the distribution within the university.

Table 1: Sample Separately Reported by Faculty and Expected Degree

	Students [%]
Faculty	
Faculty of Humanities, Social Sciences, and Theology	27.6
Faculty of Sciences	12.8
Faculty of Business, Economics, and Law	20.7
Faculty of Engineering	23.7
Faculty of Medicine	15.2
Expected degree	
Bachelor	37.2
Master	24.3
State exam	35.4
Doctoral exam	1.8
Others	1.4

3.3 Data Analysis

Descriptive statistics are reported to illustrate the current pandemic situation in higher education with regard to student digital equipment, relevant experience, and skills. Digital skills was correlated with (a) equipment and (b) e-learning experience. Finally, multivariate (co-)variance analyses were calculated to investigate possible group differences between subject cultures (operationalized via the respective faculty), gender, and required child care.

4. Results

4.1 Descriptive Statistics

Table 2 illustrates the availability of each workstation element and whether students had already used available e-learning features.

Table 2: Availability of Technical Devices and Use of Previous E-Learning Offerings

Equipment	*M*	*SD*
Work area	.91	.29
Internet access	1.00	.07
Desktop-PC	.31	.46
Notebook	.91	.29
Tablet-PC	.48	.50
Sum score PC devices	1.70	0.71
E-learning offering		
Downloadable lecture notes/literature	.92	.27
Lecture recordings	.67	.47
Live streams of lectures	.06	.24
Live media in courses	.49	.50
Online learning modules	.52	.50
Online communication and collaboration	.63	.48
Other online-supported learning opportunities	.35	.48
E-tests	.43	.50
Online self-tests for self-control	.53	.50
Sum score e-learning experiences	4.57	2.09

Note. 0 = no; 1 = yes.

Less than 1% had neither a notebook, nor a desktop/tablet PC available. About half of the e-learning offerings had already been used by students, with a maximum for downloadable scripts and a minimum for live streams.

With regard to the assessed skills, the students considered themselves to be rather competent in the use of digital tools (DTA; $M = 4.64$, $SD = 0.91$) and in the sharing of digital information (ISB; $M = 5.03$, $SD = 0.94$).

Bivariate correlations indicate significant but weak correlations between equipment and previous e-learning experiences with self-assessed skills ($.16 \leq r \leq .20$, $ps < .001$).

4.2 Interindividual Differences in Equipment, Experience, and Self-Assessed Digital Skills

4.2.1 Differences Relative to Desired Degree

First, this study explored whether the desired degree – Bachelor's (BA) vs. Master's (MA) degree – had an influence on the handling of e-learning offers or digital skills. As expected, MA students reported a higher experience with universi-

ty e-learning offerings (M = 4.83, SD = 1.84) than did BA students (M = 4.33, SD = 2.16), $F(1, 3350) = 50.76$, $p < .001$, $\eta^2 = .02$. Although both groups of students considered themselves relatively competent, the values among MA students were significantly higher (DTA: M = 5.00, SD = 0.80; ISB: M = 5.30, SD = 0.86) than among BA students (DTA: M = 4.67, SD = 0.89; ISB: M = 5.03, SD = 0.92), $F(1, 3350) = 116.26$, $p < .001$, $\eta^2 = .03$, and $F(1, 3350) = 72.14$, $p < .001$, $\eta^2 = .02$. No significant differences were found with respect to computer equipment, $F(1, 3350) = 1.72$, $p = .19$.

4.2.2 Gender Differences

Male and female students significantly differed with regard to computer equipment, $F(1, 4707) = 71.40$, $p < .001$, $\eta^2 = .02$. Male students were slightly better equipped with digital devices (M = 1.81 devices, SD = 0.73) than were female students (M = 1.63 devices, SD = 0.68). There were no significant gender differences in the experience of university e-learning offerings, $F(1, 4707) = 1.97$, $p = .16$. However, differences were found with regard to self-assessed computer skills. Male students rated their digital information sharing skills higher (M = 5.26, SD = 0.88) than did female students (M = 4.92, SD = 0.95), $F(1, 4707) = 146.25$, $p < .001$, $\eta^2 = .03$. A strong effect was shown with regard to the use of digital tools in favor of male students (M = 5.14, SD = 0.76) when compared to female students (M = 4.38, SD = 0.84), $F(1, 4707) = 957.25$, $p < .001$, $\eta^2 = .17$.

4.2.3 Differences Relative to Subject Groups

Significant subject differences of small effect size were found, DTA: $F(4, 5481) = 136.14$, $p < .001$, $\eta^2 = .09$; ISB: $F(4, 5481) = 40.12$, $p < .001$, $\eta^2 = .03$; experience with university e-learning offerings: $F(4, 5481) = 25.40$, $p < .001$, $\eta^2 = .02$; computer equipment: $F(4, 5481) = 9.92$, $p < .001$, $\eta^2 = .01$) (see Table 3 for descriptive values). Tukey post hoc tests indicate that for DTA and ISB, students from the Faculty of Engineering have higher scores than all other students. Students from the Faculty of Humanities, Social Sciences, and Theology as well as from the Faculty of Medicine have lower scores than do other students (ps < .01). Students from the Faculty of Sciences and of Business, Economics, and Law have similar self-assessed skills.

Similarly, students of the Faculty of Engineering are best equipped, and significantly better than those of the Faculty of Humanities, Social Sciences, and Theology and the Faculty of Business, Economics, and Law (ps < .05). Finally, students of the Faculty of Humanities, Social Sciences, and Theology are worse equipped than those of the Faculty of Medicine ($p < .001$).

Deviating from this pattern, results concerning experiences with university e-learning offerings are highest in the Faculty of Business, Economics, and Law

($ps < .001$) and lower in the Faculty of Sciences than in all other faculties ($ps < .01$) – with the exception of the Faculty of Humanities, Social Sciences, and Theology, where students similarly reported less experience with e-learning.

All effects remained significant when controlling for gender, although the effect sizes were smaller.

Table 3: Descriptive Values of Digital Skills, Previous E-Learning Experience, and Technical Equipment, Separately for the Students of Each Faculty

	DTA		ISB		E-Learning		Equipment	
Faculty	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Faculty of Humanities, Social Sciences, and Theology	4.43	0.90	4.88	0.97	4.38	2.11	1.62	0.70
Faculty of Sciences	4.68	0.90	5.03	0.93	4.25	1.87	1.69	0.70
Faculty of Business, Economics, and Law	4.61	0.86	5.08	0.92	5.10	1.97	1.69	0.72
Faculty of Engineering	5.09	0.80	5.28	0.88	4.67	1.98	1.77	0.71
Faculty of Medicine	4.34	0.91	4.86	0.93	4.63	2.47	1.77	0.67
Total	4.64	0.91	5.03	0.94	4.62	2.10	1.70	0.70

Note. DTA = digital tool application; ISB = information-sharing behavior.

4.2.4 Domestic Conditions

Students with children may face particular challenges due to corona-related school and daycare closures. Significant differences were found with regard to the workplace. While 92% of the students without children had a quiet workplace, this was true for only 64% of the students with children, $F(1, 5500) = 217.46$, $p < .001$, $\eta^2 = .04$. No significant differences were found in PC equipment, $F(1, 5500) < 1$, or in experience with university e-learning offerings, $F(1, 5500) = 2.29$, $p = .13$. However, students with children considered themselves less digitally skilled (DTA: $M = 4.64$, $SD = 0.91$; ISB: $M = 5.05$, $SD = 0.93$) than students without children (DTA: $M = 4.49$, $SD = 0.96$; ISB: $M = 4.69$, $SD = 1.18$), $F(1, 5500) = 7.32$, $p = .01$, $\eta^2 = .00$, and $F(1, 5500) = 33.10$, $p < .001$, $\eta^2 = .01$, for DTA and ISB respectively.

5. Discussion

The present study arose in the context of the radical switch to digital teaching exacerbated by the corona pandemic (cf. Kerres, 2020). An important feature of the survey is that, to obtain an inventory that was as unbiased as possible, the study was carried out before the onset of digital classes in the 2020 summer semester.

Consistent with previous surveys (e.g., Sailer et al., 2018; Zawacki-Richter et al., 2016), almost all students appear to have access to a desktop PC, notebook or tablet PC. This is a promising precondition for the start of a digital semester (RQ1). Still, universities should not leave any student behind, and take account of students without access to technology for participating in an ad hoc digital semester. Individual digital skills were rated as rather high, with the ability to share digital information more pronounced than the ability to use different programs (Rubach & Lazarides, 2019).

The correlative analyses indicate an interrelation between equipment features and previous e-learning experiences with self-assessed skills. Although both constructs were surveyed simultaneously and thus, no causal conclusions can be drawn, this can still be interpreted as a possible indication that both the use of technical equipment and the experience gained lead to higher competences (Senkbeil et al., 2019).

Finally, group comparisons were made based on existing subject groups, study experience, gender, and domestic conditions to identify which groups of students start the digital semester well prepared (RQ2). Regarding technical equipment, differences were found in line with expectations. Overall effect sizes can be regarded as small. Still, due to differences with regard to gender, degree (BA vs. MA), and subjects, university management and deans of the faculties need to be aware that some groups of students might need support in order to achieve equity in digital distance learning. While no significant computer equipment differences were found for students with or without children, students with children were less likely to have an available and quiet workplace, that is, these students face more difficult study conditions not exclusively attributable to digital teaching (because this is precisely where digital teaching formats could also have advantages for students with children; see Zawacki-Richter et al., 2016), but rather due to a lack of childcare facilities.

Furthermore, significant differences in self-assessed digital skills were found that are in line with the pattern of results of *actual* measured competencies (Senkbeil et al., 2019). Hence, assuming that self-reported skills correlate with actual competencies, students enrolled in specific faculties might need training to achieve ICT literacy. Particularly with regard to the use of digital applications, considerable and large gender differences were found to the disadvantage of female students, which is in agreement with results regarding computer self-concept (Lim, 2001).

5.1 Limitations and Implications

A first limitation relates to the sample, which might be a selective sample – either due to personal interest in the survey, or due to lacking internet access. Second, the results refer to only one university and do not provide detailed information on individual courses of study, socioeconomic status, or migration background. Third, digital skills were not recorded on the basis of a detailed competence test (cf. Senk-

beil et al., 2019) or behavioral data, but rather on two selected scales of self-assessed digital skills.

Still, the study is of relevance because it underscores the circumstances imposed upon students at the beginning of the 2020 summer semester. The survey provides important observations of student groups that need special support, in particular – either through courses on the use of digital media or through the consideration of their individual study situation, like one in which supplemental childcare is lacking or altogether absent. Of course, digital teaching formats that were developed within a few weeks cannot be compared with established digital or blended learning formats that have been planned by expert committees for the long term (cf. Hetzner & Schmidt, 2018; Kerres, 2020; Moskal et al., 2013). Nonetheless, the distance learning experiences brought about by the corona pandemic might contribute to a stronger and more equitable foundation for digital learning services in the long term. In addition to this stocktaking, a continuous and process-oriented investigation of the study conditions in this exceptional semester is required. That is, what (additional) resources are used by students during the semester; how do students develop computer skills; and what are the consequences of (sudden) digitalization in terms of student engagement (Bond et al., 2020) or student achievement.

Acknowledgements

We would like to thank the university management for supporting our study. We thank Désirée Bender and Nicole Froebus for their support in the preparation and implementation of the survey. We furthermore thank Miguelina Nuñez for her support in proof reading the manuscript.

This paper reports on a pre-registered study: https://osf.io/68aus.

References

Bedenlier, S., Händel, M., Kammerl, R., Gläser-Zikuda, M., Kopp, K., & Ziegler, A. (2021). Akademische Mediennutzung Studierender im Corona-Semester 2020 – Digitalisierungsschub oder weiter wie bisher? [Student academic media use in the corona semester 2020 – digitalisation boost or business as usual?]. *Medienpädagogik, 40*, 229–252. https://doi.org/10.21240/mpaed/40/2021.11.18.X

Bond, M., Bedenlier, S., Marín, V. I., & Händel, M. (2021). Emergency remote teaching in higher education: Mapping the first global online semester. *International Journal of Educational Technology in Higher Education, 18,* Article 50. https://doi.org/10.1186/s41239-021-00282-x

Bond, M., Buntins, K., Bedenlier, S., Zawacki-Richter, O., & Kerres, M. (2020). Mapping research in student engagement and educational technology in higher education: A systematic evidence map. *International Journal of Educational Technology in Higher Education, 17*, Article 2. https://doi.org/10.1186/s41239-019-0176-8

Dolch, C., & Zawacki-Richter, O. (2018). Are students getting used to Learning Technology? Changing media usage patterns of traditional and non-traditional students in higher education. *Research in Learning Technology*, *26*. https://doi.org/10.25304/rlt.v26.2038

Froebus, N., & Bender, D. (2019). *Abschlussbericht zur FAU-Studierendenbefragung FAU-St 2019* [Final report on the FAU student survey FAU-St]. Friedrich-Alexander-Universität Erlangen-Nürnberg.

Händel, M., Stephan, M., Gläser-Zikuda, M., Kopp, B., Bedenlier, S., & Ziegler, A. (2022). Digital readiness and its effects on higher education students' socio-emotional perceptions in the context of the COVID-19 pandemic. *Journal of Research on Technology in Education, 54*(2), 267–280. https://doi.org/10.1080/15391523.2020.1846147

Hetzner, S., & Schmidt, C. (2018). Pimp your lecture: Digitalisierung der Lehre als Motor für Veränderung. Auftrag an die Organisationsentwicklung der Hochschule [Pimp your lecture: Digitisation of teaching as a driver for change. A mandate for the organisational development of the university]. *Synergien(n!). Beiträge zum Qualitätspakt Lehre im Jahre 2017, 24*, 33–44. https://www.universitaetskolleg.uni-hamburg.de/publikationen/uk-schriften-024.pdf

Hodges, C. B., Moore, S., Lockee, B. B., Trust, T., & Bond, M. A. (2020, March 27). The difference between emergency remote teaching and online learning. *EDUCAUSE Review*. https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning

Hong, A. J., & Kim, H. J. (2018). College students' digital readiness for academic engagement (DRAE) scale: Scale development and validation. *The Asia-Pacific Education Researcher, 27*(4), 303–312. https://doi.org/10.1007/s40299-018-0387-0

Kerres, M. (2020). Against all odds: Education in Germany coping with Covid-19. *Postdigital Science and Education, 2*(3), 690–694. https://doi.org/10.1007/s42438-020-00130-7

Lim, C. K. (2001). Computer self-efficacy, academic self-concept, and other predictors of satisfaction and future participation of adult distance learners. *American Journal of Distance Education, 15*(2), 41–51. https://doi.org/10.1080/08923640109527083

Middendorff, E., Apolinarski, B., Becker, K., Bornkessel, P., Brandt, T., Heißenberg, S., & Poskowsky, J. (2017). *Zusammenfassung zur 21. Sozialerhebung des Deutschen Studentenwerks – durchgeführt vom Deutschen Zentrum für Hochschul- und Wissenschaftsforschung* [Summary of the 21st social survey of the German student union – conducted by the German Centre for Higher Education Research and Science Studies]. Bundesministerium für Bildung und Forschung. https://www.studentenwerke.de/sites/default/files/se21_zusammenfassung_hauptbericht.pdf

Moskal, P., Dziuban, C., & Hartman, J. (2013). Blended learning: A dangerous idea? *The Internet and Higher Education*, *18*, 15–23. https://doi.org/10.1016/j.iheduc.2012.12.001

Muilenburg, L. Y., & Berge, Z. L. (2005). Student barriers to online learning: A factor analytic study. *Distance Education, 26*(1), 29–48. https://doi.org/10.1080/01587910500081269

Paulsen, J., & McCormick, A. C. (2020). Reassessing disparities in online learner student engagement in higher education. *Educational Researcher, 49*(1), 20–29. https://doi.org/10.3102/0013189X19898690

Rubach, C., & Lazarides, R. (2019). Eine Skala zur Selbsteinschätzung digitaler Kompetenzen bei Lehramtsstudierenden [A measure for teacher students' self-evaluated digital competencies scale]. *Zeitschrift für Bildungsforschung, 9*(3), 345–374. https://doi.org/10.1007/s35834-019-00248-0

Sailer, M., Schultze-Pernice, F., Chernikova, O., Sailer, M., & Fischer, F. (2018). *Digitale Bildung an bayerischen Hochschulen – Ausstattung, Strategie, Qualifizierung und Medieneinsatz* [Digital education at Bavarian universities – Equipment, strategy, qualification, and media use]. Vereinigung der Bayerischen Wirtschaft e. V. https://www.vbw-bayern.de/Redaktion/Frei-zugaengliche-Medien/Abteilungen-

GS/Bildung/Archiv/2018/Downloads/FINAL_Digitale_Bildung_an_bayerischen_Hochschulen-2018.pdf

Senkbeil, M., Ihme, J. M., & Schöber, C. (2019). Wie gut sind angehende und fortgeschrittene Studierende auf das Leben und Arbeiten in der digitalen Welt vorbereitet? [Are first-semester and advanced university students ready for life and work in the digital world? Results of a standard setting method to describe ICT-related proficiency levels]. *Zeitschrift für Erziehungswissenschaft, 22*(6), 1359–1384. https://doi.org/10.1007/s11618-019-00914-z

Stöter, J., Bullen, M., Zawacki-Richter, O., & von Prümmer, C. (2014). From the back door into the mainstream. Learner characteristics in the context of lifelong learning. In T. Anderson & O. Zawacki-Richter (Eds.), *Online distance education. Towards a research agenda* (pp. 421–457). Athabasca University Press. https://www.aupress.ca/app/uploads/120233_99Z_Zawacki-Richter_Anderson_2014-Online_Distance_Education.pdf

Weber, H., Becker, D., & Hillmert, S. (2019). Information-seeking behaviour and academic success in higher education: Which search strategies matter for grade differences among university students and how does this relevance differ by field of study? *Higher Education, 77*(4), 657–678. https://doi.org/10.1007/s10734-018-0296-4

Zawacki-Richter, O., Kramer, C., & Müskens, W. (2016). Studiumsbezogene Mediennutzung im Wandel – Querschnittdaten 2012 und 2015 im Vergleich [Study-related media use in transition. Cross-sectional data 2012 and 2015 in comparison]. *Schriftenreihe zum Bildungs- und Wissenschaftsmanagement*, *1*. https://openjournal.uni-oldenburg.de/index.php/bildungsmanagement/article/view/101