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

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.