------Reading comprehension? Reading ability?-------

------Zeitformen?----

-----konsistent: Students/participants?-------

-----level or scale?-----

-----Reinbringen, dass wir ELFE vielleicht auch nicht als Gold-Standard sehen, den es zu erreichen gilt-

-----growth or change?-----

-----Time points or testing points?-----

-----Formative assessment, CBM and LPA-----

----How to operationalize reading fluency/what is reading fluency?-----

**Abstract**

* Aim: To draw a comprehensive picture of the Testreihe‘s? validity
* Construct validity
  + Factorial validity
  + Convergent/divergent validity
* Criterion validity
  + Predictive validity

**Introduction**

* Importance of developing reading ability
* Difficulty: individual student needs
* Need for individual facilitation
* Basis for individual facilitation: information about students’ needs
* Monitoring student’s progress in reading useful to get this information
* CBM
* LPA: new one developed (needs to meet psychometric standards like any other test)
* Aim: To draw a comprehensive picture of the LPA’s validity
* Validity (Comprising among other aspects)
  + Construct validity
    - Factorial validity
    - Convergent/divergent validity
  + Criterion validity
    - Predictive validity  
      (define status validity/ validity of change measurement)
* Challenges when validating LPA opposed to single time point test scores (what is special here?)

The ability to read provides a foundation to achieve any further knowledge and skills. As of today, professional as well as personal life in most domains is unthinkable without knowing how to read. Without reading, we would not be able to receive any information in written form or use a computer, let alone navigate in public transport or purchase everyday goods. Therefore, teaching this ability represents a main concern in primary schools. Starting from the beginning of the first grade, a huge amount of time is invested in laying the foundations for this skill, ever developing it further from then on. A helpful tool to facilitate this process can be to monitor students’ progress, being useful to adapt instructional programs to students’ needs and to support student learning (Förster & Souvignier, 2014; Stecker, Fuchs, & Fuchs, 2005). A potential approach to progress monitoring is learning progress assessment (LPA). In LPA, short, equivalent tests are applied over the school year to assess students’ development. Like any other assessments, these tests must meet psychometric standards. In this article, we aim to draw a comprehensive picture of a newly developed LPA series’ validity. While the validity of any instrument is critical, the validity of progress assessments comprises two different stages (Fuchs, 2004). In a first step, the validity of a single test administered at a single point in time needs to be examined (status validity). In a second step, it needs to be verified that an increase in test scores indeed goes along with an increase in the targeted competence (validity of change measurement). This second step is deemed important as data about changes in scores (i.e. slope data) is used to inform instructional decisions.

*Learning Progress Assessment*

Teaching a whole class of students how to read is a demanding task. Students are a heterogenous group, resulting in individually different requirements and challenges in learning how to read. This heterogeneity even is on the rise (Mayr, J., Neuweg, G. H., 2009), possibly due to factors such as migration, especially impacting the level of early reading ability (Suchan, Wallner-Paschon, Stöttinger, & Bergmüller, 2007). Consequently, individual learning progress can easily differ. For example, the choice of instruction type can facilitate or disrupt the learning progress in reading depending on the individual student. Connor, Morrison and Petrella (2004) found that students already showing higher reading comprehension as opposed to their fellow students, benefited from learning tasks managed by the children themselves, while these learning tasks had negative effects for lower-performing students. These students, in turn, benefited more from learning tasks managed by the teacher. [Zweites Beispiel dafür, wie leicht sich der Lernfortschritt unterscheidet]. As illustrated by these examples, adapting instruction for learning how to read to individual students’ needs seems important to ensure a beneficial learning progress for all students in a class. In Germany, this need even has a legal basis. Here, individual facilitation is prescribed by law in most federal states (Fischer & Rott, 2014, p.40f). In order to match individual students’ needs, as a first step, we must know what specifically those needs are. A means for providing this diagnostic information in the case of reading ability can be to explicitly monitor learning progress. Monitoring learning progress provides valuable diagnostic information [continuous feedback] (Förster & Souvignier, 2014) – both to teachers and to students. Teachers can evaluate whether their instructional program works as expected and students can judge how precisely they estimated their own performance. A framework for assessing learning progress with respect to reading ability was built by Deno (Quelle einfügen) with the formulation of curriculum-based measurement (CBM). In CBM, short, equivalent tests are applied at short intervals over the course of the school year to assess students’ learning progress (Fuchs 2004, Quelle einfügen). These tests measure reading ability as established by a curriculum (defining designated end of period performance), thus allowing to model learning progress over testing points. Using data from this kind of learning progress assessment (LPA) as diagnostic information to adapt instructional programs has proven to be effective. In a meta-analysis of 21 studies, student performance was higher when using CBM, with an average effect size of .70 (Fuchs & Fuchs, 1986). Sure enough, this advantage of using CBM comes at a cost, i.e. the time necessary to realize the repeated LPAs (Fuchs & Fuchs, 1986). What is more, there is the cost of developing the LPAs, especially in ensuring the single tests to be equivalent. This aspect is crucial to model learning progress over testing points (Tagungsbeitrag Earli 2017, Quelle einfügen).

*Newly developed LPA series*

Tagungsbeitragsautoren efficiently tackled the task of developing a series of equivalent LPAs for reading ability by using a rule-based item design. For each item, four different properties (like the number of syllables) were varied according to construction rules allowing to generate items with specific difficulties and thus equivalent tests. This offers a promising approach for inexpensively developing LPA series assuredly meeting the important psychometric property of having equivalent testing points.

*Validity of LPA*

Another important psychometric property of the newly developed LPA series, like that of any assessment, is the instrument’s validity. While the validity of any instrument is critical, the validity of an LPA series comprises two different stages (Fuchs, 2004, Quelle einfügen).

Studien aus Fuchs 2004 zu research stages von CBM

Recherche: Validity of LPA

Verbleibende zwei Studien quop

Anforderungen an die Konstruktion von LPA (Salaschek 2014) reinbringen und damit auch eine Beschreibung, wie "reading ability" eigentlich genau erfasst wird und warum und dann iwann vllt auch Souvigniers Anmerkung von wegen nicht zu ähnlich zu ELFE?

In the years since Deno (1985) conceptualized the approach of CBM, many progress assessments have been developed and validated, thereby mainly focusing on the aspect of status validity (e.g. Fore, Boon, Burke, & Martin, 2008; Valladolid, 2015; Vanderheyden et al., 2016). Tolar and colleagues (2014), in contrast, examined in how far the slope of a progress assessment can predict student outcomes at the end of an academic year. Still, validity here was judged with reference to a single status measurement. Considering the validity of change measurement, research in the field is scarce.

**Research Questions**

1. Status Validity: Does one static test score at one specific point in the school year reflect the current level of reading ability (the level of reading ability at that specific point in the school year)?
2. Validity of Change Measurement: Does the change in test scores over the school year (time?) reflect the change in reading ability over the school year (time?)?
3. Predictive Validity: Does the change of test scores over the school year (time?) predict end of period reading ability?

**Method**

Participants

* Sample A: 86 classes with 1635 students (only LPA)
* Sample B: 18 classes with 368 students (LPA + pre- and post-assessment)
* A sample of N = 2003 elementary school students

Participants and Design

A sample of N = 2003 second-grade elementary school students (Mage, SDage, % female) completed eight LPA reading tests over the course of their school year with gaps of approximately three weeks between LPA testing points. Additionally, we administered standardized paper-pencil tests at the beginning and in the end of the schoolyear to a subsample of 368 students. Prior to the LPA, this subsample completed a standardized reading ability measure, as well as one for mathematic ability and one for intelligence. Besides, we collected teacher ratings of the student’s reading ability. After the LPA, the students again completed the reading ability measure.

Instruments

* LPA

The LPA consisted of a series of eight parallel computer-based tests assessing reading comprehension on word-, sentence- and text-level. On word-level, in a word/pseudo-word tasks, participants had to decide whether a word existed or not. On sentence-level, in a sentence verification task, participants had to judge whether a sentence made sense or not. On text-level, in a text completion task, students had to decide whether a sentence reasonably continued a short story. For examples of items for all levels see […]. For each item, both accuracy (correct/incorrect answer) and response time were recorded. Students had to complete one LPA test approximately every three weeks during the school year.

* Standardized Tests

Prior and subsequently to the LPA, we assessed the student’s reading comprehension on word-, sentence- and text-level using ELFE. On word-level, in a […]-task, participants had to decide which of three words corresponded best to an image. On sentence-level, in a […]-task, participants had to […]. On text-level, in a […]-task, participants had to […].

Prior the the LPA, we assessed the student’s mathemtical ability and intelligence using DEMAT and CFT. The DEMAT measures mathematical ability as performance on […]. The CFT measures intelligence […].

* Teacher Ratings

Prior tot he LPA, we collected teacher ratings of the student’s reading comprehension consisting of one dimensional and one criterial judgment per student. Dimensionally, teachers judged their student’s reading comprehension on a scale of one to seven. Criterially, teachers judged how many words, sentences and texts their students would be able to read within two minutes of time.

Analytic Strategy

* LPA Scores

To integrate accuracy (how correctly items were answered) and response time (how fast items were answered) into an efficacy measure of reading comprehension, we calculated the correct item summed residual time (CISRT) per word-, sentence- and text-level (see equation 1). Doing so, we determined a range of valid processing times: For each item, a lower threshold was set to the 5%-quantile and an upper threshold was set to the 95%-quantile of all response times thus incorporating the average 90% of response times into the valid range[[1]](#footnote-1). The CISRT then represents the amount of processing time from this range one still has left after correctly responding to an item. As an example, if the thresholds for an item are set to 1 and 11 seconds, the range of valid processing times covers 10 seconds. For a correct answer after 5 seconds, 6 of 10 seconds or 60% of the time are left. As we can see from this example, higher scores correspond to a better performance, as they imply faster correct responses.

* Reliability

Before investigating the LPA’s validity, we briefly looked into its reliability constituting an essential basis for further validity analyses. We calculated the LPA’s retest reliability between testing points for each level as well as the split-half reliability for each time point and level, dividing the item pool by odd and even item numbers. Regarding retest reliability, we want to point out that high retest reliability coefficients might conflict with capturing growth between testing points.

* Construct Validity

*Factorial Validity*. Regarding the LPA’s dimensionality, we postulate a three-factor model, where the items of each level (word, sentence and text) load on their own factor, respectively (see figure […]). We conducted a CFA assuming this model structure for each testing point of the LPA to assure that the dimensionality doesn’t change over time. In order to estimate the CFA models, we split the items for each level into three parcels built by counterbalancing item positions. Accordingly, for e.g. six items, the first and the fourth item would go into the first parcel, the second and the fifth item would go into the second parcel and the third and the sixth item would go into the third parcel.

*Status Validity.* To investigate the convergent validity of LPA scores from a single point in time, firstly, we correlated them with ELFE pre and post scores (obtained prior and subsequently to the LPA) on each level. Secondly, we correlated the LPA scores with the dimensional and criterial teacher ratings on each level. As indicators of divergent validity, the total LPA scores (summarized over all levels) were correlated with scores for mathematical ability and intelligence, respectively.

*Validity of Change Measurement.* To investigate the validity of change measurement, we are interested to see in how far growth in scores over the course of the LPA corresponds to growth between ELFE pre and post scores. Using Structural Equation Modeling, in order to describe growth in the LPA, we constructed a latent growth model with a slope factor representing linear growth over testing points. For the growth between ELFE pre and post test, we constructed a latent change model incorporating a difference factor representing linear growth between pre and post test. To obtain a convergent validity indicator, we considered the correlation between the LPA slope factor and the ELFE pre-post difference factor.

* Criterion Validity

*Predictive Validity.* Predicting end of period reading comprehension, we intend to explain performance in the ELFE post test through the ELFE pre score, and the LPA intercept and slope factors. This represents a rather strict operationalization, as we predict the ELFE post score with an earlier score from the same test. Finding an incremental influence of the LPA slope factor on the ELFE post score would thus provide a strong indication of predictive validity.

**Results**

**Discussion**

* Change validity only convergent (because no divergent measures at the end of the school year)

**Fragen**

* Warum nehmen wir den Intercept bei der prädiktiven Validität mit auf?
* Zeitform Methodenteil?

**Archiv**

They are on the one hand expected to cover a range of contents in a limited amount of time while on the other hand making sure that all students sufficiently learn without leaving any student behind.

As students have individual needs in learning how to read, teaching this skill to a whole class of students is challenging for teachers.

We assessed reading comprehension on word-, sentence- and text-level prior and subsequently to the LPA using ELFE. To capture word comprehension, an image and a choice of three words are presented to participants who have do decide which word corresponds best to the image. On sentence-level, participants had to […]. The test operationalizes text comprehension with a task […]

To address individual students’ needs in a class environment limiting resources for intensive individual support While Addressing student heterogeneity, explicitly monitoring learning progress in reading has been found to be useful to adapt instructional programs to students’ needs and to support student learning (Förster & Souvignier, 2014; Stecker et al., 2005). What is more, in Germany, tailoring instructional programs to students’ needs even is stipulated by law . To meet these requirements, progress monitoring

using a standardized treatment of all students .As we know, students have different needs in learning how to read, so that not every approach works best with every student.] Students already showing higher reading comprehension as opposed to other students from their class benefited from child-managed instructional activities, while these had negative effects for lower-performing students. These students, in turn, benefited more from teacher-managed instructional activities .

Based on the concept of CBM (Deno, 1985; Fuchs, Fuchs, & Hamlett, 2007), a potential approach to progress monitoring is learning progress assessment (LPA). In LPA, short, equivalent tests are applied over the school year to assess students’ development. Like any other assessments, these tests must meet psychometric standards. While the validity of any instrument is critical, the validity of progress assessments comprises two different stages (Fuchs, 2004). In a first step, the validity of a single test administered at a single point in time needs to be examined (status validity). In a second step, it needs to be verified that an increase in test scores indeed goes along with an increase in the targeted competence (validity of change measurement). This second step is deemed important as data about changes in scores (i.e. slope data) is used to inform instructional decisions.

In the years since Deno (1985) conceptualized the approach of CBM, many progress assessments have been developed and validated, thereby mainly focusing on the aspect of status validity (e.g. Fore, Boon, Burke, & Martin, 2008; Valladolid, 2015; Vanderheyden et al., 2016). Tolar and colleagues (2014), in contrast, examined in how far the slope of a progress assessment can predict student outcomes at the end of an academic year. Still, validity here was judged with reference to a single status measurement. Considering the validity of change measurement, research in the field is scarce.

The CISRT represents the amount of time that one still has left after successfully responding to an item. Therefore, higher scores correspond to a better performance. The CISRT is measured in percent of the total processing time thus ranging from zero to 100. We determined the amount of total processing time to incorporate the average 90% response times per item. Consequently, the fastest and slowest 5% of response times per item were excluded from the analysis.

The CISRT is based on a minimal and maximal processing time per item: For each item, a lower threshold is set to the 5%-quantile and an upper threshold is set to the 95%-quantile of all response times thus incorporating the average 90% of response times into the CISRT. The CISRT represents

The CISRT is based on a minimal and maximal processing time, which are determined by considering the average 90% of response times.

References

Cecil Fore, III, Boon, R. T., Burke, M. D., & Martin, C. (2008). Validating Curriculum-Based Measurement for Students With Emotional and Behavioral Disorders in Middle School. *Assessment for Effective Intervention*, *34*(2), 67–73. https://doi.org/10.1177/1534508407313234

Connor, C. M., Morrison, F. J., & Petrella, J. N. (2004). Effective Reading Comprehension Instruction: Examining Child x Instruction Interactions. *Journal of Educational Psychology*, *96*(4), 682–698. https://doi.org/10.1037/0022-0663.96.4.682

Deno, S. L. (1985). Curriculum-Based Measurement: The Emerging Alternative. *Exceptional children*, *52*(3), 219–232.

Fischer, C., & Rott, D. (2014). *Individuelle Förderung als schulische Herausforderung* (1. Aufl.). *Schriftenreihe des Netzwerk Bildung: Vol. 31*. Berlin: Friedrich-Ebert-Stiftung Abt. Studienförderung.

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

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, *53*(3), 199–208. https://doi.org/10.1177/001440298605300301

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (2007). Using curriculum-based measurement to inform reading instruction. *Reading and Writing*, *20*(6), 553–567. https://doi.org/10.1007/s11145-007-9051-4

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 : the Official Journal of the Division of School Psychology, American Psychological Association*, *31*(3), 369–382. https://doi.org/10.1037/spq0000138

Mayr, J., Neuweg, G. H. (2009). Lehrer/innen als zentrale Ressource im Bildungssystem: Rekrutierung und Qualifizierung. In Werner Specht (Ed.), *Nationaler Bildungsbericht Österreich 2009: 2. Fokussierte Analysen Bildungspolitischer Schwerpunktthemen* (pp. 341–360). Graz: Leykam.

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

Suchan, B., Wallner-Paschon, C., Stöttinger, E., & Bergmüller, S. (2007). *PIRLS 2006. Internationaler Vergleich von Schülerleistungen. Erste Ergebnisse. Lesen in der Grundschule*: Leykam.

Valladolid, V. C. (2015). Development and Validation of Curriculum-Based Measurement (CBM) for Identifying Students with Reading Difficulties. *Online Submission*, *6*(1), 2–10.

Vanderheyden, A. M., Broussard, C., Fabre, M., Stanley, J., Legendre, J., & Creppell, R. (2016). Development and Validation of Curriculum-Based Measures of Math Performance for Preschool Children. *Journal of Early Intervention*, *27*(1), 27–41. https://doi.org/10.1177/105381510402700103

1. Consequently, the fastest and slowest 5% of response times per item were excluded from the analysis. [↑](#footnote-ref-1)