-------Am Ende: Möglichst wenig Passiv?------

------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, 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. While the validity of any instrument is critical, the validity of progress assessments comprises more than one stage (Fuchs, 2004). In a first step, like with any test, the validity of a single test administered at a single point in time needs to be examined (status validity). Secondly, 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. The current study aims to draw a comprehensive picture of a newly developed LPA series’ validity.

*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, & 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). 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). CBM was originally introduced to monitor poorly performing students by frequently measuring their skill level with brief tests. This monitoring has proven to be effective, as in a meta-analysis of 21 studies, student performance was higher when using CBM, with an average effect size of .70 (L. S. Fuchs & Fuchs, 1986). As another, yet closely related approach to CBM (Förster & Souvignier, 2014), learning progress assessment (LPA) aims to monitor the progress of all students in a class. Implementing this in an efficient manner, every three weeks, a web-based test is administered to each student. These equivalent tests measure reading ability as established by a curriculum (defining designated end of period performance), thus allowing to model learning progress over testing points.

Sure enough, this advantage of using CBM comes at a cost, i.e. the time necessary to realize the repeated LPAs (L. S. 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, 1985b) conceptualized the approach of CBM, many progress assessments have been developed and validated, thereby mainly focusing on the aspect of status validity (Cecil Fore, Boon, Burke, & Martin, 2008; Furey, Marcotte, Hintze, & Shackett, 2016; 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 equivlaent 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 (Werner Specht, 2009)year)
* Research on optimizing trend estimation hasgenerally focused on three questions: (1) Howshould we calculate slope (van Norman, Christ, & Zopluoglu, 2013). Vielleicht in Diskussion, dass auch wenn slope valide ist, das trotzdem auch geklärt werden muss. In genannter Quelle wird OLS als favorisierte Methode genannt, aber später revidiert der Autor das in einer anderen Veröffentlichung:

**Fragen**

* Warum nehmen wir den Intercept bei der prädiktiven Validität mit auf?
* Zeitform Methodenteil?
* **Ist unsere Art den Slope zu schätzen wie OLS?** Denn das ist laut mehrere neuerer Studien kein guter Weg (z.B. (Solomon & Forsberg, 2017))
  + In jedem Fall aber: Wie man den Growth für die Lehrer darstellt (z.B. durch Linie aus OLS oder anders) hat impact (unterschiedlich technically adequate).

**Neuer Ansatz 030220**

The ability to read provides a foundation to achieve any further knowledge and skills. 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, providing objective information about reading achievement. Based on this information, instructional programs can be adapted to students’ needs to support student learning (Förster & Souvignier, 2014, 2014; Stecker et al., 2005). A potential approach to progress monitoring is learning progress assessment (LPA) (Förster & Souvignier, 2014). 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 poses a somewhat special case. Here, we need not only to examine the validity of a single test score at a single point in time. Rather, we must shift our attention to a change in scores over time (i.e. slope data) (L. S. Fuchs, 2004). This shift is deemed important as slope data is especially used to inform instructional decisions. The current study aims to draw a comprehensive picture of the validity of a newly developed LPA series targeting reading ability.

*Assessing Learning Progress*

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 et al., 2007). Consequently, individual learning progress can easily differ. To nonetheless succeed in facilitating each student individually, teachers need data about student achievement. Using this data to inform instructional decisions, teachers can improve student performance (Borman et al., 2016; Carlson, Borman, & Robinson, 2017; Slavin, Cheung, Holmes, Madden, & Chamberlain, 2013). Especially collecting student achievement data on a regular basis, i.e. monitoring learning progress repeatedly, provides valuable diagnostic information in a temporal resolution (Förster & Souvignier, 2014). Teachers can evaluate how well students respond to their instructional program (Johnson, Jenkins, & Petscher, 2010) and whether it should be adapted (Shapiro, Edwards, & Zigmond, 2005). 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). CBM was originally introduced to monitor poorly performing students by frequently measuring their skill level with brief tests (Deno, 2003). [Einfügen, dass growth/slope/change genutzt wird und dass das effective ist] This monitoring has proven to be effective, as in a meta-analysis of 21 studies, student performance was higher when using CBM, with an average effect size of *d*=.7 (L. S. Fuchs & Fuchs, 1986). Also Stecker et al. (2005) conclude in their review that CBM is associated with higher student performance. As an alternate, yet closely related approach to CBM, LPA aims to monitor the progress of all students in a class. Implementing this with respect to the usually tight schedule of a school day, every three weeks, a web-based test taking approximately 20 minutes is administered to each student. These equivalent tests measure reading ability as established by a curriculum (defining designated end of period performance), thus allowing to model learning progress over testing points. Like CBM, the application of LPA turns out to be beneficial for student achievement (Förster, Kawohl, & Souvignier, 2018; Förster & Souvignier, 2015). Obviously, this can only be true if the LPA scores represent a valid operationalization of learning progress.

*Validitiy of LPA*

In order to examine the validity of a progress monitoring instrument, traditional validity analyses need to be extended. In the case of CBM, (L. S. Fuchs, 2004) identifies two research stages with respect to an instrument’s validity. In a first step, like with any test, we must ascertain that a single test score from a single point in time reflects the current level of the targeted competence (status validity). Secondly, we must verify that change in test scores over time reflects change in the targeted competence (validity of change measurement). Only then, valuable diagnostic information offered by CBM can be used beneficially. Integrating the concepts of status validity and validity of change measurement into the traditional triad of content, construct and criterion validity, we can locate both concepts under the aspect of construct validity. Status validity corresponds to traditional examinations of convergent and divergent validity, i.e. comparing a test with other tests measuring similar or different competences. Here, we want to consider convergent and divergent validity with respect to one static score from one specific point in time. Validity of change measurement conveys the concepts of convergent and divergent validity into a temporal dimension. Here, we want to consider convergent and divergent validity not with respect to a static score, but with respect to a change in scores over time. In addition to extending construct validity, validity of change measurement also impacts criterion validity. When using CBM, teachers evaluate the effectiveness of their instructional program based on the slope of CBM scores over time (Deno, 2003). Therefore, criterion-related validity indices should be based on slope data, too. In this process, we cannot simply assume that a positive slope is equivalent to an increasing competence. (Christ & Ardoin, 2009) identified a huge variation in CBM scores within a short amount of time, destabilizing growth estimates. For an overview of further sources of instability, see (van Norman et al., 2013). Consequently, in addition to verifying status validity, ensuring validity of change measurement should be a default step when dealing with CBM and other progress monitoring instruments.

While there are studies incorporating slope data when analyzing criterion validity as well as studies focusing on status validity, there are nearly no studies dealing with the validity of change measurement, let alone all three aspects.

Tolar 2014 ab 1.2 lesen!!!!

However, we can find a lot of studies focusing on status validity or predicting outcomes with growth, but nearly none judging validity of change measurement. mehr oder fast ausschließlich was zu typischen Validitätssachen wie Statusvalidität (hoffentlich).

Es gibt einige Simulationsstudien, die sich damit beschäftigen, wie man change in scores over time schätzen sollte (z.B. OLS oder Bayesian) (und dann Validität als correlation of true and observed scores) (also true slope vs. observed slope?). Die können zwar Hinweise darauf geben, wie der slope technisch bestimmt werden sollte, aber nicht darauf, ob ein konkretes reales Verfahren valide ist. Bsp. Für so eine Studie ist Christ, Desjardins 2012

Literatursuche mit cbm and validity: Viele Studien zu Kriteriumsvalidität (Nutzen die static scores oder growth: Es gibt auch welche mit growth (s. citavi)). Also vielleicht: Viele zu statusvalidität und Kriteriumsvalidität, aber wenig zu Validität der Veränderungsmessung (hoffentlich)!

Dann Studien aus conference paper zu status validity

Shifting focus from regarding a single test score to interpreting change in test scores over time 🡪 Auch bei prädiktiver Validität.

* **What is LPA**
  + (Förster & Souvignier, 2014)
    - „learning progress assessment (LPA) is one prominent tool in the field of formative assessment which can serve teachers as well as students to optimize learning and instruction ([Black and Wiliam, 2009](https://www.sciencedirect.com/science/article/pii/S0959475214000218?via%3Dihub" \l "bib5), [Clark, 2012](https://www.sciencedirect.com/science/article/pii/S0959475214000218?via%3Dihub" \l "bib10)). Reviews of the literature on effects of LPA show that this approach has a high potential to foster student learning (e.g. [Stecker, Fuchs, & Fuchs, 2005](https://www.sciencedirect.com/science/article/pii/S0959475214000218?via%3Dihub" \l "bib65))“
    - „research has primarily focused on low-achieving students and it usually was applied to single children of a classroom“
  + (Förster & Souvignier, 2015)
    - No real distinction between LPA and CBM, vaguely both as forms of progress monitoring, or LPA is improved CBM in general instead of special education?
  + (Zeuch, Förster, & Souvignier, 2017):
  + (Förster et al., 2018):
  + CBM LPA Formative Assessment
    - (Förster & Souvignier, 2014): CBM is one form of LPA
    - (Förster & Souvignier, 2015): no real distinction, maybe between the lines: LPA is improved CBM (as ORF fails to give differentiated information) transferred to general education
* **Reading**
  + (Förster & Souvignier, 2014)
    - “theoretical models of reading competence ([Förster and Souvignier, 2011](https://www.sciencedirect.com/science/article/pii/S0959475214000218?via%3Dihub" \l "bib18), [Mullis et al., 2003](https://www.sciencedirect.com/science/article/pii/S0959475214000218?via%3Dihub" \l "bib50))”
    - Some people: ORF not good indicator reading fluency, as you need “not only identify words but concurrently need to construct their meaning to comprehend text. He (Samuels, 2007) emphasized this simultaneity of decoding and comprehension” to be most important for reading fluency 🡪 good technical reader not necessarily has high comprehension level (Lerkkanen, 2014) 🡪 Samuels: “new test-concepts are needed that require the reader to simultaneously decode and comprehend text” 🡪 “Hierarchical models of text comprehension ([Kintsch, 1998](https://www.sciencedirect.com/science/article/pii/S0959475214000218?via%3Dihub" \l "bib37)) and reading competence models ([Mullis, Martin, Gonzalez, & Kennedy, 2003](https://www.sciencedirect.com/science/article/pii/S0959475214000218?via%3Dihub#bib50)) may provide the theoretical basis for a new test concept”
      * ORF: Worte werden nur vorgelesen. Kann auch Wort, das ich nicht verstehen, einfach vorlesen. Bei uns: Auch schon auf Wort-Level: Entscheiden, ob es das Wort gibt 🡪 comprehension
  + (Förster & Souvignier, 2011)
    - “reading competence is very important for the acquisition of knowledge and skills in academic settings (Daneman, 1991; Mullis, Martin, Gonzalez, & Kennedy, 2003)”
    - “The most common strategy to monitor reading progress is the measure-ment of reading ability by indicators that strongly correlate to standardized reading tests, namely reading fluency and maze tasks (Fuchs, 2004).” “But no information on specific needs of a poor reader”
    - “Since collection of data does not promote student achievement until instructional modifications on the basis of diagnostic information are made (Stecker et al., 2005), tests should not only identify poor achievers but also give information that can be used in the instructional decision-making process” 🡪 Provide feedback to teachers (not only he is good, he isn’t)
    - “Thus, a key challenge in assessing (reading) progress is the identification of measurement tasks that (a) simultaneously integrate the various skills required for competent year-end performance, (b) are sensitive to student growth, and (c) are designed to be given on a frequent and repeated basis in school.”
      * “One approach, termed curriculum sampling, involves systematic sampling of the skills constituting the annual curriculum in such a way that each test represents the curriculum equivalently. This method is often used in math (Foegen, Jiban, & Deno, 2007), whereby each test includes the same sub skills in the same proportion; for instance, addition, subtrac-tion, multiplication, and division problems”
        + Fuchs 2004 zu curriculum sampling: “It offers the added benefit [(as compared to robust indicators)] of informing instruction by providing descriptions of individual skill mastery, because each skill in the annual curriculum is systematically assessed on every weekly test (eg Fuchs, Fuchs, Hamlett & Allinder, 1991; Fuchs, Fuchs, Hamlet & Stecker, 1990)”
      * “For the second approach, termed robust indicators, measures are identified that represent broadly defined proficiency and cor-relate robustly with the various component skills that constitute the academic domain. In the area of reading, oral reading fluency and the maze task have been identified as valid indicators for reading progress (Reschly et al., 2009; Wayman et al., 2007)”
        + ORF: correlated with standardized tests of reading achievement, but results for relation to reading comprehension are mixed. Also: Feasibility (only 1-to-1 possible)
        + Maze: More feasible, good criterion validity, sensitive to changes over time
* Both valid indicators of reading achievement, but both do not provide “differentiated diagnostic information about single components of reading achievement that can be used to adapt instruction”
  + - Comprehension most important goal. Text comprehension: mental representation of information in the text (Kintsch, 1998; van Dijk & Kintsch, 1983): hierarchical models of text comprehension: textbase and situation model
      * Textbase: semantic relations between text elements, local representation; microstructure; lower-level reading processes
      * Situation model: Connections between text content and prior knowledge; macrostructure; higher-level reading processes
      * Textbase contatining propositions explicitly taken from the text (automated for good readers) is basis for situation model (where meaning needs to be actively constructed)
* For ORF/maze no situation model is needed, only textbase from lower-level reading processes
* For reading achievement, both are needed, situation model becoming especially important when reading experience/performance increases (Danemann, 1991)
* Provide teachers with information about both textbase and situation model to fully depict reading achievement
  + Kintsch, 1998 (Quelle noch einfügen)
    - “As long as one tests for information that is directly given in the text, one is measuring textbases. […] They may involve some semantic knowledge, as in generalizing a concept, but that still remains at the textbase level”, “As soon as one tests for things not directly in the text, on is testing for aspects of the situation model. Inference questions and sorting tasks are obvious examples” (p. 293)
  + Tagungsbeitrag Förster 2017
    - “Items were designed to assess the efficiency of component processes of reading comprehension on word, sentence, and text level”
  + (Schneider, 2010)(Sitzung 11 Päda II)
    - Wenn bei Wortidentifikation Probleme auftreten, hat das negative Konsequenzen für den gesamten Leseprozess 🡪 Reading comprehension on word level to intercept potential problems here
    - Stand in der Quelle nicht auch, dass erstmal hierarchieniedrige Prozesse wichtiger beim Lesenlernen?!
    - Man geht heute allgemein von zwei Komponenten der Lesekompetenz aus
      * Erste Komponente: Basale Leseprozess wie Rekodieren (Lautieren) und Dekodieren (Sinnentnahme des Gelesenen/Lautierten) bilden sich zunächst aus
      * Zweite Komponente: Leseverständnis, wird mit zunehmendem Alter bedeutsamer. Aussagen eines Textes aktiv mit ihrem Vor-, Welt- und Sprachwissen verknüpfen
* Das mit Kintsch-Modell verbinden
  + Further possible information about reading as a construct: (LaBerge & Samuels, 1974; National Reading Panel (U.S.), 2000)
  + Is Reading proficiency/ability the same as reading comprehension or is reading comprehension part of it?
  + Robust indicators like ORF as well as our comprehension measure used to judge reading proficiency?
  + Reading made up of fluency and comprehension?
  + Oral Reading Fluency as an Indicator of Reading Competence: A Theoretical,Empirical, and Historical Analysis (Quelle noch hinzufügen)
    - “Second, gains or performance slopes can track the development of reading competence within an individual. These strategies for characterizing reading competence and improvement have been shown to be more sensitive to inter- and intraindividual differences than those offered by other well-accepted, more broadly conceptualized reading tasks (e.g., Marston, Fuchs, & Deno, 1985). For example, as Frederiksen (1981) demonstrated, the number of word reading errors in context does not as a rule distinguish groups of high- and low-ability readers as well as the chronometric aspect of processing, as reflected in oral reading rate, which consistently provides a basis for distinguishing levels of reading expertise” 🡪 Argument for using slope
  + Müller Richter 2014:
* **Studien Validität und Slope** ToDo: Welche Maße genau (hoffentlich robust indicators)
* Bei allen wichtig

1. How was slope calculated?
2. What measures were used (for progress monitoring and as outcome measures)
3. How were outcome measures predicted from slope (eg one model, separate models, controlling for initial performance or not, …)?
   * (L. S. Fuchs, Fuchs, & Compton, 2004)
     + CBM word identification slopes correlated with outcome measures for predictive validity
   * (Cho, Capin, Roberts, & Vaughn, 2018)
     + “The purpose of this study was to better understand the value of using slope data to monitor student progress by examining the predictive validity of oral reading fluency (ORF) on reading comprehension of struggling fifth-grade readers.”
     + “Does ORF growth predict reading outcomes for struggling fifth-grade readers beyond the information provided by initial performance?” 🡪 nur struggling readers
     + Nützlicher Abschnitt zu “Predictive Utility of ORF slope” “We identified only six studies that have examined the predictive utility of ORF slope on reading comprehension performance ([Baker et al., 2008](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6233895/#R2); [Kim, Petscher, Schatschneider, & Foorman, 2010](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6233895/#R30); [Schatschneider et al., 2008](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6233895/#R44); [Tolar et al., 2014](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6233895/#R59); [Wanzek et al., 2010](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6233895/#R67); [Yeo, Fearrington, & Christ, 2012](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6233895/#R69)). Four studies evaluated the predictive utility of ORF slope for reading outcomes among students in grades 1–3.”
     + Slopes: individual growth curve model of random linear slope and random intercept with ORF scores measured at three time points
   * (Christ, Zopluoglu, Long, & Monaghen, 2012)
     + Simulation study, no validity
   * (Karay & Schauber, 2018) **SEM**
     + Medical students
     + Combined latent growth models for intercept and slope in preclinical and clinical phase (with interrelated slopes and intercepts) and correlations of slopes and intercepts with performance on national licensing exam
     + Ganz gute Infos zu LGM
   * (Keller-Margulis, McQuillin, Castañeda, Ochs, & Jones, 2017) **SEM**
     + Latent classes of readers based on growth on some test. Least-performing class failed end of year state test
     + Growth could be promising alternative to static measures
   * Keller-Margulis, Shapiro, Hintze (2008) (Quelle noch einfügen)
     + Titel: Long Term Diagnostic Accuracy of Curriculum-BasedMeasures in Reading and Mathematics
     + “What is the relation between growth rates from CBM data and large-scale assessments 1 and 2 years later?”
     + Slope and 1 year later assessment no correlation, slope and 2 years later assessment correlation
     + Slope modelled with OLS
   * (Oslund et al., 2014)
     + Kindergarten children at risk of developing reading difficulties
     + Using static scores to predict reading outcomes at the end of first and second school year
   * (Petscher, Cummings, Biancarosa, & Fien, 2013)
     + ORF growth may be curvilinear and linear growth a flawed practice (quite some studies)
   * (Shapiro, Dennis, & Fu, 2015) **Two models**
     + Computer adaptive test and cbm for math
     + Two level hierarchical linear models (Bayesian) yielded individual math slopes. These then fed as independent variables into regression to predict some math outcome
     + Slopes of CBM nearly no prediction incremental to static score, slope of computer test only a bit
     + Unser Modell nicht zweischrittig, sondern slope und Zusammenhang mit outcome in einem modell?
   * (Tolar, Barth, Fletcher, Francis, & Vaughn, 2014) **SEM wie pr. Val. Bei uns!**
     + “Unless the progress monitoring measure is highly aligned with outcome, slope may be an inefficient method for evaluating progress in a Response-To-Intervention context.”
     + 6-8 grade students
     + “Group level studies of progress monitoring slope have typically focused on psychometric properties of slope or sensitivity of slope for measuring group level differences (Ardoin, Christ, Morena, Cormier, & Klingbeil, 2013).” “None of these studies evaluated the relation between slope and reading outcomes using a model-based approach.” Three studies that did that ((Kim et al., 2010; Schatschneider et al.,2008; Wanzek et al., 2010)), but they did not control for beginning of year reading ability and growth was estimated from one model, then the estimates were entered into separate regression models
     + Predictive validity may be higher when outcome measure is aligned with progress monitoring measure, i.e. both text comprehension and not one text comprehension and one word list 🡪 Maybe for concurrent validity, word scales are not as aligned as sentence and text scales (discussion)?
     + Measures: Progress Monitoring Measures: Reading comprehension subtest (Maze CBM), passage fluency (ORF measure), word fluency (ORF measure); Outcome measures: Passage Comprehension Test and Word Reading Efficacy Test
     + LGM over progress monitoring time points, then predicting post-outcome-test with pre-outcome test and intercept and slope of LGM. So like our predictive model, only that outcome pre and post here are not latent variables! All combinations of progress monitoring and outcome measures tested, setting intercept either as first or last progress monitoring testing point
     + Results: For two of three progress monitoring measures, slope did generally not predict either of the two outcome tests (when controlling for initial/final progress monitoring status (intercept)), only three exceptions of which in one slope is negative predictor. Only case where slope is robust predictor: word fluency slope predicts word reading efficacy test (intercept at initial time point)
   * (Yeo, Fearrington, & Christ, 2011) (2012) **SEM**
     + CBM Oral Reading and Maze Reading
       - Meta-Analysis zu Validity of these two measures wrt performance on outcome test: (Shin & McMaster, 2019)
     + “When compared to traditional achievement tests, one notable feature of CBM is that it allows teachers to document the academic growth of students over time. CBM measures have numerous alternate forms and are designed to be administered on a weekly or biweekly basis, making it possible to calculate growth rates for individual students (Deno, 2003).”
     + So-called response to intervention used to identify learning disabilities: slope data is used to determine students (so valid slopes are important!)
     + “A recent review on CBM reading measures by Wayman et al. (2007) indicated a need for research on the validity of CBM slopes.” (‼) 🡪 We contribute to this
     + “First, much of the prior research examined reliability of slopes and not the validity of slopes. Second, the few studies that examined the validity of slopes typically used estimates of level, rather than change, as the criterion variable” 🡪 Argument dafür, dass es nicht nur ausreicht, slope auf ein statisches Outcome zu beziehen (warum nicht?)
     + “Estimates of slope should be evaluated not with static scores but with other estimates of slope” “estimates of level and trend are not interchangeable (Wayman, 2007)”
     + Section “Applications of CBM Slope Estimates”: Punkte dafür, was man mit dem slope machen kann
     + “However, one limitation of the Baker et al. study was that they investigated validity of slope with the criterion of high-stakes reading tests, which provide estimates of student achievement at a single time point. Such estimates lack status as criterion variables because estimates of level and trend are not interchangeable (Wayman et al., 2007). Estimates of slope should be evaluated not with static scores but with other estimates of slope”
     + Result: “Results indicate that CBM-R growth estimates are not correlated with CBM-mR growth estimates as derived with bivariate latent growth modeling (BLGM). In addition, results indicate that CBM growth estimates do not contribute to predictions of student performance on statewide assessments.”
     + Section “Data Analysis”: Quellen für LGM und fit index Daumenregeln
   * Wanzek et al., 2010 (Quelle fehlt noch) s. Tolar 2014 **Two models**
   * Schatschneider et al., 2008 (Quelle fehlt noch) s. Tolar 2014 **Two models**
   * Kim, Petscher, Schatschneider, & Foorman, 2010 (Quelle fehlt noch) s. Tolar 2014 **Two models**
   * (Baker et al., 2008) noch lessen **SEM, similar to ours for pr. Val.**
     + ORF
     + Intro: ORF widely used as screening (maybe argument for our measure: ORF good for screening, but our better for facilitating students over time in more depth?)
     + And maybe: Response to Intervention: ORF used to select students for Special Education(?) 🡪 Not sophisticated enough, a measure like ours would be better?
     + Intro: Good description of how ORF is actually measured
     + Intro: Hint at model for reading development (LaBerge and Samuels, 1974)
     + ORF strongly related to comprehension (more than decoding) 🡪 Many sources: ORF valid measure of overall reading proficiency
     + We need to look at intercept and slope, not slope alone, as often in academic tests, slope and intercept are related. Discussion: “To accurately interpret slopes, it is helpful to … control for for the starting point”
     + “In the current study, we were interested in the contribution of slope, controlling for initial level of performance” 🡪 Contribution of slope to predicting outcome (incremental to intercept)
     + “The third objective was to test how well various models that included ORF and performance on specific high-stakes reading tests in Year 1 predicted performance on specific high-stakes reading tests in Year 2. In particular, we were interested in testing how well ORF stood up in prediction models that included a comprehensive measure of reading in the model. We expected that even under pre-diction models that included a comprehensive measure of reading, ORF would still provide important information in the prediction, consistent with the findings of Wood (2006).”
     + SEM model like ours for predictive validity: slope and intercept over 2 years of progress monitoring (classes 1 and 2 (?)), then slope, intercept and outcome test after class one predict same outcome test after class 2. But: outcome tests as indicators, not as latent variables
     + Results:
       - Static scores: Correlation of static scores and outcome measure: Between .58 to .82
       - Slope in addition to intercept: “ORF intercept and slope predicted a statistically significant portion of performance on the Grade 2 SAT-10 (p < .0001). Together, ORF level and ORF slope explained 70% of the variance on the SAT-10 high-stakes reading test at the end of Grade 2. In addressing Research Question 2, ORF slope accounted for an additional 10% of the variance on the Grade 2 SAT-10, after controlling for initial level of performance.”
       - Pre outcome measure and intercept and slope: “ORF intercept and slope and the high-stakes reading measure in Year 1. In both models, ORF slope accounted for a statistically significant amount of the variance in predicting the high-stakes measure. In first and second grade, the contribution of slope was greater than in second and third grade.”
     + Not really progress monitoring, only three measurement points over school year 🡪 More measurement points and slope is predictive 🡪 Use slope data to guide instructional decisions regarding individual students (with our measure) to correct wrong developments
     + First study to relate growth to important outcome measures (high-stakes reading tests)
   * (Wayman, Wallace, Wiley, Tichá, & Espin, 2007) **Review**
     + „ If teachers were to use the measures to make instructional decisions, the measures would have to have demonstrated reliability and validity “
     + Review Marston, 1989: word identification and reading aloud valid with respect to criterion: “Criterion-related validity coefficients with published measures of reading ranged from .63 to .90, with most above .80.Criterion-related validity coefficients with basal reading series criterion mastery tests ranged from .57 to .86, with half above .80” (14 studies) 🡪 Mostly static scores?!
     + This paper: Review update of Marston, 1989
     + “Establishing the validity of a measure is an ongoing and recursive process (Messick,1989b)” 🡪 Wsl der Messick, den Natalie mit “Das was man interpretiert muss valide sein” meinte
     + “Messick (1989a, 1989b) described construct validity as a multifaceted, but unified, concept that takes into account the evidential and consequential factors related to test interpretation and test use.”
     + Research related to growth: Mostly reading-aloud, only a little of the other two measures, maze selection and word identification, that were examined here
       - Little research wrt validity of slope (defined as prediction of performance on external outcome measures, so criterion-related validity)
       - Shin, Deno, Espin, 2000: Hierarchical Linear Models too estimate growth on maze selection measures. Growth related to performance on outcome test. (As two separate models?!)
       - Tichà, 2009 and Espin 2007: Hierarchical Linear Models to estimate growth for maze selection measures, was related to outcome test
       - Speece and Ritchey 2005: growth-curve analysis for reading aloud, predicted end of year performance
       - Highlighting need for more research with respect to slope validity
   * Yeo (2010) **Metaanalysis**
     + “Another feature of using CBM is that such data are sensitive to the effects of program change, thereby allowing teachers to evaluate the effectiveness of prereferral interventions (Deno, 2003) and providing them with sufficient instructional information for improving students’ academic performance in reading.”
     + “The use of CBM in reading has become increasingly important with the current focus on school accountability and high-stakes assessments in research and practice (Shaprio, Keller, Lutz, Santoro, & Hintze, 2006)” 🡪 Arguing, that our measure would be better suited for decisions like this?
     + “With the national emphasis on adequate yearly progress, many schools and districts have created CBM benchmarks to identify students who are at risk of failing statewide achievement tests (Silberglitt & Hintze, 2005; Wood, 2006). Yet, the effectiveness of using CBM reading measures as alternative indicators of school success is dependent on build-ing predictive validity evidence with statewide achievement tests (Wood, 2006).”
     + Incorporating studies using reading aloud (majority) and maze and statewide achievement tests as outcome: What is predictive validity of CBM for the outcome tests?
     + 27 studies included
     + Overall effect size: cbm measures correlated with outcome test performance (all studies only reported correlations): .689
     + Moderator analysis, as significant variance between studies

* **Measures**
  + CISRT: Reading comprehension on word, sentence and text level (word comprehension, sentence comprehension, text comprehension)
  + Reliability: Retest and Split-Half
* Studie Francis 2008
  + ORF: read text for typically one minute, ORF is WCPM
  + ORF passage forms can vary although high retest reliability and apparent equivalance because of readability formulas
  + Link between fluency and comprehension: fluent readers are better at comprehension. Therefore ORF as proxy for comprehension
  + FIVE ESSENTIAL CHARACTERISTICS FOR PROGRESS MONITORING ASSESSMENTS
    - CBM proposed to have these properties
  + Challenge in progress monitoring (CBM or not): large number of tests needed which have consistent difficulty levels as we want to assess the rate of progress made in a skill 🡪 parallel tests needed, otherwise we cannot know someone’s rate of progress. This has often been viewed as a reliability problem, but it is unrelated to reliability. Rather, scores need to be scaled in a way that differences in text difficulty across passages are factored out of the reported score distribution
  + Eg DIBELS test relies on stimulus equating (readability) to establish equivalence
  + Readability indices
    - Readability eg as measured by the Spache index and on other indices vary remarkably (Good, Kaminski, 2002b)
    - Only modest, if any, correlations with WCPM scores, so not good for evaluating equivalence

**Neuer Ansatz 180220**

The ability to read provides a foundation to achieve any further knowledge and skills and represents a main concern in primary schools. Explicitly monitoring students’ progress in reading has been found to be useful to support student learning. Providing objective information about reading achievement, teachers can use progress monitoring to adapt instructional programs to students’ needs (Förster & Souvignier, 2014; Stecker, Fuchs & Fuchs, 2005). 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. We propose a newly developed instrument within the framework of LPA. This instrument offers differentiated diagnostic information on students’ reading comprehension on word, sentence and text level, thereby maximizing progress monitoring’s benefits. Naturally, to be used in this way, the instrument must meet psychometric standards. While the validity of any instrument is critical, the validity of progress assessments poses a somewhat special case. Here, we need not only to examine the validity of a single test score at a single point in time. Rather, we must shift our attention to a change in scores over time (i.e. slope data) (L. S. Fuchs, 2004). This shift is deemed important as slope data is especially used to inform instructional decisions. The current study aims to draw a comprehensive picture of our newly developed LPA instrument’s validity.

Introduction

The ability to read provides a foundation to achieve any further knowledge and skills, thus representing a main concern in primary schools. Explicitly monitoring students’ progress has been found to be useful to support students in learning how to read. Repeatedly providing diagnostic information about reading achievement, teachers can use progress monitoring to adapt instructional programs to students’ needs, resulting in higher growth for students (Förster & Souvignier, 2014; Shapiro et al., 2015; Stecker et al., 2005). Therefore, supplying teachers with tools helping them to monitor student progress can improve instruction and students’ learning outcomes. The most widely used progress assessments in reading are oral reading fluency (ORF) and maze tasks (Deno, 2003; Reschly, Busch, Betts, Deno, & Long, 2009). Located within the framework of Curriculum-Based Measurement (CBM) (Deno, 1985a; L. S. Fuchs, Fuchs, & Hamlett, 2007), these assessments rely on the assumption that test forms are equivalent. Only then, it is possible to validly infer progress from test scores (Francis et al., 2008). Test equivalence here is usually established using readability indices and verified employing retest reliability. However, research indicates that neither readability indices nor high retest reliability guarantee equivalent test forms (Ardoin, Suldo, Witt, Aldrich, & McDonald, 2005; Francis et al., 2008). Additionally, while these assessments seem to represent a valid indicator for reading comprehension, they fail to provide insight into specific processes readers are struggling with and consequently lack differentiated diagnostic information (Förster & Souvignier, 2011). Surmounting these shortcomings, we consider a new web-based reading progress assessment developed using rule-based item design. In rule-based item design, we can purposefully construct equivalent test forms by equating item difficulty: Item properties influencing item difficulty are systematically combined to ensure identical levels of difficulty between test forms and thus test equivalence already during the construction stage. What is more, our instrument assesses the efficiency of component processes of reading comprehension on word, sentence and text level, thereby offering differentiated diagnostic information as a basis for instructional decisions. Like any other assessment, our instrument must meet psychometric standards. While the validity of any assessment is critical, validating progress assessments requires special considerations. Here, our focus of interest is not mainly a single test score at a specific point in time. Rather, we are interested in change in scores over time (i.e. slope data) that is used to inform instructional decisions (L. S. Fuchs et al., 2004; L. S. Fuchs, 2004). Consequently, we need to ensure that this change in scores over time is a valid indicator of change in reading comprehension. This study aims to draw a comprehensive picture of our newly developed instrument’s validity, highlighting special considerations required for the validation of progress assessments along the way.

Theoretical Background

Students are a heterogenous group, resulting in individually different requirements and challenges in learning how to read. This heterogeneity especially impacts the level of early reading ability (Suchan et al., 2007). Consequently, individual learning outcomes in reading can easily differ. Data about student achievement helps teachers to nonetheless facilitate each student individually. Using this data to inform instructional decisions, teachers can improve student performance (Borman et al., 2016; Carlson et al., 2017; Slavin et al., 2013). Especially collecting student achievement data repeatedly, i.e. monitoring learning progress, provides valuable diagnostic information (Förster & Souvignier, 2014). In this way, teachers do not only obtain information about a student’s current achievement level at multiple points in time. Rather, they can also consider change in level over time, i.e. learning progress. This information is valuable, as students do not only differ in base level, but also in progress of competencies (Salaschek, Zeuch, & Souvignier, 2014). Based on learning progress, teachers can repeatedly evaluate how well students respond to their instructional program and whether it should be adapted in order to improve student performance (Deno, 2003; L. S. Fuchs, 2004; L. S. Fuchs, Fuchs, Hamlett, & Stecker, 1991; Johnson et al., 2010; Parker, Vannest, Davis, & Clemens, 2012; Shapiro et al., 2005; Stecker et al., 2005). Francis et al. (2008) highlight five fundamental features necessary for successful learning progress assessments. Firstly, the assessments must be applied in regular intervals to create a sound data basis. Secondly, due to their high frequency, carry-out of the assessments must be brief and easy. Thirdly, assessments must be standardized, so supply scores on a consistent metric in order to capture progress reliably and validly. Fourthly, both performance level of single assessments and rate of progress throughout assessments must be prognostically valid, i.e. predict end-of-term performance. Only then, instructional decisions can be based upon assessment performance. Lastly, assessments must be equivalent, that is free from practice or form effects, to yield unbiased estimations of progress. A framework for assessing learning progress with respect to reading ability proposedly possessing these features was built by Deno (1985a) with the formulation of CBM. Originating in special education, CBM repeatedly assesses student performance with brief, easy to administer tests sampled from curriculum materials at short, eg weekly, intervals of time (Deno, 2003). The standardized tests enable teachers to reliably and validly monitor learning progress (L. S. Fuchs & Deno, 1991). Coming in equivalent forms, the tests probe desired end-of-term performance to depict students’ progress on the curriculum (L. S. Fuchs, 2004). Thus, slope can be used to quantify learning progress. Yeo (2010) found this learning progress to be associated with criterion measures with a correlation of *r*=.69 in a meta-analysis of 27 studies. Overall, using CBM for progress monitoring 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 *d*=.7 (L. S. Fuchs & Fuchs, 1986). Also Stecker et al. (2005) conclude in their review that CBM is associated with higher student performance.

Constructing CBM tasks in a way to retain this positive effect, there are mainly two approaches (L. S. Fuchs, 2004). In one approach, termed curriculum-sampling, each CBM test represents the curriculum equivalently. This approach is often used in mathematics, where each test contains the same problem types representing specific curricular competencies in the same proportion. For basic arithmetic, this could for example be addition and subtraction tasks (for a review of CBM in mathematics, see Foegen, Jiban, and Deno(2007). The second approach, termed robust indicators, is mostly used in reading. Here, a task type correlating with the competencies rooted inside the curriculum is measured. Two of the most popular task types in this domain are ORF and maze tasks (Deno, 2003; Reschly et al., 2009). In assessing ORF, a student reads aloud from a text passage, usually for a short time interval like one minute, while the teacher notes the amount of words read correctly (Deno, 1985a). In maze tasks, single words are deleted from a text passage which students must fill in based on eg three response options during a time interval of two to three minutes (L. S. Fuchs & Fuchs, 1992). In order to yield unbiased estimations of learning progress, text passages for these tasks should be selected in a way that they are of equivalent difficulty (Ardoin et al., 2005). Mostly, passage equivalence is established using readability indices and verified by means of retest reliability (Francis et al., 2008). Readability indices quantify the difficulty of a text passage based on specific properties of the text so that passages with a similar readability can be selected as equivalent test forms. Researchers employ numerous different readability indices (for a brief description of eight of them, see (Ardoin et al., 2005). They include properties like sentence length (Fry-index; (Fry, 1989)), the amount of words per sentence (Flesh-Kincaid-index; (FLESCH, 1948)) or the total number of words containing more than three syllables (FOG-index; (Gunning, 1971)). Usually, several of these properties are used within a formula specifically weighting them to compute a passage’s readability. However, research indicates that readability indices cannot guarantee passages of equivalent difficulty. Depending on which readability index is used, rank ordering passages according to their readability produces varying results (Good & Kaminski, 2002), hinting away from consistently establishing equivalent passage difficulty via readability indices. Ardoin et al. (2005) identified five readability indices commonly used in CBM, of which four did not exhibit correlations with ORF scores, leading them to discourage the use of readability indices to estimate passage difficulty. In replication, Ardoin, Williams, Christ, Klubnik, and Wellborn (2010) could not predict ORF scores using three readability indices, casting doubt regarding their utility in establishing passage difficulty. Likewise, Begeny and Greene (2014) conclude from studying the appropriateness of readability to gauge passage difficulty, that difficulty should be quantified with means other than readability indices. Also in evaluating passage equivalence, there are reservations regarding the common procedure of employing retest reliability. Francis et al. (2008) argue that researchers misleadingly interpret a high retest reliability as an indicator of passage equivalence although equivalence is a problem unrelated to reliability. In summary, there is evidence discouraging the use of readability indices and retest reliability to infer passage equivalence for CBM test forms. Nonetheless, equivalent test forms remain a central requirement to correctly estimate learning progress (Ardoin et al., 2005; Francis et al., 2008). In search of an alternate approach to constructing equivalent CBM test forms, rule-based item design emerges as a promising candidate. In rule-based item design, item properties affecting item difficulty are systematically varied to purposefully manipulate item difficulty. To link item properties to item difficulty, the cognitive processes needed to solve an item are made explicit, thus integrating cognitive psychology into psychometry (Irvine, Kyllonen 2002). Assuming that, working on an item, cognitive processes are employed, success depends on the processes specific for an item. Thus, items vary in difficulty depending on the cognitive processes necessary for their solution. Explicitly knowing these cognitive processes and what item properties they are linked to establishes a clear connection between item properties and item difficulty, allowing to purposefully manipulate item difficulty (Holling, 2009). In this way, it is possible to generate items without needing to calibrate them (Embretson, 1999). This has, among other things, been successfully used to generate algebra word problem items (for a review of rule-based item design in algebra word problems, see Holling, 2009). As an example, Enright and Sheehan (2002) systematized properties of algebra word problems affecting item difficulty. They made out three item properties: mathematical complexity, context and “algebraicness”. Mathematical complexity comprises attributes like the number of mathematical operations or the level of parentheses used in an item. Context refers to an item’s content basis, like covering the units time or distance. “Algebraicness” describes how many variables are needed in order to solve an item. The authors could show that for example items with a higher level of “algebraicness” were significantly more difficult to solve. Also in developing figural matrix items and word problems in probability theory, rule-based item design has been successfully employed to predict item difficulty from item properties (Freund,…,Holling,2008; Holling,Zeuch,2009). In addition to purposefully influencing item difficulty, rule-based item design offers further benefits. Knowing the cognitive processes linked to an items’ properties, it is possible to design items assessing specific competencies. In addition, a test score will not only be an ability estimation but include feedback about the cognitive processes a subject is struggling with (Holling, 2009). Summarizing, rule-based item design provides two major advantages. It enables test constructors to firstly purposefully generate items of varying difficulty while secondly assessing specific cognitive processes including feedback about their mastery. With these properties, rule-based item design seems to be particularly beneficial for constructing learning progress assessments. To this end, we can use rule-based item design to inexpensively construct an arbitrary number of equivalent test forms equating item difficulty, without the need to calibrate the items psychometrically. What is more, differentiated feedback about the mastery of the cognitive processes necessary to complete the tests can be obtained, extending the data basis about learning progress for instructional decisions. However, combining two of the most prevalent CBM tasks in reading, ORF and maze, with the approach of rule-based item design, two types of complications come to life. Practically, ORF tasks contain whole text passages. In effect, it is difficult to make out single items, as language segments like sentences or words are interconnected. Thus, items made out on any semantic level are nested inside the text passage, complicating rule-based item generation. Also in terms of maze tasks, the gaps to be filled in are nested inside a text passage. Consequently, it is unclear how to generate equivalent text passages by operating on item level. Conceptually, ORF and maze don’t directly measure the competencies we would like to assess, as reading comprehension. Instead, tasks types robustly correlating with these competencies are measured. As a result, ORF and maze provide an informative indicator of reading comprehension, but don’t directly address any component processes constituting reading comprehension (Förster, Souvignier, 2011). However, exactly this is one major advantage of rule-based item design. Using rule-based item design, we are not only able to assess specific competencies, but also to receive differentiated feedback about the mastery of these competencies. Combining rule-based item design with CBM could thereby drive the assessment of learning progress in reading towards a curriculum-sampling approach, measuring explicitly defined curricular competencies while at the same time providing insight into processes students are still struggling with. Taken together, test formats other than CBM or maze could best exploit the advantages of rule-based item design for assessing learning progress in reading. Practically, we need to operate on item level. Conceptually, refining the assessment of reading comprehension towards a curriculum-sampling approach, a cognitive perspective offers a basis to identify specific competencies comprising reading comprehension. From a cognitive perspective, reading comprehension can be conceptualized as an efficient mastery of cognitive subprocesses on the word, sentence, and text level (Richter 2009). Efficiency here refers to the fact that an efficient result is not only correct, but also obtained with rather low effort. (Richter, 2012). Accordingly, an efficient subprocess works with both high accuracy and low cognitive load. Important subprocesses AM ANFANG MUSS STEHEN, DASS ES UM VALIDITÄT DES SLOPES GEHT (Notizen zu Paper Tobias passen dazu, warum slope relevant)! So was wie: Validität ist immer wichtig, bei der Validität von Progress Assessments muss man aber besondere Facetten beachten. Struktur: Validität des Slopes. Wichtig weil. Hier mit folgendem Test untersucht. Dann kurz erklären, warum Test konstruiert. To use rule-based item design, we further need to operate on item level. Dann Richter. Dann unser Test (und hier web- and computer-based). Dann Validität. a large item pool can be inexpensively generated without the need constructeUsing Equating item difficulty -concise good description of rule-based item design and knackpunkt: “cognitive components” (darüber Brücke zu refining reading construct)-

* Richter 2012
  + Lesefähigkeiten = effiziente Bewältigung von kognitiven Teilprozessen des Lesens auf Wort-, Satz- und Textebene (Richter 2009, Vellutino 2004)
  + Oft wird auch Reaktionszeit erfasst 🡪 Indikator für Effizienz (ein effizienter Prozess löst das Item nicht nur korrekt, sondern auch mit geringer kognitiver Beanspruchung)
  + Erster standardisierter Test mit kognitionspsychologischer Perspektive, der nicht produktorientiert arbeitet, sondern prozessorientiert, um Rückschlüsse auf die Effizienz von Teilprozessen des Leseverstehens ziehen zu können
  + Zentrale Begriffe
    - Lesefähigkeit: Ganz global die Fähigkeit, Wörter, Sätze und Texte zu verstehen
    - Leseverstehen: Prozess des Leseverstehens, der sich aus kognitionspsychologischer Sicht in Teilprozesse untergliedern lässt
    - Leseverständnis: Produkt des Leseverstehens: Qualität und Umfang mentaler Repräsentationen, die beim Lesen von Wörtern/Sätzen/Texten konstruiert werden
  + Kognitionspsychologische Perspektive: Leseverständnis beruht auf der effizienten (zuverlässigen, also korrekten und zugleich ressourcenschonenden) Bewältigung von kognitiven Teilprozessen auf Wort-, Satz- und Textebene
  + Teilprozesse Wortebene
    - Visuelle Worterkennung: Eintrag in mentalem Lexikon einem geschriebenen Wort zuordnen
    - Zentrale für Leseverständnis (Hypothese lexikalische Qualität und interaktiv-kompensatorisches Modell)
    - Wenn nicht ausreichend effizient, Bindung kognitiver Ressourcen, die dann bei hierarchiehöheren Prozessen fehlen (Theorie verbale Effizienz und Anwendungsbeispiel Naumann 2008)
    - Visuelle Worterkennung hat drei Teilprozesse (Dual Route Cascaded Model)
      * Phonologische Rekodierungsprozesse: Einer Graphemfolge phonologische Repräsentation zuordnen (bes. wichtig in frühem Stadium des Lesens und bei seltenen Wörtern)
      * Orthographische Vergleichsprozesse: direkter Vergleich der Schreibweise eines Wortes mit seiner Repräsentation im mentalen Lexikon (orthographische Strategie); besonders für häufige Wörter und solche mit irregulärer Schreibweise
      * Aktivierung von Wortbedeutungen als Basis für das Verständnis zusammenhängender Wörter (Sätze, Texte). Defizite in der Effizienz dieses Prozesses können auch bei guten orth. Vergleichsprozessen zu Leseschwierigkeiten führen
    - Also Worterkennung = phonologische Rekodierung oder orthographische Vergleichsprozesse plus Aktivierung von Wortbedeutungen
  + Teilprozesse Satzebene
    - Syntaktische (Satzbauanalyse) und semantische (Integration von Wortbedeutungen in Satzsinn) Prozesse, unklar, ob diese modular oder interagierend ablaufen
    - Teilweise wohl auch Informationen aus dem Satzkontext zur Worterkennung genutzt 🡪 Wort- und Satzebene keine getrennten Verarbeitungsstufen, sondern interaktive Prozesse
    - In jedem Fall spielen syntaktische und semantische Prozesse eine Rolle für das Leseverstehen
  + Teilprozesse Textebene
    - Integration von bereits Gelesenen und weiteren, neuen Informationen zu einer Repräsentation des Texts.
    - Kintsch
      * Textbase: semantic relations between text elements, local representation; microstructure; lower-level reading processes
      * Situation model: Connections between text content and prior knowledge; macrostructure; higher-level reading processes
      * Textbase contatining propositions explicitly taken from the text (automated for good readers) is basis for situation model (where meaning needs to be actively constructed)
    - Zuerst lokal also zwischen verschiedenen Sätzen (Textbase) und dann globaler (situation model). Zu Beginn des Leseerwerbs textbase wichtig, die Grundlage für situation model bildet.
    - Es müssen bspw. kausale Beziehungen zwischen Sätzen nachvollzogen oder mithilfe von Vorwissen erschlossen werden; das können Grundschüler unterschiedlich gut (effizient)
* Holling 2009
  + Explizite Verbindung zwischen kognitiven Prozessen, die nötig sind, um ein Item zu lösen und systematischem Testdesign wichtig, um valide zu messen (Borsboom, …, 2004)
  + Rule-based item design integrates cognitive psychology and psychometric theory, increasingly common (Irvine und Kyllonen, 2002)
  + Rule-based item design
    - Cognitive processes for solving an item are made explicit. Thus construct validity of an item can easily be tested and items assessing specific sub-abilities can be designed
    - Generation of large item pools that don’t necessarily need to be calibrated (Embretson, 1991)
    - The cognitive processes needed for solving a problem affect how difficult an item is. If we have item properties significantly affecting item difficulty, these properties must thus affect the cognitive processes determining item difficulty 🡪 Via item properties affecting the cognitive processes, item difficulty can be influenced
    - And Items of differing difficulty are necessary to estimate a subject’s ability
  + Enright and Sheehan 2002: Three item properties influencing item difficulty for algebra word problems
  + Aside from the item properties on the task side, there may be influences on item difficulty on the person side: Person properties may also play an important role
  + Authors’ pilot study
    - Identify item properties affecting difficulty of statistical word problems. These properties then have to be implemented as rules for constructing items 🡪 The levels of the factors (properties) are systematically varied over items. Eg context a vs b, proability vs absolute numbers, irrelevant information yes vs no, …
    - Design matrix of items can be specified
    - Da ich weiß, welche Itemeigenschaften welche kognitiven Prozesse, die zur Lösung wichtig sind, beeinflussen, erlaubt die Testbearbeitung Rückschlüsse darauf, mit welchen kognitiven Komponenten ein Teilnehmer noch Probleme hat 🡪 Feedback über individuelle Schwierigkeiten mit kognitiven Komponenten der Items
    - Ergebnis: statistical word problems can be generated based on predefined difficulty factors allowing to easily generate items with a wide range of difficulties
  + Durch systematische Variation der item properties kann ich systematisch die difficulty beeinflussen (aber Achtung: person side properties beeinflussen auch die difficulty; aber das Problem habe ich bei readability ja auch) und somit äquivalente Testformen herstellen
* Embretson 1999
  + With the right item design principles, it is possible to produce items with specific difficulties on the fly
  + Usually, items are calibrated to meaningfully assess ability. To go without calibration when generating items on the fly, all stimulus properties must be specified for each item to the predict the psychometric properties from the item
  + Cognitive design system approach (Embretson, 1998): Establish which cognitive components play a role in solving an item. These cognitive processes are linked to certain item properties. Depending on the cognitive processes, and thus on their associated item properties, items’ difficulty is affected 🡪 design principles predict item difficulty and contains information about cognitive processes necessary for correctly solving an item
  + Eigene Verständnisbeschreibung des Cognitive design system approach: Depending on an item’s properties, different cognitive components are necessary to a different extent to correctly solve an item, making items vary in difficulty. Knowing which cognitive components play a role in the solution process, we can draw inferences about an item’s difficulty. Knowing which item properties are linked to which cognitive components, we can predict item difficulty based on the item properties an item has got
  + Four advantages of cognitive design system approach

Rule-based item design als Alternative beschreiben

Das ist aber schwierig mit ORF und maze. Außerdem: Brücke zu component processes schlagen: Wie in Präsentation: Looking for other readings tasks. This also allows to refine the reading construct measured by these tests. Away from robust indicators, more curriculum-sampling like, which provide informative approximation of reading comprehension but lack in-depth information.

Theoretisches Problem Reading Aloud: can read aloud fluently without comprehension? S. Wayman 2007 S. 105

La Berge D, Samuels SJ. Toward a theory of automatic information processing in reading. Cognitive Psychology 1974;6:293–323. (in Francis 2008 gefunden)

CBM enables teachers to reliably and validly monitor learning progress in a standardized fashion (L. S. Fuchs & Deno, 1991). Originating in special education, CBM repeatedly assesses student performance with brief tests of constant task difficulty at short intervals of time (Deno, 2003). The equivalent tests probe desired end-of-term performance to depict students’ progress on the curriculum (L. S. Fuchs, 2004). Thus, slope can quantify learning progress.

Robust indicators und sampling the skills themselves

Anforderungen nach Francis

In practice, ORF and maze

* Special education 🡪 general education (Deno, 2003: in general nicht so helpful wie in tailored special; am Ende)
* Robust indicators vs curriculum sampling
* Förster 2014 S.2 oben rechts: growth rate of reading comprehension different across students (Förster, Souvignier 2011), so progress data useful
* Deno, 1985: “CBM improves the data base for making educational decisions for several reasons. all of which relate to the fact that measures of student achievement may be obtained frequently (daily if desired). Since CBM requires only that the teacher regularly obtain a brief sample of curriculum-based performance, we have seen that repeated measurements can be conducted to produce data that may be graphically displayed. When a series of datapoints is graphed as in Figure 5, not only is it possible to de-scribe the student's current level of performance in thecurriculum.it is also possible to reliably estimate the rate of a student's increase in performance by computing the slope of the data (ef. White&Haring,1980). The accessibility of slope data through CBM is one of the unique contributions of this approach to achievement measurement since it yields a statistic describing student growth that is unavailable through using conventional achievement tests. As will be illustrated later. the data on slope makes it possible to move instruction into the arena of applied science.”
* 2014 Souvignier Förster Salaschek aus literature\_notes bzgl. Schwierigkeit und parallele Tests und später wg. Lehrerurteilen

This monitoring has proven to be effective, as in a meta-analysis of 21 studies, student performance was higher when using CBM, with an average effect size of *d*=.7 (L. S. Fuchs & Fuchs, 1986). Also Stecker et al. (2005) conclude in their review that CBM is associated with higher student performance.

*Assessing Learning Progress*

- Motivation for progress monitoring

- Progress monitoring offers data about student achievement that can be used to inform instructional decisions

- CBM as one form of progress monitoring (equivalent tests)

- Widely used measures, even for selecting students for special education/screening

- What these measures lack

- equivalence

- facettes of reading (hier Richter Prodi-l einbauen)

- LPA (or not???)

- Our instrument (with rule-based item design)

- Fluency measures like ORF good proxy for comprehension because close link between Fluency and Comprehension (Francis, 2008, Einleitung)

- Test equivalence here is usually evaluated using retest reliability. However, Francis et al. (2008) argue that equivalence is unrelated to reliability. Rather, test forms need to show identical levels of difficulty. Unser Ansatz: Purposefully construct equivalent test forms über gleiche Itemschwieriegkeiten

Students are a heterogenous group, resulting in individually different requirements and challenges in learning how to read. This heterogeneity especially impacts the level of early reading ability (Suchan et al., 2007). Consequently, individual learning progress in reading can easily differ. Data about student achievement helps teachers to nonetheless facilitate each student individually. Using this data to inform instructional decisions, teachers can improve student performance (Borman et al., 2016; Carlson et al., 2017; Slavin et al., 2013). Especially collecting student achievement data repeatedly, i.e. monitoring learning progress, provides valuable diagnostic information (Förster & Souvignier, 2014). In this way, teachers do not only obtain information about a student’s current achievement level at multiple points in time. Rather, they can also consider change in level over time, i.e. learning progress. Fuchs et al. (1991), noch hinzufügen: Focus on growth important 🡪 Why is it benefical to use learning progress (in Abgrenzung zu “data about student achievement” von oben)?. Evaluating and adapting instructional programs based on data about learning progress has been found to improve student learning.

Based on learning progress, teachers can repeatedly evaluate how well students respond to their instructional program and whether it should be adapted in order to improve student performance (Deno, 2003; L. S. Fuchs et al., 1991; L. S. Fuchs, 2004; Johnson et al., 2010; Parker et al., 2012; Shapiro et al., 2005; Stecker et al., 2005). A framework for assessing learning progress with respect to reading ability was built by Deno (1985a) with the formulation of curriculum-based measurement (CBM).

[concise description of CBM]

- originally developed for special education

- curriculum

- quantify rate of learning

- scores plotted as learning curve (so static scores and growth)

- learning curve (slope) is of primary concern

- norms

🡪 Was ist der Vorteil davon, slope (zusätzlich) zu benutzen?

Fuchs, Fuchs, Compton 2004:

"CBM iscollected frequently (weekly or monthly); studentscores are graphed; a slope is derived from thegraphed scores to quantify reading improvement;and the teacher applies decision rules to the slopeto formulate instructional decisions."

"The methods by whichCBM informs reading instruction rely on thegraphed performance indicator. Decisions arebased on a graph displaying time on the horizon-tal axis and performance (number of words readcorrectly from text in 1 min) on the vertical axis.If a student's growth trajectory is judged to be ad-equate, the teacher increases the student's goal foryear-end performance; if not, the teacher revisesthe instructional program. Research shows thatthese decision rules produce more varied instruc-tional programs, which are more responsive to in-dividual needs (Fuchs et al., 1989b), with moreambitious student goals (Fuchs, Fuchs, & Ham-lett, 1989a) and stronger end-of-year scores oncommercial reading tests (e.g., Fuchs, Deno, &Mirkin, 1984; Wesson, 1991)." --> Wouldn't it be even better if we used more differentiated feedback?

Paper Tobias:

"Ziel der Lernverlaufsdiagnostik ist weniger die Klassifikation, sondern eher das Abbilden der Lernentwicklung und dasdaran orientierte formative Anpassen pädagogischer Aktivitäten (Souvignier, Förster & Zeuch, 2016)."

"Der Vorteil dieses Ansatzes (curriculum sampling) ist, dass mathematische bzw. arithmetische Kompetenzen sehr curriculumsnah erfasst werden können, wodurch zugleich eine kriteriumsorientierte Messung ermöglicht wird (Klauer & Strathmann, 2013)."

"Um die individuellen Lernverläufe einzelner Schüler abzubilden, werden meist lineare Trends auf Schülerebene geschätzt. Dazu wird auf lineare Regressionsmodelle oder gemischte lineare Modelle zurückgegriffen (Snijders & Bosker, 2012). Bei der Schätzung individueller linearer Trends sind die beiden Parameter Regressionsinterzept (Intercept) sowie Regressionssteigung (Slope) von Interesse. Der Slope stellt dabei den mittleren Lernzuwachs einer Person dar, in den Intercept gehen sowohl Informationen zum individuellen Lernausgangsniveau („Startpunkt“) als auch zum mittleren Leistungsniveau („Durchschnitt“) ein."

"Einige Studien untersuchten die prädiktive Validität individueller Slopes und berichten heterogene Ergebnisse. Die moisten Studien wurden zudem im Zusammenhang mit Lesefertigkeiten durchgeführt. So fanden Stage und Jacobsen (2001) eine Korrelation von r = .26 zwischen dem Slope bei Oral Reading Fluency (Lernverlauf Lesen) und dem Textverständnis. Schatschneider, Wagner und Crawford (2008) berichten hingegen, dass aus individuellen Lernverläufen im Lesen geschätzte Slopes keine inkrementelle Validität für die Vorhersage des Leseverständnisses am Ende des Schuljahres beitragen (s. auch Yeo, Fearrington & Christ, 2012)."

[design of CBM-tasks: curriculum sampling vs robust indicators]

[ORF and maze as widely used robust indicators in reading]

[no differentiated diagnostic information, so even more helpful with that]

[In line with Förster (2011), we argue that more differentiated diagnostic information creates a more profound foundation to adapt instructional program upon]

[Because our test measures … based on … model of reading and therefore better for adapting instruction]

[ORF: Worte werden nur vorgelesen. Kann auch Wort, das ich nicht verstehen, einfach vorlesen. Bei uns: Auch schon auf Wort-Level: Entscheiden, ob es das Wort gibt 🡪 comprehension]

Read (Aarnoutse, van Leeuwe, Voeten, & Oud, 2001) 🡪 Why do I have this source? Sth about reading construct?

Read (Deno, 2003), (Shapiro et al., 2015) and (Tindal, 2013) to then concisely describe CBM (important: slope data is used!)

Natalie: “die wesentliche Frage ist, inwiefern die differenzierte Erfassung hilfreich für instruktionale Entscheidungen ist”

„Was ein Modell betrifft, haben wir uns bislang immer an Kintsch orientiert oder an den Arbeiten aus der Arbeitsgruppe von Tobias Richter“

* Validity
  + (Guion, 1980):
    - More than one score
    - Content construct and criterion as three aspects (but not independent!): “Technical Recommendations for Psychological Tests and Diagnostic Techniques (American Psychological Association [APA] et al., 1954)”
    - (American Psychological Association, American Education Research Association, & National Council on Measurements Used in Education, 1954)
    - Criterion validity: Two categories: “(a) to investigate the meaning of scores as measures of a certain attribute and (b) to investigate the scores as concomitants or predictors of other attributes (APA et al, 1974)” 🡪 Does our test measure what we intended to measure? / more practical interest of predicting sth meaningful
  + validity trias may not be regarded as three distinct "type" of validity. They are facets of a unitary construct, we can for example never rely on criterion related validity indicators alone but have to base it on construct validity arguments (Messick, 1989a, 1989b) 🡪 We need to regard multiple facets of validity to integrate them into a unified concept of validity
  + Ursprünglich nach APA trias, was mistankely regarded as three distinct components by researchers (Guion, 1980 and Messick 1989 “Meaning …”). Better: Unitary perspective adopted by APA (American Psychological Association, American Educational Research Association, & National Council of Measurement Used in Education, 1985). Validity as unitary construct integrating all evidence helpful in interpreting test scores. According to Messick (1989 “Meaning …”) we can subsume all this evidence under the label of construct validity, where a test represents an indicator for the construct we attempt to measure. Consequently, we can regard content and criterion-related validity aspects as part of construct validity as they contribute to our understanding/interpretation of test scores. 🡪 We try to offer different validity indicators to accumulate evidence for different facets of construct validity, thereby also considering features unique to progress assessments
  + Nützlich: Section “Comprehensiveness of Construct Validity” (Messick, 1989, “Meaning …”), p.7
  + (Messick, 1995)
    - "The traditional concept of validity divides it into three separate types; content, criterion, and construct validities. This view is fragmented and incomplete, failing to take into account evidence of the value implications of score meaning as a basis for action and of the social consequences of score use"
    - "In particular, what needs to be valid is the meaning or interpretation of the scores as well as any implications for action that this meaning entails(Cronbach, 1971)"
    - “Indeed, broadly speaking, validity is nothing less than an evaluative summary of both the evidence for and the actual as well as potential consequences of score interpretation and use (i.e., construct validity conceived comprehensively). This comprehensive view of validity integrates considerations of content, criteria, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and utility. Fundamentally, then, score validation is empirical evaluation of the meaning and consequences of measurement. As such, validation combines scientific inquiry with rational argument to justify (or nullify) score interpretation and use.”
    - Fundamental feature of construct validity: construct representation: “decompose the task into requisite component processes and assembling them into a functional model or process theory” (based on Embretson, 1983)
    - Threats: construct underrepresentation (zu schmale Operationalisierung) and construct-irrelevant variance (zu breite Operationalisierung)
    - “construct-irrelevant difficulty for individuals and groups is a major source of bias in test scoring and interpretation as well as of unfairness in test use.” 🡪 Bei unseren Tests kein Vorwissen nötig, das construct-irrelevant difficulty und somit construct-irrelevant variance wäre?
    - “In its simplest terms, construct validity is the evidential basis for score interpretation. As an integration of evidence for score meaning, it applies to any score interpretation -- not just those involving so-called "theoretical constructs."”
    - “Historically, primary emphasis in construct validation has been placed on internal and external test structures – that is, on the appraisal of theoretically expected patterns of relationships among item scores or between test scores and other measures.”
    - “empirical relationships between predictor scores and criterion measures should make theoretical sense in terms of what the predictor test is interpreted to measure and what the criterion is presumed to embody (Gulliksen, 1950).”
    - “An important form of validity evidence still remaining bears on the social consequences of test interpretation and use.” 🡪 nicht nur intended consequences als strands im nomological network, sondern auch unintended consequences 🡪 Das haben wir nicht
    - „Thus, the process of construct validation evolves from these multiple sources of evidence a mosaic of convergent and discriminant findings supportive of score meaning.“
    - “In sum, the construct validity of score interpretation comes to undergird all score-based inferences not just those related to interpretive meaningfulness but including the content- and criterion-related inferences specific to applied decisions and actions based on test scores.”
    - “validity becomes a unified concept and the unifying force is the meaningfulness or trustworthy interpretability of the test scores and their action implications, namely, construct validity”
    - Six aspects of construct validity making up unified validity approach
      * Content aspects
        + Content validity (boundaries and structure of construct)
      * Substantive aspects
        + Cover domain processes in addition to domain content and scientific evidence beyond expert judgment that participants really engage into these processes when doing test tasks
        + “empirical evidence of response consistencies or performance regularities reflective of domain processes (Loevinger, 1957).”
        + “not just to the content representativeness of the construct measure but also to the process representation of the construct and the degree to which these processes are reflected in construct measurement.” 🡪 content and substantive aspects together = representativeness of the construct
      * Structural aspects
        + “scoring models should be rationally consistent with what is known about the structural relations inherent in behavioral manifestations of the construct in question (Loevinger, 1957; Peak, 1953)”
        + “the internal structure of the assessment (i.e., interrelations among the scored aspects of task and subtask performance) should be consistent with what is known about the internal structure of the construct domain (Messick, 1989)” 🡪 substantive aspects: in addition to content, domain processes should be represented in the test. Structural aspects: Scoring should be reflect what we know about how the structure/interplay of the domain processes 🡪 content, substantive and structural aspects together: abgedeckt, Konstruktinhalt und Konstruktprozesse abzubilden, und das Scoring gemäß der Struktur des Konstrukts zu machen 🡪 Wir nehmen an: Lesefähigkeit unterteilt in Wort-, Satz- und Textebene (related but distinct factors in reading ability) so scores for each level, these grow linearly when measured repeatedly in equivalent form (quantify rate of learning) and from the growth we are able to infer reading development so that we can base instructional adaptations on the growth score (Haben allerdings nicht getestet, dass growth linear, höchstens indirect dadurch dass das Modell gut passt, aber haben kein konkurrierendes Modell das z.B. quadratischen growth annimmt)
      * Generalizability aspects
        + Test set of indicators of construct. Generalizability: Meaning/Interpretation of score applies to concept more broadly, not only specifically in the are of the assessed tasks 🡪 generalizability: boundaries of score meaning 🡪 Erfasst unser Test Lesefähigkeit generell?
        + Not only generalizability across tasks, but limits of score interpretation also affected by generalizability across time, occasions, raters, …
      * External aspects
        + In how far do relationships of the score with other measures (or the lack of relation) support the score interpretation
        + “Discriminant evidence is particularly critical for discounting plausible rival alternatives to the focal construct interpretation. Both convergent and discriminant evidence are basic to construct validation.”
        + “Of special importance among these external relationships are those between the assessment scores and criterion measures pertinent to selection, placement, licensure, program evaluation, or other accountability purposes in applied settings. Once again, the construct theory points to the relevance of potential relationships between the assessment scores and criterion measures, and empirical evidence of such links attests to the utility of the scores for the applied purpose.”
      * Consequential aspects
        + Evaluating consequences of score interpretation (positive consequences should be maximal, negative ones minimal) 🡪 benefits of instructional adaptation based on test scores? No/minimal disadvantages because of it?
    - “validation is empirical evaluation of the meaning and consequences of measurement”
    - Welche Aspekte wir haben
      * Content
        + Anything?
      * Substantive
        + Somehow CISRT? But what evidence could there be?
      * Structural
        + CFA
      * Generalizability
        + Correlations with ELFE (also growth), correlations with teacher ratings as different task?
      * External
        + Correlations with ELFE (also growth), CFT and DEMAT; Criterion-related validity (growth) (also predict ELFE post with static scores per time point?)
      * Consequential
        + Nothing

Fuchs 2004 als Basis. Regarding validity, two of three research stages mentioned important

* Studien zu Stage 1 aus Fuchs
* Auch in neueren Studien und dann die aus conference paper
* Studien Stage 2
* Studien Stage 1
  + Ardoin et al., 2004
  + Clarke & Shinn, 2004
  + McGlinchey & Hixson, 2004 ("assessing CBM correlations and decision utility with respect to a highstakes reading test.")
  + Fore, Boon, Burke, & Martin, 2008 (from conference paper)
  + Valladolid, 2015 (from conference paper)
  + Vanderheyden et al., 2016 (from conference paper)
* Studien Stage 2
  + Hintze and Christ (2004)
  + Fuchs, Fuchs, Hamlett, Walz & Germann, 1993
  + Tolar and colleagues (2014) (from conference paper)
  + (Yeo et al., 2011)
* Literatursuche zu beiden Stages
  + Progress monitoring | cbm | lpa und validity und vielleicht slope
    - (Oslund et al., 2014) (“This study examined the changing role and longitudinal predictive validity of curriculum-embedded progress-monitoring measures (CEMs )”)
    - Christ, Zopluoglu, Long, & Monaghen, 2012

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 et al., 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 et al., 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). 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). CBM was originally introduced to monitor poorly performing students by frequently measuring their skill level with brief tests. This monitoring has proven to be effective, as in a meta-analysis of 21 studies, student performance was higher when using CBM, with an average effect size of .70 (L. S. Fuchs & Fuchs, 1986). As another, yet closely related approach to CBM (Förster & Souvignier, 2014), learning progress assessment (LPA) aims to monitor the progress of all students in a class. Implementing this in an efficient manner, every three weeks, a web-based test is administered to each student. These equivalent tests measure reading ability as established by a curriculum (defining designated end of period performance), thus allowing to model learning progress over testing points.

Sure enough, this advantage of using CBM comes at a cost, i.e. the time necessary to realize the repeated LPAs (L. S. 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).

Obviously, this is only true if tests correctly capture learning progress. In other words, like any other test, a test measuring learning progress must meet psychometric quality criteria. Regarding validity as a central quality criterion, repeated tests aiming to capture learning progress pose a somewhat special case. In addition to verifying that a test score reflects the current level of the targeted competence, we need to investigate in how far a change in test scores over time reflects the change in the targeted competence over time. Only then can we confidently use the information about learning progress offered by such a series of repeated tests. The current study aims to draw a comprehensive picture of the validity of a newly developed test series measuring learning progress, thereby contributing to the so far scarce research about this topic.

Traditionally, when examining a test’s validity, test scores are interrelated with other test scores or criteria to determine in how far a test score represents the quality intended to measure. In the case of learning progress, however, verifying that single test scores indeed measure the strength of some ability is not enough. Rather, we would like to verify that the change in test scores over time indeed represents a change in ability.

**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 (L. S. Fuchs et al., 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, 1985b) conceptualized the approach of CBM, many progress assessments have been developed and validated, thereby mainly focusing on the aspect of status validity (Cecil Fore et al., 2008; Furey et al., 2016; 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.

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, 2014; Stecker et al., 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.

Here, we need not only to examine the validity of a single test score at a single point in time. Rather, we must additionally verify that an increase in test scores indeed goes along with an increase in the targeted competence (L. S. Fuchs, 2004). This second step is deemed important as data about changes in scores (i.e. slope data) is especially used to inform instructional decisions. The current study aims to draw a comprehensive picture of the validity of a newly developed LPA series targeting reading ability.

References

Aarnoutse, C., van Leeuwe, J., Voeten, M., & Oud, H. (2001). Development of decoding, reading comprehension, vocabulary and spelling during the elementary school years. *Reading and Writing*, *14*(1), 61–89. https://doi.org/10.1023/A:1008128417862

American Psychological Association, American Education Research Association, & National Council on Measurements Used in Education (1954). Technical Recommendations for Psychological Tests and Diagnostic Techniques. *Psychological Bulletin*, *51*(2), 201–238.

American Psychological Association, American Educational Research Association, & National Council of Measurement Used in Education (1985). *Standards for educational and psychological testing*. Washington DC: American Psychological Association. Retrieved from http://worldcatlibraries.org/wcpa/oclc/13560935

Ardoin, S. P., Suldo, S. M., Witt, J., Aldrich, S., & McDonald, E. (2005). Accuracy of Readability Estimates' Predictions of CBM Performance. *School Psychology Quarterly*, *20*(1), 1–22. https://doi.org/10.1521/scpq.20.1.1.64193

Ardoin, S. P., Williams, J. C., Christ, T. J., Klubnik, C., & Wellborn, C. (2010). Examining Readability Estimates' Predictions of Students' Oral Reading Rate: Spache, Lexile, and Forcast. *School Psychology Review*, *39*(2), 277–285. Retrieved from https://www.researchgate.net/publication/233861173\_Examining\_Readability\_Estimates'\_Predictions\_of\_Students'\_Oral\_Reading\_Rate\_Spache\_Lexile\_and\_Forcast

Baker, S. K., Smolkowski, K., Katz, R., Finn, H., Seeley, J. R., Kame'enui, E. J., & Beck, C. T. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review*, *37*(1), 18–37.

Begeny, J. C., & Greene, D. J. (2014). Can readability formulas be used to successfully gauge difficulty of reading materials? *Psychology in the Schools*, *51*(2), 198–215. https://doi.org/10.1002/pits.21740

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2016). Final Reading Outcomes of the National Randomized Field Trial of Success for All. *American Educational Research Journal*, *44*(3), 701–731. https://doi.org/10.3102/0002831207306743

Carlson, D., Borman, G. D., & Robinson, M. (2017). A Multistate District-Level Cluster Randomized Trial of the Impact of Data-Driven Reform on Reading and Mathematics Achievement. *Educational Evaluation and Policy Analysis*, *33*(3), 378–398. https://doi.org/10.3102/0162373711412765

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

Cho, E., Capin, P., Roberts, G., & Vaughn, S. (2018). Examining Predictive Validity of Oral Reading Fluency Slope in Upper Elementary Grades Using Quantile Regression. *Journal of Learning Disabilities*, *51*(6), 565–577. https://doi.org/10.1177/0022219417719887

Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology*, *47*(1), 55–75. https://doi.org/10.1016/j.jsp.2008.09.004

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

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

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

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

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

FLESCH, R. (1948). A new readability yardstick. *The Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Foegen, A., Jiban, C., & Deno, S. (2007). Progress Monitoring Measures in Mathematics. *The Journal of Special Education*, *41*(2), 121–139. https://doi.org/10.1177/00224669070410020101

Förster, N., Kawohl, E., & Souvignier, E. (2018). Short- and long-term effects of assessment-based differentiated reading instruction in general education on reading fluency and reading comprehension. *Learning and Instruction*, *56*, 98–109. https://doi.org/10.1016/j.learninstruc.2018.04.009

Förster, N., & Souvignier, E. (2011). Curriculum-Based Measurement: Developing a Computer-Based Assessment Instrument for Monitoring Student Reading Progress on Multiple Indicators. *Learning Disabilities: A Contemporary Journal*, *9*(2), 65–88.

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

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

Fry, E. B. (1989). Reading Formulas: Maligned but Valid. *Journal of Reading*, *32*(4), 292–297. Retrieved from www.jstor.org/stable/40029925

Fuchs, L. S. (2004). The Past, Present, and Future of Curriculum-Based Measurement Research. *School Psychology Review*, *33*(2), 188–192.

Fuchs, L. S., & Fuchs, D. [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. [D.], & Compton, D. L. (2004). Monitoring Early Reading Development in First Grade: Word Identification Fluency versus Nonsense Word Fluency. *Exceptional Children*, *71*(1), 7–21. https://doi.org/10.1177/001440290407100101

Fuchs, L. S., Fuchs, D. [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

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic Distinctions between Instructionally Relevant Measurement Models. *Exceptional Children*, *57*(6), 488–500. https://doi.org/10.1177/001440299105700603

Fuchs, L. S., & Fuchs, D. [Douglas] (1992). Identifying a Measure for Monitoring Student Reading Progress. *School Psychology Review*, *21*(1), 45–58. https://doi.org/10.1080/02796015.1992.12085594

Fuchs, L. S., Fuchs, D. [Douglas], Hamlett, C. L., & Stecker, P. M. (1991). Effects of Curriculum-Based Measurement and Consultation on Teacher Planning and Student Achievement in Mathematics Operations. *American Educational Research Journal*, *28*(3), 617–641. https://doi.org/10.2307/1163151

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

Good, R. H., & Kaminski, R. A. (2002). *DIBELS Oral Reading Fluency Passages for First through Third Grades (Technical Report No. 10)*. Eugene, OR: University of Oregon.

Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology*, *11*(3), 385–398. https://doi.org/10.1037/0735-7028.11.3.385

Gunning, R. (1971). *The technique of clear writing* (Rev ed.). New York, N.Y.: McGraw-Hill.

Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the Accuracy of a Direct Route Screening Process. *Assessment for Effective Intervention*, *35*(3), 131–140. https://doi.org/10.1177/1534508409348375

Karay, Y., & Schauber, S. K. (2018). A validity argument for progress testing: Examining the relation between growth trajectories obtained by progress tests and national licensing examinations using a latent growth curve approach. *Medical Teacher*, *40*(11), 1123–1129. https://doi.org/10.1080/0142159X.2018.1472370

Keller-Margulis, M., McQuillin, S. D., Castañeda, J. J., Ochs, S., & Jones, J. H. (2017). Identifying Students at Risk: An Examination of Computer-Adaptive Measures and Latent Class Growth Analysis. *Journal of Applied School Psychology*, *34*(1), 18–35. https://doi.org/10.1080/15377903.2017.1328627

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*(2), 293–323. https://doi.org/10.1016/0010-0285(74)90015-2

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.

Messick, S. (1989a). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.2307/1175249

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

National Reading Panel (U.S.) (2000). *Report of the National Reading Panel : teaching children to read : an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction : reports*. [Washington D.C.]: National Institute of Child Health and Human Development National Institutes of Health. Retrieved from http://worldcatlibraries.org/wcpa/oclc/47848860

Oslund, E. L., Simmons, D. C., Hagan-Burke, S., Kwok, O.‑M., Simmons, L. E., Taylor, A. B., & Coyne, M. D. (2014). Can Curriculum-Embedded Measures Predict the Later Reading Achievement of Kindergarteners at Risk of Reading Disability? *Learning Disability Quarterly*, *38*(1), 3–14. https://doi.org/10.1177/0731948714524752

Parker, R. I., Vannest, K. J., Davis, J. L., & Clemens, N. H. (2012). Defensible Progress Monitoring Data for Medium- and High-Stakes Decisions. *The Journal of Special Education*, *46*(3), 141–151. https://doi.org/10.1177/0022466910376837

Petscher, Y., Cummings, K. D., Biancarosa, G., & Fien, H. (2013). Advanced (Measurement) Applications of Curriculum-Based Measurement in Reading. *Assessment for Effective Intervention*, *38*(2), 71–75. https://doi.org/10.1177/1534508412461434

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

Salaschek, M., Zeuch, N., & Souvignier, E. (2014). Mathematics growth trajectories in first grade: Cumulative vs. compensatory patterns and the role of number sense. *Learning and Individual Differences*, *35*, 103–112. https://doi.org/10.1016/j.lindif.2014.06.009

Schneider, W. (2010). Lesenlernen. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (4th ed.). Weinheim: Beltz.

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

Shapiro, E. S., Edwards, L., & Zigmond, N. (2005). Progress Monitoring of Mathematics Among Students with Learning Disabilities. *Assessment for Effective Intervention*, *30*(2), 15–32. https://doi.org/10.1177/073724770503000203

Shin, J., & McMaster, K. (2019). Relations between CBM (oral reading and maze) and reading comprehension on state achievement tests: A meta-analysis. *Journal of School Psychology*, *73*, 131–149. https://doi.org/10.1016/j.jsp.2019.03.005

Slavin, R. E., Cheung, A., Holmes, G., Madden, N. A., & Chamberlain, A. (2013). Effects of a Data-Driven District Reform Model on State Assessment Outcomes. *American Educational Research Journal*, *50*(2), 371–396. https://doi.org/10.3102/0002831212466909

Solomon, B. G., & Forsberg, O. J. (2017). Bayesian asymmetric regression as a means to estimate and evaluate oral reading fluency slopes. *School Psychology Quarterly : The Official Journal of the Division of School Psychology, American Psychological Association*, *32*(4), 539–551. https://doi.org/10.1037/spq0000206

Stecker, P. M., Fuchs, L. S., & Fuchs, D. [Douglas] (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.

Tindal, G. (2013). Curriculum-Based Measurement: A Brief History of Nearly Everything from the 1970s to the Present. *ISRN Education*, *2013*(2), 1–29. https://doi.org/10.1155/2013/958530

Tolar, T. D., Barth, A. E., Fletcher, J. M., Francis, D. J., & Vaughn, S. (2014). Predicting reading outcomes with progress monitoring slopes among middle grade students. *Learning and Individual Differences*, *30*, 46–57. https://doi.org/10.1016/j.lindif.2013.11.001

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

Van Norman, E. R., Christ, T. J., & Zopluoglu, C. (2013). The effects of baseline estimation on the reliability, validity, and precision of CBM-R growth estimates. *School Psychology Quarterly : The Official Journal of the Division of School Psychology, American Psychological Association*, *28*(3), 239–255. https://doi.org/10.1037/spq0000023

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

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

Werner Specht (Ed.) (2009). *Nationaler Bildungsbericht Österreich 2009: 2. Fokussierte Analysen Bildungspolitischer Schwerpunktthemen*. Graz: Leykam.

Yeo, S. (2010). Predicting Performance on State Achievement Tests Using Curriculum-Based Measurement in Reading: A Multilevel Meta-Analysis. *Remedial and Special Education*, *31*(6), 412–422. https://doi.org/10.1177/0741932508327463

Yeo, S., Fearrington, J. Y., & Christ, T. J. (2011). Relation Between CBM-R and CBM-mR Slopes. *Assessment for Effective Intervention*, *37*(3), 147–158. https://doi.org/10.1177/1534508411420129

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

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