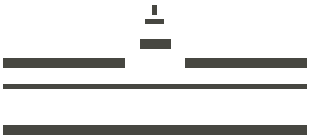


Westfälische
Wilhelms-Universität
Münster

University of Münster

Institute of Psychology

Work Unit Assessment and Evaluation in Schools

Master's Thesis in Psychology

Static Score and Slope

A Comprehensive Validity Analysis of the Quop-L2 Reading Progress Assessment

Mathis Erichsen

Student Number: 418182

mathis.erichsen@uni-muenster.de

August 4, 2020

First Reviewer: Dr. Natalie Förster

Second Reviewer: Prof. Dr. Mitja Back

Contents

List of Figures

List of Tables

1 Abstract

Learning progress assessments (e.g. assessing progress in reading) are used to inform instructional decisions; however, to provide correct information, they must be valid. This study presents a comprehensive validity analysis of the newly developed quop-L2 reading progress assessment. Satisfying special considerations linked to validating learning progress assessments and building on best-practice recommendations, the analyses with data from N = 1989 and a subsample of n = 354 second grade students were threefold: Firstly, the proposed structure (i.e. structural validity) of the quop-L2 assessing the efficiency of component processes of reading comprehension on the word, sentence and text level, thereby offering differentiated diagnostic feedback, was investigated employing Confirmatory Factor Analyses. Secondly, status validity (a static test score from one time point constituting a valid measurement) was examined computing convergent correlations between quop-L2 scores and standardized reading tests as well as teacher judgments. Discriminant correlations were calculated with respect to standardized measures of intelligence and mathematics. Thirdly, slope validity (validly capturing change in test scores over time, i.e. slope, representing learning progress) was scrutinized. To this end, latent quop-L2 slope was related to a) latent pre-post change in standardized reading performance and to b) latent standardized reading end-of term performance in a Structural Equation Modeling Context. Results confirmed quop-L2's structural validity, its status validity, and its slope validity (regarding a) on sentence and text level and regarding b) on all levels). Hence, quop-L2 seems to validly deliver differentiated diagnostic information about reading progress and thus suited to base instructional decisions upon.

2 Introduction

A psychometric test can be a helpful tool to inform a decision. For this, the interpretation of the test score must be valid (Cronbach, 1971). In educational contexts, special test formats, so-called learning progress assessments, are used. Here, beyond evaluating a static score, teachers interpret a change in scores over multiple assessment points, i.e. slope. Based on slope, learning progress is evaluated and instructional decisions are made (L. S. Fuchs et al., 2004). Consequently, slope needs to validly capture learning progress. Ensuring this extends traditional validity analyses to a two-step approach (L. S. Fuchs, 2004). In a first step, one needs to examine whether a static test score obtained at a single point in time validly measures the current ability (status validity). In a second step, one must verify that change in test scores over time is a valid indicator of learning progress (slope validity).

Considering learning progress in reading, two of the most popular test formats to assess reading progress are oral reading fluency (ORF) and maze tasks (Deno, 2003; Reschly et al., 2009). Although these test formats represent valid indicators of overall reading progress, they do not directly address the component processes constituting reading comprehension[1] (Förster & Souvignier, 2011). From a cognitive psychology perspective, being a competent reader means to efficiently execute component processes on the word, sentence, and text level (Richter & Christmann, 2009). Assessing how the efficiency of these component processes develops could provide more differentiated diagnostic information than indicators of overall reading progress like ORF or maze tasks can offer, potentially proving valuable to guiding instructional decisions. Closing this gap, the quop-L2 reading progress assessment (Förster et al., 2017) has been developed to follow this structure, measuring the efficiency of component processes of reading comprehension on the word, sentence, and text level. In this study, I present a comprehensive validity analysis of this newly developed instrument, specifically investigating quop-L2's (1) structural validity, (2) status validity and (3) slope validity. I thus aim to ensure it can be used as intended, i.e. to inform instructional decisions based on differentiated diagnostic information about reading progress. Along the way, I highlight special considerations required for validating learning progress assessments.

[1] In this study, I consider reading ability in terms of comprehension, as I regard comprehension as the overall goal in learning how to read.

3 Theoretical Background

3.1 Assessing Reading Progress

Why is it at all beneficial to assess learning progress? Using learning progress as an information source for instructional decisions is a form of data-based decision making (DBDM). DBDM endorses the idea of basing educational decisions on systematically collected and evaluated data. DBDM can be used at different levels to attain different goals, for example at district level to investigate the effects of policy on schools and students (Levin & Datnow, 2012) or at school level to hold schools accountable for educational outcomes (Carlson et al., 2017). At classroom level, DBDM can improve student learning (Borman et al., 2016; Carlson et al., 2017; Slavin et al., 2013). This is achieved by collecting student data and evaluating it, to then adapt instruction based on the data (Mandinach, 2012). Considering the area of reading, elementary students notably differ in their early reading ability (Suchań et al., 2007). Hence, using data to inform instructional decisions seems to be a promising approach for facilitating each student individually in learning how to read. Following the rationale of DBDM, especially the repeated collection of student achievement data can provide valuable diagnostic information (Förster & Souvignier, 2014). In this way, teachers do not only obtain information about a student's current performance 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 et al., 2014). Based on learning progress, teachers can repeatedly evaluate how well students respond to their instructional program and whether it should be adapted to improve student performance (Deno, 2003; L. S. Fuchs, 2004; L. S. Fuchs et al., 1991; Johnson et al., 2010; Parker et al., 2012; Shapiro et al., 2005; Stecker et al., 2005).

To successfully inform instructional decisions, learning progress assessments should possess five fundamental features (Francis et al., 2008). 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 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 (for example reading comprehension at the end of the school year). 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 comprehension that should possess these features was developed by Deno (1985) with the formulation of Curriculum-based Measurement (CBM). Originating in special education, CBM repeatedly assesses student performance with brief, easy to administer tests sampled from curriculum materials at short, e.g. weekly, intervals of time (Deno, 2003). The standardized tests enable teachers to reliably and validly monitor reading 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 reading progress. Overall, using CBM to assess reading progress 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.

Providing a promising framework to assess reading progress, how can CBM assessments be validated? Considering the requirements posed by Francis et al. (2008), not only single assessment scores but also their rate of progress must predict end-of-term performance. Especially in the light of this progress being used to base instructional decisions upon (L. S. Fuchs et al., 2004; Souvignier et al., 2016; Zeuch et al., 2017), validly measuring reading progress should be a primary area of concern. As Messick (1995b) puts it, validity ultimately addresses the question if the interpretation of a test score is justified. Consequently, adapting instruction based on inferences about reading progress is only reasonable if this progress is captured validly. In line with this, L. S. Fuchs (2004) argues that the validity of CBM assessments comprises two stages. As mentioned above, in a first step - like with any test - researchers must examine status validity. In a second step, they must verify slope validity. If validity can be ensured on both stages, the reading progress captured by the CBM assessments provides valuable diagnostic information to teachers they can base instructional decisions upon. Teachers can use this information to for example identify students not developing as expected and formatively adapt their instructional program to facilitate these students' learning (Souvignier et al., 2016).

3.2 Status Validity

Regarding CBM research in reading, status validity has been widely investigated. In the earliest review I could identify, Marston (1989) reported convergent correlations of static CBM scores with standardized reading tests ranging from r = .57 to r = .86 across 14 studies. Since then, multiple meta-analyses investigating convergent CBM status validity have been conducted. All these meta-analyses focused on the relation between static CBM scores and standardized respectively federal state reading tests[2]. In their meta-analysis, Reschly et al. (2009) considered 41 studies yielding an overall median correlation of r = .68 between CBM scores and the reading tests mentioned above. Consistent results could be found in a meta-analysis published one year later, integrating 27 studies to an overall mean correlation of r = .69 (Yeo, 2010). In the most recent meta-analysis in this domain, Shin and McMaster (2019) identified an overall mean correlation of r = .63 across 61 studies. Another convergent measure researchers have used on some occasions is teacher judgments of reading comprehension, also showing correlations with static CBM scores (e.g. Baker & Good, 1995; Jenkins & Jewell, 1993). As an example, Jenkins and Jewell (1993) reported coefficients of r = .56 and r = .66 for the relation between two different CBM measures and teacher judgments. Baker and Good (1995) even found correlations of r = .80 between static CBM scores and teacher judgments on a 7-point Likert scale ranging from *way below average* to *way above average*. Teacher judgments cannot only be used as a convergent but also as a discriminant measure. In their study, Baker and Good (1995) also gathered discriminant teacher judgments rating bilingual Hispanic students' English language competence on the same scale. These judgments correlated by r = .62 with static CBM scores. It should be noted, however, that the authors did not explicitly exclude reading from language competence when asking the teachers for their ratings. Thus, convergent and discriminant teacher judgments might be highly correlated in this study. A standardized test from the same study assessing English language competence without the subjects needing to read (e.g. by listening to a tape) perhaps constitutes a more meaningful indicator of discriminant status validity here. It correlated by r = .47 with static CBM scores. A few other studies investigated discriminant status validity

[2]When referring to federal state reading tests, I address standardized reading tests that find application and feature reference norms across at least one US federal state. These can be existing standardized tests used in this context or standardized tests developed by a federal state themselves.

with respect to static CBM scores (e.g. Fewster & Macmillan, 2002; Kranzler et al., 1998). Findings include no significant and a low correlation of $r = .24$ for mental speed and intelligence, respectively (Kranzler et al., 1998), as well as correlations between $r = .16$ and $r = .39$ with social studies marks (Fewster & Macmillan, 2002). In summary, there is strong evidence supporting convergent validity of static CBM scores in relation to standardized and federal state reading tests. Also with respect to teacher judgments, at least a few studies indicate convergent validity. Discriminant status validity has not yet been as intensively studied, although single results using different measures point into the direction of its existence. Consequently, convergent status validity can be regarded as well-established. At most, some extending research on convergent teacher judgments or other measures could be used to corroborate it even further. Likewise, more studies applying discriminant measures might be desirable to fully complete our image of CBM status validity.

3.3 Slope Validity

Considering CBM slope validity, research in the field appears less numerous and not as consistent as for status validity. The lower consistency may be due to higher complexity of slope as opposed to static scores, leading to a variety of design choices. One such choice pertains to the way of modeling slope. Common practice in CBM is to model slope via ordinary least squares regression (OLSR; Christ & Desjardins, 2017). As an example, in their validity study, L. S. Fuchs et al. (2004) used OLSR to calculate slopes from CBM scores which were positively correlated with criterion measures of reading comprehension. Similarly, Keller-Margulis et al. (2008) related slopes obtained from OLSR to achievement on federal state reading tests administered two years later. They found that CBM slopes obtained in first and second grades correlated with federal state test achievement collected two years later, but CBM slopes obtained in third grade did not. Results like these, obtained by applying OLSR to calculate slope, have been subjected to criticism. Some researchers judge OLSR slope estimation as imprecise due to a high residual error (Christ, 2006; Hintze & Christ, 2004) and discourage from its use. In addition, OLSR models are rather inflexible, not allowing for an adequate implementation of statistically and conceptually important aspects like random effects modeling or dealing with

non-observable, latent variables. As a result, more statistically advanced approaches to estimating slope overcoming these restrictions might be preferable – such as Hierarchical Linear Modeling (HLM) or Structural Equation Modeling (SEM).

What is more, both L. S. Fuchs et al. (2004) and Keller-Margulis et al. (2008) investigated slope without including an estimate of performance level, i.e. an intercept. As Baker et al. (2008) argue, to accurately interpret the effect of CBM slope on an outcome, an intercept must be included into the model. This is deemed important as the authors note that in reading, slope and intercept are often correlated. The intercept can for example represent initial or final performance level, enabling the researcher to judge slope independent of it. Accordingly, slope validity models should include an intercept to control for performance level when interpreting slope. Using HLM, Shin et al. (2000) modeled CBM slope and its relation to a federal state reading test by including the federal state test as a level-two predictor into their model. Controlling for initial performance level, they found that higher CBM slopes were associated with higher federal state test scores. Likewise, in a small exploratory study, Espin et al. (2010) could establish this relation for one out of two CBM measures under consideration. In a slightly different approach, Kim et al. (2010) calculated initial performance level and slope using HLM to then use the obtained estimates in a separate model to predict performance on a federal state reading test. They found slope for a CBM measure applied in first grade to best predict federal state test scores at the end of first and third grade. On the contrary, this relation did not hold across grade levels, as it could not be reproduced for second and third grade CBM slope. In another study, Wanzek et al. (2010) however reported CBM slopes across grade levels one to three to predict failure on two federal state reading tests administered at the end of third grade beyond initial performance level. Initial performance level and slope were calculated in the same fashion as in the study by Kim and colleagues (2010). Not controlling for initial but for final CBM performance level, Schatschneider et al. (2008) concluded that slope estimates obtained via HLM make little to no contribution beyond final performance level when predicting end-of-term performance on a federal state reading test. Stage and Jacobsen (2001) found HLM slope estimates to have no effect in predicting federal state reading test performance when controlling for performance level either at the beginning, in the middle, or at the end of the school year. Taken together, validity results are also mixed using more advanced modeling techniques like HLM and incorporating

an intercept when modeling slope. In the absence of systematic methodological flaws like small sample sizes that could explain the observed heterogeneity, it remains unclear whether choosing different time points to control for performance level (like initial vs. final performance level) plays a role in the inconsistent results. At any rate, however, one can conclude that evaluating slope should include controlling for performance level.

In addition to controlling for performance level, Baker et al. (2008) incorporated the score on a former federal state test as a further predictor in their study. Estimating intercept and slope in a SEM framework, they seek to investigate the predictive power of slope incremental to the intercept and the same federal state test from the year before on the current year's federal state test performance. A contribution of slope while controlling not only for CBM performance level but also for former performance on the criterion measure could provide a strong indicator of slope containing unique information about reading performance. Indeed, Baker and colleagues (2008) found that slope explained federal state test performance at the end of second and third grade, respectively, beyond initial performance level and performance on the same federal state test from one year before. Likewise controlling for prior performance on the criterion measure, Tolar et al. (2014) showed slope to predict reading performance on a federal state test for typical, but not for struggling readers. Due to the additional control variable used in the analyses, these findings gain in robustness in a field of inconsistent results.

So far, all mentioned studies validated slope with respect to static scores, i.e. estimates of student performance at a single point in time. However, while performance level (i.e. intercept) and change (i.e. slope) may be correlated (Baker et al., 2008), they are conceptually different. Therefore, it makes sense to evaluate slope not only with respect to static scores but also with respect to other estimates of slope (Yeo et al., 2011). In this way, beyond predicting end-of-term performance, it can be ensured that slope indeed represents change in performance over time. I could identify three studies validating slope with respect to change (Speece & Ritchey, 2005; Tichá et al., 2009; Yeo et al., 2011). Tichá et al. (2009) administered the same reading achievement test in the beginning (pre) and at the end (post) of their CBM data collection period. They incorporated the pre-post-difference of reading performance as a level-two predictor when modeling CBM scores using HLM. The authors found the pre-post-change to be related to slope for one of two CBM measures under investigation. In a similar analysis model, Speece and

Ritchey (2005) found slope from first grade calculated via OLSR to predict slope from second grade modelled via HLM. While these two studies show some indications of slope validity, Yeo et al. (2011) did not find corroborating evidence. In their study, the authors found no relation between slopes for two CBM measures obtained in an SEM-approach. In total, the few studies conducted so far evaluating slope with estimates of change yield heterogenous results.

Next to the rather technical, design-related choices mentioned here (slope modeling choice, controlling for performance level, controlling for former criterion performance and validating slope with estimates of change), naturally, there is a variety of other factors potentially impacting slope validity. These could for example be students' grade level or the conceptual alignment of CBM and criterion measure (Cho et al., 2018; Tolar et al., 2014). However, before systematically investigating these factors, it might be useful to agree on certain best-practice standards when designing a slope validity study. Like that, researchers would be better equipped to tackle the inconsistencies uncovered so far. Contributing to this goal, I argue that to obtain a comprehensive study design satisfying the requirements posed by slope validity, at the current state of research, four recommendations can be offered. Firstly, when modeling slope, more statistically advanced approaches than OLSR should be used as OLSR has been criticized for imprecision (Christ, 2006; Hintze & Christ, 2004) and is rather inflexible as explained above. Quite some studies do this, applying HLM or SEM. Secondly, when validating slope, performance level should be modeled as an intercept to disentangle it from slope. Again, quite some studies do this. However, it remains yet unclear how the time point chosen in controlling for performance level (e.g. initial or final) affects results. Thirdly, when relating slope to a static criterion measure, it might be beneficial to include prior performance on the criterion into the model to obtain more robust results. This was so far only done by single studies. Fourthly, being intended to represent progress, slope should not only be evaluated with respect to static scores but also with respect to other estimates of change. Only three studies doing this could be identified. These recommendations will be subsequently implemented. Prospectively, they could maybe also be used in other studies to help shed light on the so far heterogenous results in the domain of slope validity. As a final remark about the current state of research I can conclude in line with Wayman et al. (2007) that above everything else, more research with respect to slope validity is needed.

3.4 Differentiated Diagnostic Information

Debating validity of CBM reading assessments, how do these tests precisely measure reading after all? The two most widely used measures in this area are ORF and maze tasks (Deno, 2003; Reschly et al., 2009). Consequently, most of the validity studies referred to above applied these assessment types. In 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, 1985). In maze tasks, single words are deleted from a text passage which students must fill in based on several response options during a short time interval of two to three minutes (L. S. Fuchs & Fuchs, 1992). ORF and maze tasks are termed robust indicators (L. S. Fuchs, 2004) as these task types robustly correlate with the competence actually intended to be measured, i.e. reading comprehension. As such, ORF and maze tasks provide valid and informative indicators of reading comprehension and thus, given status and slope validity, also of reading progress.

Although being valid indicators of overall reading progress, ORF and maze tasks do not directly address the component processes constituting reading comprehension (Förster & Souvignier, 2011). These component processes could however provide more differentiated diagnostic information than robust indicators of overall reading progress. If wanting to break down reading comprehension into such processes, cognitive psychology offers a useful framework. From a cognitive perspective, reading comprehension can be conceptualized as an efficient mastery of cognitive component processes on the word, sentence, and text level (Richter & Christmann, 2009). *Efficient* here refers to the fact that an efficient result is not only correct but also obtained with rather low effort (Richter et al., 2012). Accordingly, an efficient component process works with both high accuracy and low cognitive load. To tap efficiency, in this process-based view of reading comprehension, additionally to the subject's response itself, the reaction time needed to produce the response is captured. The idea behind this operationalization is that an efficient component process operates faster than a non-efficient one, thus leading to a faster response. Richter and colleagues (2012) presented such a process-oriented assessment for primary school students, ProDi-L (Prozessbezogene Diagnostik des Leseverstehens bei Grundschulkindern [Process-based assessment of reading skills in primary schoolchildren]). ProDi-L assesses the efficiency of component processes of reading comprehension on

the word, sentence, and text level. For example, on word level, a reader must visually recognize a written word to match it to its mental representation. A component process achieving this, amongst others, is orthographic comparison, where a word's spelling is compared to that of its mental representation. Thus, an efficient reader can recognize words both fast and correct. On sentence level, words have to be connected to form a whole sentence. An exemplary component process responsible for this is semantic integration incorporating word meanings into that of a sentence. On text level, meaning must propagate even further, integrating it across sentences to form a representation of the whole text. Being especially important for early readers, the process of local coherence uses semantic relations between text elements to accomplish this goal. For a detailed overview of all component processes on the three levels, see Richter et al. (2012). ProDi-L covers each of these processes in a separate subtest. With this test structure, it seems possible to specifically inform teachers about component processes of reading comprehension students are still struggling with, thereby offering differentiated diagnostic information to base instructional decisions upon.

However, ProDi-L was not developed as a reading progress assessment, therefore not satisfying the requirements posed for learning progress assessments by Francis et al. (2008). For example, test administration takes too long, and there are not enough equivalent test forms available. Utilizing the process-based structure of ProDi-L for measuring reading progress, Förster et al. (2017) adapted the convincing task types offered by Richter et al. (2012) to construct the quop-L2 reading progress assessment. Like ProDi-L, it aims to assess the efficiency of component processes of reading comprehension on the word, sentence, and text level hence offering more differentiated diagnostic information about reading progress than currently used robust indicators like ORF or maze tasks. In doing so, the quop-L2 progress assessment is designed in a CBM-fashion to fulfill the requirements demanded above.

3.5 The Current Study

When developing a reading progress assessment, a validity analysis is necessary to ensure the new instrument's psychometric quality. The quop-L2 reading progress assessment is to my knowledge the first process-based assessment of reading progress.

Therefore, it is an important step to ensure its structural validity, i.e. to verify its proposed three-dimensional structure of conceptualizing reading comprehension as efficient component processes on the word, sentence, and text level. Only then, it can offer the assumed differentiated diagnostic information. Furthermore, many studies are reviewing reading progress assessments with respect to status validity, but fewer studies do so with respect to slope validity. In this area, most studies evaluated slope with respect to a static criterion. Therefore, especially relating slope to other estimates of slope has been scarcely pursued. What is more, most studies have only focused on specific aspects of status and/or slope validity.

Following L. S. Fuchs (2004) two-step approach, this study thus aims to comprehensively validate the quop-L2 reading progress assessment, targeting structural validity as well as both status validity and slope validity in their entirety. In addressing slope validity, the study design adheres to the best-practice recommendations given above. These are namely (1) using an SEM-approach to model slope instead of OLSR while (2) controlling for performance level, (3) controlling for prior performance when relating slope to a static criterion and (4) additionally evaluating slope with an estimate of change. I consider slope to validly measure reading progress when if fulfills two criteria. Firstly, slope needs to indeed measure progress in reading comprehension. To verify this, I relate slope to another estimate of change. Secondly, this progress then needs to predict end-of-term performance as demanded by Francis et al. (2008). To examine this, I use slope as a predictor of end-of-term performance. If both criteria are fulfilled on top of status validity, slope represents performance-relevant progress justifying its use as a basis for instructional decisions.

Taken together, this validity study intends to draw a comprehensive picture of the newly developed instrument's validity, taking into account special considerations required for validating progress assessments. In doing so, it is the first study to use a comprehensive design adhering to best-practice recommendations derived from the current state of research instead of only focusing on single validity aspects. Specifically, in contribution to Wayman and colleagues' (2007) call for more research on slope validity, I address the following three research questions in establishing validity of the quop-L2 reading progress assessment.

1. Structural validity: Does the test reflect the proposed three-dimensional structure of reading comprehension conceptualized as efficient component processes on the word, sentence, and text level?
2. Status validity: Does a single test score at a single point in time validly measure the current level of reading comprehension?
3. Slope validity: Does change in test scores over time validly measure progress in reading comprehension?

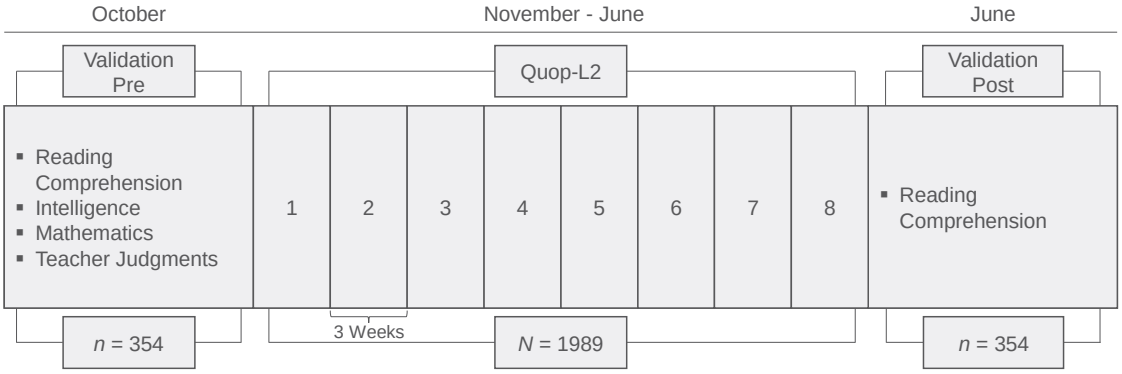
4 Method

The investigation of the research questions was based on a sample of German second graders who were monitored for reading progress across the school year using the quop-L2 reading progress assessment. Additionally, standardized measures of reading, intelligence and mathematics were applied to a subsample at the beginning (pretest) and in parts at the end (posttest) of the school year. To review research question one, Confirmatory Factor Analyses (CFAs) were calculated. For answering research question two, convergent and discriminant correlations between static quop-L2 scores and the standardized measures were computed. Moving on to research question three, latent quop-L2 slope was related to latent pre-post-change in standardized reading performance (change score) and to latent reading performance in the standardized posttest (end-of-term performance; static score) in an SEM framework.

4.1 Participants and Design

The study was based on all N = 1989 (47.76% female, M_{age} = 7.90, SD_{age} = 0.48) German second graders participating in the web-based learning progress assessment system quop (Souvignier et al., 2019) in reading during the school year 2015/16. Students originated from 109 classrooms in 47 schools situated in six federal states. Most schools were located in Hesse (55.32%) and North Rhine-Westphalia (31.92%). The average classroom consisted of 18.24 students, which is a bit lower than the average classroom size in Gemany (21.9 students; Tarelli et al., 2012). Students completed eight equivalent quop-L2 reading progress assessments in intervals of approximately three weeks over the course of the school year.

Figure 1
Overview of the Study Design.



In a subsample of n = 354 (48.59% female, M_{age} = 7.74, SD_{age} = 0.44) students from 18 classrooms in eight schools, additional validation measures were administered. These schools were located around a medium sized town in North Rhine-Westphalia. The subsample size was large enough so that also after dropout and data exclusion, at least 250 students could be analyzed to ensure stability of correlation coefficients (Schönbrodt & Perugini, 2013). Students from the subsample completed standardized measures of reading comprehension (ELFE II; W. Lenhard, Lenhard, et al., 2017), intelligence (CFT 1-R; Weiss & Osterland, 2012) and mathematics (DEMAT 1+; Krajewski et al., 2002) in October prior to the quop-L2 reading progress assessments (pretest). Further, teachers were asked to judge their students' reading skills. Subsequently to the reading progress assessments, reading comprehension was assessed again in June (posttest). For an overview of the design, see Figure 1 (p. 19). I excluded students who had missed all eight reading progress assessments, students who were too old or too young to be in second grade (less than six or more than twelve years at the time of the pretest) and students whose parents did not consent study participation from further analysis. The final samples consisted of N = 1913 students for the total sample and n = 354 students for the subsample, respectively.

Standardized measures at pre- and posttest were administered to whole classrooms in a group setting in paper-pencil format by trained university student assistants. While the students completed the standardized measures, teachers judged their reading skills. Reading progress assessments were provided in the *quop* system developed at the University of Münster (Souvignier et al., 2019).

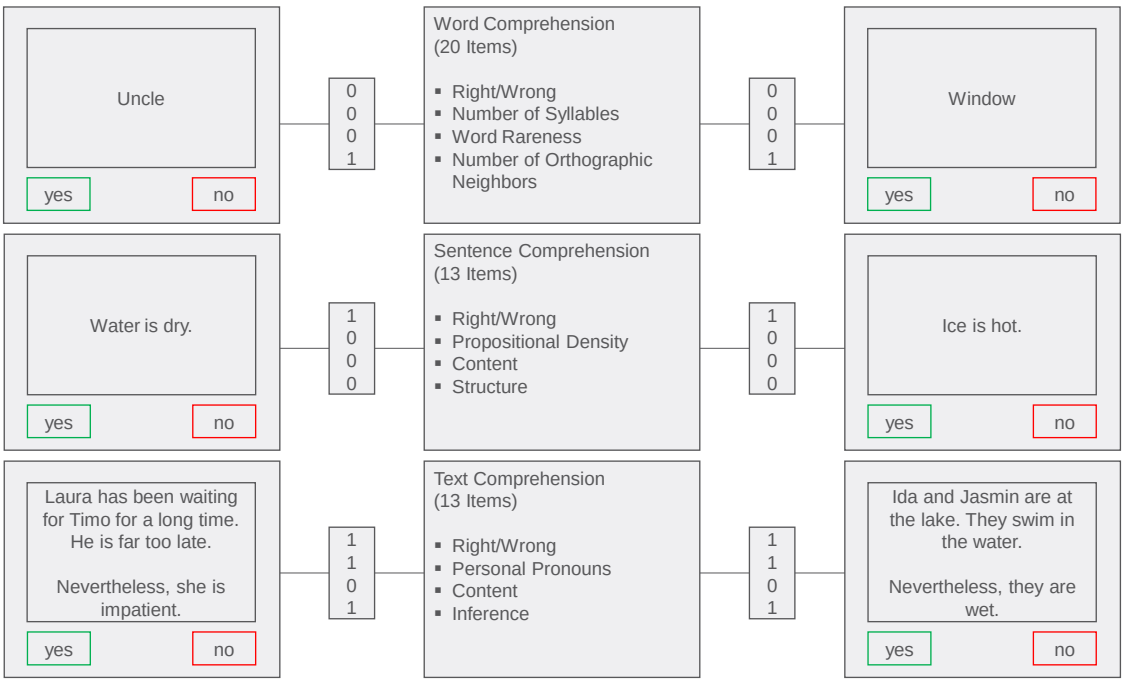
Running in a browser, *quop* is a web-based approach to assess learning progress in reading and mathematics for first to sixth grade. Over the course of a school year, eight equivalent tests are available for intervals of three weeks each. The short tests, taking approximately 10 to 15 minutes, are completed during self-study periods or in group sessions, depending on the number of computers available in the classroom or a computer room. Before the first test, students complete a tutorial to familiarize themselves with the online assessment procedure. After each test, they receive feedback about their performance on the current and former tests. Teachers are equipped with their own login to access test results on student and class level, where progress is depicted in graphs. Norm values are available based on all students that have ever completed a specific test in *quop*.

4.2 Quop-L2 Reading Progress Assessment

In administering the quop-L2 tests, for every item, both response accuracy (i.e. if a response is correct) and response speed (i.e. how fast a response is given) are recorded. This allows for analyzing response efficiency as a function of response accuracy and response time. In this sample, median test duration across all measurement time points amounted to 7.6 minutes. Each test consists of three subscales inspired by ProDi-L to assess the efficiency of component processes of reading comprehension on the word, sentence, and text level, respectively. On word level, to assess orthographic comparison, in a word/pseudoword discrimination task, students have to decide whether a presented word is real or not. On sentence level, capturing semantic integration, a sentence verification task prompts students to determine whether a sentence presented to them makes sense or not. Finally, on text level, in a story verification task measuring local coherence, students have to judge whether a third sentence complements a short story consisting of two sentences in a meaningful way. For sample items see Figure 2 (p. 21).

All subscales were constructed using rule-based item design. In rule-based item design, item properties affecting item difficulty are systematically varied to purposefully manipulate item difficulty. In this way, it is possible to explicitly construct items of desired difficulty levels without needing to calibrate every single item (Embretson, 1999; Holling et al., 2008). Consequently, test authors can cheaply construct an arbitrary number of equivalent test forms, making this approach especially beneficial for progress assess-

Figure 2
Sample Items and Item Properties.



Note. Depicted are two equivalent items per level and their design characteristics.

ments. The test authors of the quop-L2 assessment identified item properties that should theoretically influence item difficulty for every subscale (this relation was confirmed in Förster et al. (2017). Each property was coded dichotomously, with a 0 indicating no difficulty and a 1 indicating difficulty. Thus, the sum of the item properties served as an indicator of item difficulty. Apart from an item being correct or incorrect, three further item properties were used for every subscale (see Figure 2, p. 21). Systematically varying these, 20 items were constructed for the word level, and 13 items each for the sentence and text level. To generate equivalent test forms, unique items following the same design rules (i.e. the same pattern of item property characteristics) like the already existing ones were developed. With this procedure, a total of four strictly controlled equivalent test forms was created. For examples of two equivalent items per subscale, compare the items in Figure 2.

To prevent a confounding of item and measurement time point, in this study, the total sample was divided into eight groups, with each group completing a different combination of test halves per time point (cf. Klein Entink et al., 2009). Hence, over the first four time points, each group completed each item once. For the remaining four time points, the process was repeated. An illustration of the whole procedure can be found in Appendix A.

4.3 Validity Measures

4.3.1 ELFE II – Reading Comprehension

The ELFE II is a standardized achievement test assessing reading comprehension on the word, sentence, and text level that can be used throughout first to sixth grade. It is designed as a speed test, so that students are given three minutes time to solve as many of the 75 items on word level as possible. The same applies to the 36 and 26 items on sentence and text level, respectively. For each correctly solved item, a test-taker is awarded one point. On word level, students must choose one out of four words best describing a picture. All words are similar in terms of graphemes and phonemes and exhibit the same number of syllables. On sentence level, in one place in a sentence, the correct out of five possible words must be inserted. Sentences differ in complexity and length. On text level, short texts consisting of two to four sentences are presented. Students must answer a question related to the text by choosing one out of possible options. For a correct answer, information from the text must be integrated either only across adjacent or across multiple sentences. The test authors report odd-even split-half reliabilities of $r \geq .89$ for all levels and grades in the standard paper-pencil version used in this study. ELFE II total scores (normed) correlate high with another standardized reading test, the SLS 2-9 (r = .77; Wimmer & Mayringer, 2014) and lower with the CFT 1-R (r = .39).

4.3.2 CFT 1-R – Intelligence

The CFT 1-R is a standardized achievement test assessing general intelligence from first grade to the beginning of third grade with non-verbal material. It is organized into two parts measuring different aspects of general intelligence. Part one assesses perceptual speed while part two assesses basic intellectual skills. We only used part two here as an indicator of intelligence, since computing an overall score across both parts requires a certain threshold between test parts to not be exceeded which could have led to high dropout. Basic intellectual skills are assessed with three subtests. Students must solve a rule detection task, a classification task and a matrix task. Retest reliability for part two is reported as r = .94 and correlations with other intelligence tests amount to r = .63 (HAWIK; Petermann & Petermann, 2008) and r = .50 (DVET; Meis et al., 1997).

4.3.3 DEMAT 1+ – Mathematics

The DEMAT 1+ is a standardized achievement test assessing mathematical competence in first and at the beginning of second grade. It consists of nine subtests comprising a total of 36 different tasks that are rooted inside the curriculum of all German federal states. Subtests contain for example subjects like addition, subtraction, and sets. Internal consistencies for second grade are reported to be r = .88. The test correlates with teacher judgments of mathematical performance (r = .66), and correlations with an informal measure of addition and subtraction performance (DZB 1; Wagner & Born, 1994) amount to r = .77. We used both available pseudo-parallel forms in this study.

4.3.4 Teacher Judgments

Teachers judged their students' reading skills on the word, sentence, and text level in both a dimensional and a criterial way. For the dimensional judgment, teachers rated their students' reading skills per level on a 7-point Likert scale ranging from *far below average* to *far above average*. For the criterial judgment, they estimated how many word, sentence, and text items their students would correctly solve within two minutes. This judgment referred to items from a pilot study conducted earlier. As an orientation, the average performance of all students participating in the pilot study was given.

4.4 Analytic Strategy

For all analyses, the open source software R (version 3.6.3; R Core Team, 2020) was used along with a confidence level of α = .05. In a preprocessing step, items for which negative response times were recorded due to technical issues as well as any reading progress assessments completed outside regular school hours were coded as missing. Analyses with respect to structural validity comprised the total sample, while analyses with respect to status and slope validity were based on the validation subsample. As validity requires reliability, before turning to the main analyses, I ran a brief reliability analysis on the quop-L2 scores to make sure validity results had a meaningful basis. To this end, odd-even split-half reliabilities as well as retest-reliabilities between consecutive tests were computed.

4.4.1 Quop-L2 Reading Progress Assessment Scores

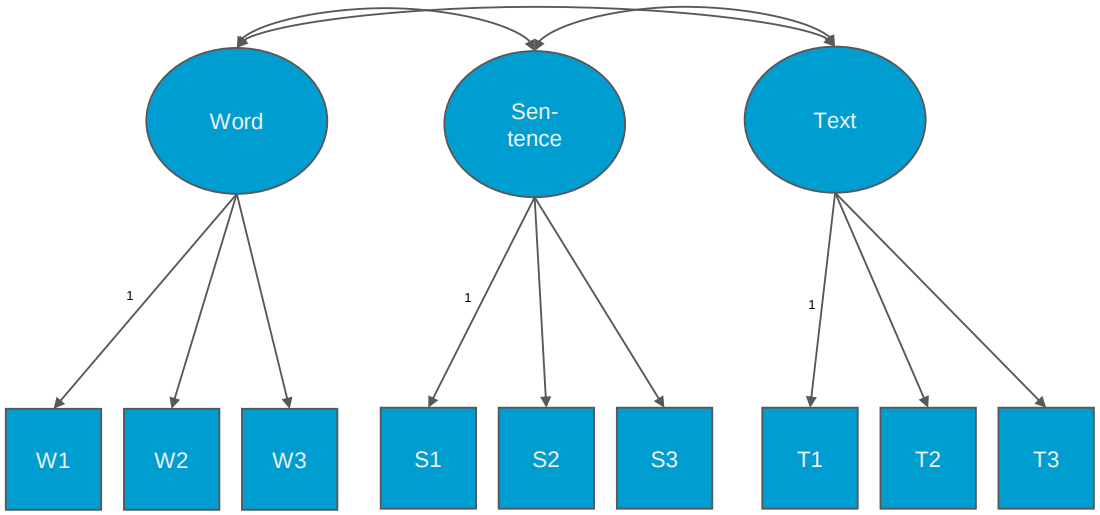
Incorporating both response accuracy and response time into efficiency scores, I implemented a variant of the correct item summed residual time (CISRT) proposed by van der Maas and Wagenmakers (2005). In this scoring procedure, based on a limit for response times, a correct response is rewarded the remaining time, while the score for an incorrect response is zero. Thus, a fast correct response, indicating higher efficiency, results in a higher score than a slow correct response, indicating lower efficiency. Scores are then just summed over items. Instead of defining a response time limit in advance, I applied quantile-based cutoffs. Specifically, scores below the 5%-quantile or above the 95%-quantile of the response time distribution were considered missing. Basing response time limits on the distribution of scores, these limits seem less arbitrary compared to setting them a priori. Moreover, this procedure helps to reduce bias introduced by guessing (CISRT scores are susceptible to guessing as a fast correct guess will be rewarded with a higher score than a slower, truthful response). By not only considering a lower limit for response times but also an upper one, implausibly long response times are handled as well. Additionally, to allow for a more straightforward interpretation regardless of absolute times, I expressed the CISRT scores as the amount of time remaining after giving a correct response. Hence, scores were not summed but averaged across items per level. In this step, item-median imputation was used in case of missing scores. Taken together, the CISRT score for one student on a scale with n items was calculated like described in equation 1,

$$\frac{\sum_{i=1}^{n} a_i * ((1 - \frac{r_i - b_l}{b_u - b_l}) * 100)}{n} \tag{1}$$

where a_i denotes the response accuracy for item i (resolving to zero ore one), r_i represents the response time for item i and b_l and b_u stand for the lower and upper limit for response time, respectively. The CISRT score thus represents the average amount of time remaining for responding to an item across the scale[3].

[3] Although averaging and not summing remaining times like the term CISRT implies, I refer to the scores calculated in this study as CISRT scores since averaging vs. summing does not change the mathematical logic behind or the statistical results based upon these scores.

Figure 3
Structural Validity Model.



Note. For simplicity, error variances and variances for observed and latent variables are not displayed in the figure.

4.4.2 Structural Validity

To verify the postulated structure of the quop-L2 reading progress assessment, I conducted CFAs[4] (R package lavaan; Rosseel, 2012) for every measurement time point. The supposed model structure assumes that scores for each level (word, sentence, and text) load on a different latent variable, representing reading comprehension on the respective levels (see Figure 3, p. 25). Covariances between the latent variables were allowed, as component processes proposedly interact between levels (Richter et al., 2012). The CFAs were estimated with three item mean parcels per level serving as indicators. Parcels were built by counterbalancing item positions, referring to W. Lenhard, Schroeders, et al. (2017). In effect, the first item was assigned to parcel one, the second to parcel two, the third to parcel three, and the fourth again to parcel one etc. This procedure should prevent bias caused by potential differential student motivation across the test duration from affecting the parcels.

[4]For these as well as the slope validity SEM models reported below, the clustered structure of the data (students being nested into classes) was considered by calculating cluster-robust standard errors. When not investigating cluster-specific effects, this procedure allows for a pragmatical way of accounting for the clustered structure while making less assumptions as compared to explicitly modeling it (McNeish et al., 2017). It is to be noted that I did not consider the school level, as I believe most clustering effects are linked to the classroom with its unique teacher rather than to the school. This assumption could however not be tested as lavaan does not (yet) support the consideration of more than one cluster variable.

4.4.3 Status validity

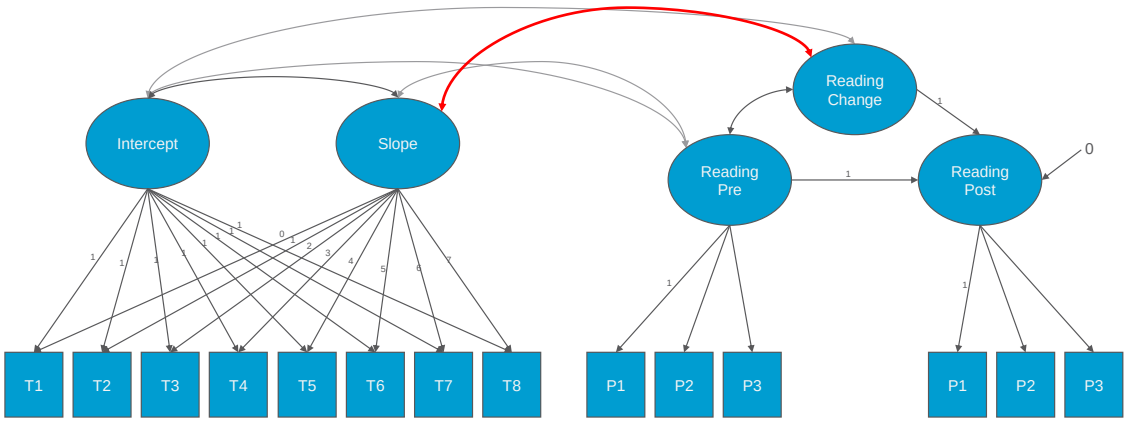
To obtain indicators of convergent status validity, Pearson correlations between both the standardized reading pre- and posttest and all quop-L2 test were computed per level. Similarly, both the dimensional as well as the criterial teacher judgments were related to the quop-L2 assessments. Here, correlations were calculated per class, Fisher-Z-transformed and then averaged, as each teacher judged the students from their respective class. Regarding discriminant status validity, quop-L2 scores were averaged across levels after standardizing them to obtain total scores per measurement time point. These were then correlated to the standardized measures of intelligence and mathematics.

4.4.4 Slope validity

As discussed above, more advanced modeling techniques like HLM or SEM may be preferable over OLSR when dealing with slopes representing reading progress. In this study, I decided to model slope in an SEM framework as I wanted to capture the latent construct of reading comprehension that can only be measured indirectly via manifest indicators. Also, measurement error occurring in this process can be separated from the assessment of the latent construct of interest in this approach. Lastly, SEM is very flexible, particularly allowing to estimate slope and relate it to other quantities within the same model instead of having to specify two separate models to do this. Like for structural validity, models were estimated using the R-package lavaan (Rosseel, 2012).

To investigate whether slope indeed captured reading progress, slope from the quop-L2 tests was related to pre-post change in the standardized reading measure per level. Modeling slope, a linear latent growth model (LGM) with the CISRT scores for all quop-L2 assessments serving as indicators was estimated. Here, a slope factor was modeled by linearly increasing factor loadings over the quop-L2 measurement time points. As previous research has yielded inconsistent results when controlling for performance level at different time points, I decided to control for performance level across all time points. Therefore, an intercept factor was modeled by fixing the factor loadings of all measurement time points to one. Like this, the intercept factor should capture mean performance level across time points. Covariances between intercept and slope factor were allowed, as level and change are often correlated in reading tests (Baker et al., 2008).

Figure 4
Combined LGM/LCM Model.

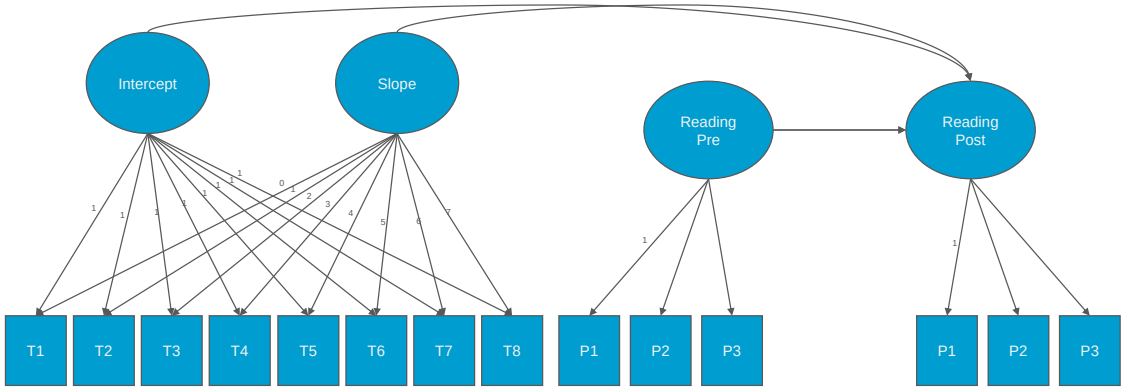


Note. For simplicity, error variances and variances as well as means and intercepts for observed and latent variables are not displayed in the figure. Indicator-specific covariances between the same parcels were allowed over measurement time points.

With only two measurement time points available for the standardized reading measure, I modeled the difference between pre and post scores in a latent change model (LCM). To this end, item parcels built in the same way as described above served as indicators for the pre- and the posttest performance, respectively. In the structural model, the post-performance was predicted in a perfect linear regression by the pre-performance and an artificial latent difference variable. As a result, change between pre- and posttest was captured in the difference variable. Again, covariances between the pre-performance and the difference variable were allowed. Finally, to relate slope to pre-post change in the standardized reading measure, both the LGM and the LCM were estimated simultaneously while additionally obtaining the covariance between the LGM slope and the LCM change as an indicator of slope validity. In this combined model, covariances between the latent variables were allowed. For a graphical illustration of the combined model, refer to Figure 4 (p. 27).

To verify that slope predicted end-of-term performance, reading performance in the standardized posttest was predicted per level by quop-L2 slope, controlling for quop-L2 intercept and standardized pretest performance. Again, slope and intercept were modeled in an LGM as described above. Likewise, standardized pre- and post-performance were estimated with three item parcels each. In the structural model, standardized posttest performance was regressed on quop-L2 slope and intercept as well as on pretest performance, allowing all covariances. This model structure is depicted in Figure 5 (p. 28).

Figure 5
Predictive Utility Model.



Note. For simplicity, error variances and variances as well as means and intercepts for observed and latent variables are not displayed in the figure. Indicator-specific covariances between the same parcels were allowed over measurement time points.

4.4.5 Missing Data

Per quop-L2 measurement time point and level, 7% to 19.86% of data was missing in the total sample, and 3.11% to 15.25% in the subsample, e.g. due to students missing a test because of illness. Also, between pre and post assessment of the validation measures, there was some fluctuation, resulting in 7.06% to 14.41% missing data for the standardized tests. For teacher judgments, missing data amounted to 2.54%. With respect to correlational analyses, I used all pairs of values that were available. In effect, every correlation reported below was based on a median of 309 students (ranging from 256 to 334) and should have stabilized according to Schönbrodt and Perugini (2013). Therefore, regarding correlations incorporating a different number of cases depending on missingness in specific variables was considered unproblematic. For an overview of how many complete pairs were available for which variables, see Appendix B. Handling missingness in the context of SEM models, I used full information maximum-likelihood (FIML; Enders, 2001). FIML is straightforward to implement in lavaan and uses all data available when estimating model parameters. Hence, for example a subject missing only one out of eight quop-L2 tests did not need to be excluded from the analysis.

Figure 6
Quop-L2 Mean Scores Across Time Points.

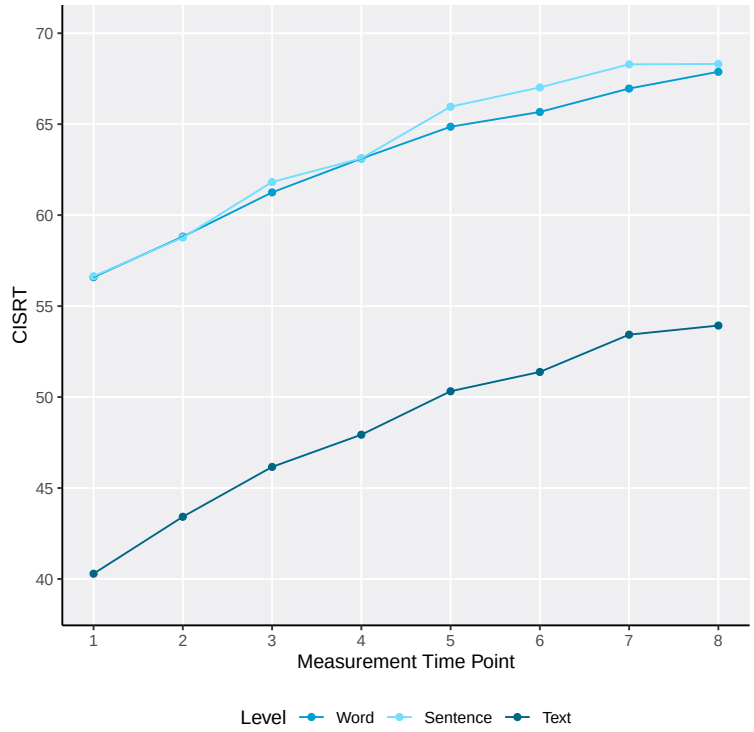
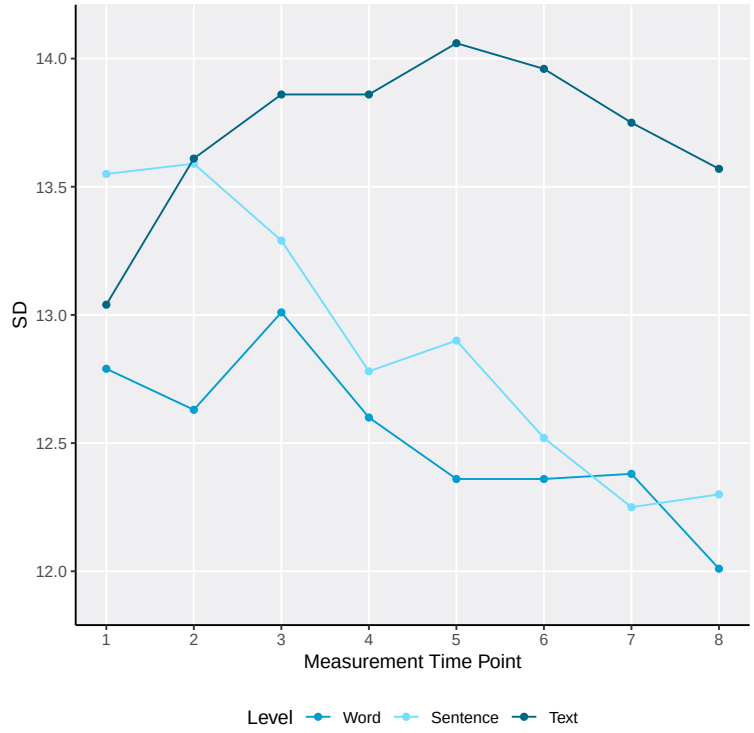


Figure 7
Standard Deviations of Quop-L2 Scores Across Time Points.



5 Results

Means and standard deviations for the quop-L2 assessments per level are shown in Figures 6 (p. 29) and 7 (p. 29; The range of standard deviations across all measurement time points per level amounted to the following: SD_{word} = 12.01-13.01, $SD_{sentence}$ = 12.25-13.59, SD_{text} = 13.04-14.06). Descriptively speaking, progress can be observed across all levels. Likewise, student's scores descriptively improved from standardized reading pretest to posttest with a mean difference of M_{diff} = 9.29 (SD_{diff} = 1.44) on word level, M_{diff} = 4.78 (SD_{diff} = 0.78) on sentence level and M_{diff} = 3.82 (SD_{diff} = 1.05) on text level. More detailed information on means and standard deviations for all measures can be found in Appendix C. Exemplary correlations of quop-L2 scores from the second time point with all validation measures are presented in Appendix D.

5.1 Reliability

Median and range of retest as well as odd-even split-half reliabilities per level across all measurement time points are depicted in Table 1 (p. 31). For an overview of all reliability coefficients please see Appendix E. Here, split-half reliabilities for the word and the sentence level were r = .82[5] and above for all measurement time points. On text level, they reached r = .74 and above, except for the first time point yielding a coefficient of r = .65. It could be the case that students needed the first test to get used to the items on text level, which appear to be the most difficult descriptively (see Figure 6, p. 29), resulting in a lower split-half reliability for this time point. Retest reliabilities were collectively lower, ranging between r = .63 and r = .71 across all levels and measurement time points. This was to be expected, as the quop-L2 assessment was constructed to be sensitive to reading progress. Too high stability across time would counteract this sensitivity. In sum, the reliability analyses indicate no reservations against investigating quop-L2 validity due to missing reliability.

[5] As classifying the absolute value of a correlation coefficient into categories like *high*, *moderate*, or *low* is always a somewhat arbitrary process, I forgo evaluations like these in this study.

Table 1
Median and Range of Split-Half and Retest Reliabilites Across Quop-L2 Time Points.

	Split-Half Reliabilites		Retest Reliabilites	
Level Quop-L2	Median	Range	Median	Range
Word	.86	.82 - .89	.66	.63 - .71
Sentence	.90	.86 - .92	.66	.64 - .68
Text	.76	.65 - .83	.66	.64 - .68

Note. All correlations are significantly different from zero with $p < .001$.

5.2 Structural Validity

For every measurement time point, a CFA examining the postulated three-dimensional structure of reading comprehension in the quop-L2 reading progress assessment was estimated. Please note that I relied on common relative and incremental fit indices to quantify model fit here. I did not report χ^2-test statistics as this test tends to overemphasize small differences between the model-implicit and the population covariance matrix for large samples (Schermelleh-Engel et al., 2003). CFA models fit well to the data across all measurement time points. The relatively worst fit was obtained at time point seven (*RMSEA* = .04, *SRMR* = .02, *CFI* = .99, *TLI* = 0.98) and the relatively best fit at time point four (*RMSEA* = .03, *SRMR* = .02, *CFI* = .99, *TLI* = 0.99). An overview of fit indices across all time points can be found in Appendix F. Results therefore provide empirical evidence for the conceptualization of reading comprehension as efficient component process on the word, sentence, and text level in the quop-L2 reading progress assessment.

5.3 Status Validity

Investigating status validity, quop-L2 reading progress assessment scores were related to performance in convergent and discriminant measures. Correlations between static quop-L2 scores and standardized reading pre- and posttest as well as teacher judgments per level served as indicators for convergent status validity. Median and range of all convergent correlations per level can be found in Table 2 (p. 33). Quop-L2 scores correlated

with the standardized reading pre- and posttest on word level with at least r = .60 across all measurement time points. On sentence level, correlations scored even higher with most being above r = .71. Coefficients on text level in turn were found to be lower with a median of r = .60 for the relation between quop-L2 scores and the reading pretest and r = .62 for the posttest. Here, across the first three measurement time points, correlations for both reading pre- and posttest ranged below the medians with the first time point yielding the lowest values of r = .52 and r = .53 for the pretest and the posttest, respectively. For time points four to eight, coefficients increased so that nearly all values lay at r = .61 and above. With respect to the teacher judgments, a similar pattern could be observed. For both dimensional and criterial judgments on word level, three quarters of the correlations with quop-L2 scores across time points reached a level of r = .61 and above. On sentence level, all correlations lay at r = .61 or above, with the majority being located at r = .65 and more. Like with standardized reading scores, correlations of quop-L2 scores and teacher judgments on text level were lower compared to the other levels and tended to rise across time points. Starting only slightly above r = .40 for the first quop-L2 test, correlations increased to r = .60 and above for later measurement time points. Discriminant correlations of total quop-L2 scores across all time points with standardized measures for intelligence and mathematics amounted to median coefficients of r = .39 (range r = .32 - .43) and r = .43 (range r = .40 - .47) respectively. As a result, convergent correlations were consistently higher than discriminant ones, except for the teacher judgments on text level in relation to the first quop-L2 measurement time point. For a detailed depiction of all convergent and discriminant correlation coefficients please consult Appendix G. Disattenuated correlation coefficients (coefficients that are corrected for reliability constraints) are additionally reported in Appendix H. Note that with respect to teacher judgments, no disattenuation could be realized as with the available data, no reliability estimation was possible.

5.4 Slope Validity

The investigation of slope validity consisted of two parts. Firstly, quop-L2 slope was related to pre-post change in the standardized reading test. Secondly, quop-L2 slope was used to predict standardized reading posttest performance while controlling for quop-L2

Table 2
Median and Range of Convergent Status Validity Correlations Across Quop-L2 Timepoints.

Reading Pre		Reading Post		TJ Dimensional		TJ Criterial	
Median	Range	Median	Range	Median	Range	Median	Range
			Word				
.63	.60 - .65	.64	.62 - .66	.63	.57 - .68	.60	.55 - .68
			Sentence				
.75	.65 - .81	.73	.69 - .76	.65	.61 - .67	.64	.62 - .68
			Text				
.60	.52 - .63	.62	.53 - .66	.59	.42 - .65	.57	.41 - .62

Note. Levels for all Measures: Word, Sentence and Text. Reading Pre = ELFE II Pretest. Reading Post = ELFE II Posttest. TJ = Teacher Judgments. All correlations are significantly different from zero with $p < .001$.

intercept and standardized reading pretest performance. As a prerequisite for meaningfully relating two estimates of progress, both slope estimated in the LGM ($M_{\delta word} = 1.56$, $SE = 0.22$, $p < .001$; $M_{\delta sentence} = 1.66$, $SE = 0.26$, $p < .001$; $M_{\delta word} = 1.73$, $SE = 0.19$, $p < .001$) and pre-post difference estimated in the LCM ($M_{\delta word} = 1.58$, $SE = 0.12$, $p < .001$; $M_{\delta sentence} = 1.74$, $SE = 0.08$, $p < .001$; $M_{\delta text} = 1.44$, $SE = 0.13$, $p < .001$) showed substantial progress for all levels. Goodness-of-fit indices of the combined models per level are shown in Table 3 (p. 34) indicating good fit to the data on word level and acceptable fit on sentence and text level. Regarding the covariance between latent quop-L2 slope and latent pre-post change in the standardized reading test, latent slope and latent change correlated significantly on sentence level ($r = .50$, $SE = 0.09$, $p < .001$) and on text level ($r = .29$, $SE = 0.12$, $p = .01$). However, on word level, no significant correlation could be found ($r = .09$, $SE = 0.14$, $p = .51$). Consequently, progress in the quop-L2 assessment was related to progress between standardized reading pre- and posttest on sentence and text level, but not on word level. Estimating the model structure predicting standardized reading posttest performance revealed a good fit to the data on all levels (see 3). All three predictors, namely latent standardized reading pretest performance, latent quop-L2 intercept and latent quop-L2 slope significantly influenced latent standardized reading posttest performance (see Ta-

Table 3
Goodness-of-Fit Indices for Slope Validity Models per Quop-L2 Level.

	Combined LGM/LCM Model				Predictive Utility Model			
Level Quop-L2	RMSEA	SRMR	CFI	TLI	RMSEA	SRMR	CFI	TLI
Word	.02	.03	1.00	1.00	.02	.03	1.00	1.00
Sentence	.07	.05	.98	0.98	.03	.03	1.00	1.00
Text	.07	.07	.96	0.96	.03	.04	.99	0.99

Note. Combined LGM/LCM relating quop-L2 slope to standardized reading pre-post change. Predictive Utility Model using quop-L2 slope to predict standardized reading end-of-term performance beyond performance level and beginning-of-term performance.

Table 4
Predicitive Utility Model Results.

	Predictor					
	Quop-L2 Slope		Quop-L2 Inter		Reading Pre	
Level Quop-L2	β	SE	β	SE	β	SE
Word	0.22**	0.08	0.47***	0.08	0.49***	0.08
Sentence	0.31***	0.05	0.37***	0.08	0.64***	0.09
Text	0.31***	0.08	0.43***	0.08	0.46***	0.07

Note. Prediction of Standardized Reading End-of-Term Performance by Quop-L2 Slope, Performance Level and Beginning-of-Term Performance across Levels. End-of-Term Performance = ELFE II Posttest. Quop-L2 inter(cept) = Performance Level. Reading Pre = Beginning-of-Term Performance (ELFE II Pretest). Significance Codes: *p < .05. **p < .01. ***p < .001.

ble 4, p. 34). Therefore, while controlling for quop-L2 intercept and standardized reading pretest performance, quop-L2 slope was found to predict standardized reading posttest performance. This held true for all levels, so for both word (β = 0.22, SE = 0.08, p < .01), and sentence level (β = 0.31, SE = 0.05, p < .001), as well as text level (β = 0.31, SE = 0.08, p < .001). Thus, quop-L2 reading progress predicted end-of-term performance in reading comprehension incremental to performance level and beginning-of-term performance.

6 Discussion

This study aimed to comprehensively validate the newly developed quop-L2 reading progress assessment to ensure it could be put to its intended use, i.e. to inform instructional decisions based on differentiated diagnostic information about reading progress. To this end, its (1) structural validity, (2) status validity, and (3) slope validity were investigated based on a sample of N = 1989 and subsample of n = 354 second grade-students. I employed CFAs verifying the assessment's proposed structure of measuring the efficiency of component processes of reading comprehension on the word, sentence, and text level (structural validity). Then, I correlated static quop-L2 scores to convergent measures of standardized reading and teacher judgments alongside discriminant measures of intelligence and mathematics (status validity). Next, adhering to best-practice recommendations developed above, I related latent quop-L2 slope to a) latent pre-post change in standardized reading performance and b) standardized reading end-of-term performance in an SEM context (slope validity). Results indicated structural validity thereby confirming that quop-L2 provides differentiated diagnostic information on reading comprehension as opposed to the robust indicators prevalent so far. In the two-step approach proposed by L. S. Fuchs (2004), first, status validity could be inferred from virtually all correlations between quop-L2 scores and convergent measures scoring higher than correlations between quop-L2 scores and discriminant measures. Second, slope validity was found a) on sentence and text level as quop-L2 slope was related to standardized reading pre-post difference as an estimate of change and b) on all levels as quop-L2 slope successfully predicted end-of-term performance. Taken together, these results build a strong case for the assessment's validity and thus encourage using quop-L2 as a differentiated diagnostic information source for instructional decisions. Due to the comprehensive and complex perspective on validity offered here, this study not only reveals the validity of quop-L2, thereby encouraging its intended utilization, but also importantly contributes to the current state of research on the validity of learning progress assessments in general.

6.1 Structural Validity

The cognitive perspective of conceptualizing reading comprehension as the efficiency of component processes on the word, sentence, and text level (Richter & Christmann,

2009) has hitherto only been available and successfully validated in status assessments (ProDi-L; Richter et al., 2012). In this study, structural validity results indicated a successful operationalization of this process-based approach to reading comprehension for the quop-L2 reading progress assessment: CFAs conducted over all eight quop-L2 measurement time points in a sample of close to 2000 second grade students fit well to the proposed structure of capturing the efficiency of reading comprehension as interconnected, yet separable component processes on the word, sentence, and text level. Consequently, under the prerequisite of intact status and slope validity, quop-L2 is the first assessment to offer differentiated information about reading progress as opposed to the robust indicators prevalent far (Förster & Souvignier, 2011). This differentiated diagnostic information comprises not only information on the development of reading comprehension on separate levels, but tapping its efficiency integrates both response accuracy and response time to a holistic perspective on reading progress. Overall, quop-L2 is the first reading progress assessment providing differentiated diagnostic information on reading comprehension thus offering a broad basis for instructional decisions.

6.2 Status Validity

Past studies have often only regarded single aspects of status validity by investigating convergent or discriminant validity alone (e.g Fewster & Macmillan, 2002; Jenkins & Jewell, 1993). This study investigated status validity in its entirety, relating multiple convergent and discriminant measures to quop-L2 scores across all time points on the highest possible level of detail. This was the word, sentence, and text level, respectively for convergent measures and the level of total scores for discriminant measures. Results showed that virtually all correlations between quop-L2 scores and convergent measures scored higher than correlations between quop-L2 scores and discriminant measures. This held true across quop-L2 levels and measurement time points for relations to the standardized reading pre- and posttest as well as teacher judgments as convergent measures and standardized tests of intelligence and mathematical competence as discriminant measures. Several meta-analyses (Reschly et al., 2009; Shin & McMaster, 2019; Yeo, 2010) investigated the relation between CBM scores (mostly ORF and maze tasks) and standardized reading tests. The average correlations found in these meta-analyses were close

to the median convergent correlations of quop-L2 scores with standardized reading tests across all levels found in this study. Therefore, these convergent quop-L2 status validity indicators fit well to those aggregated across CBM research. In other words, the differentiated feedback provided by quop-L2 does not come at the expense of lower convergent associations to other reading tests, highlighting this advantage of quop-L2 as compared to robust indicators like ORF or maze tasks. Considering convergent teacher judgments, the height of median coefficients with quop-L2 scores is close to that between teacher judgments and CBM found by Jenkins and Jewell (1993), but lower than those reported by Baker and Good (1995). However, for a meaningful integration of this study's convergent teacher judgment results, the body of research in this area is too small. The same applies for divergent status validity, which has only been investigated in a few studies so far (e.g Fewster & Macmillan, 2002; Kranzler et al., 1998). Within this study in any case, relative differences between convergent teacher and discriminant correlations also in comparison to convergent standardized reading test correlations support quop-L2 status validity.

Reviewing convergent status validity correlations in greater detail, it is striking that quop-L2 scores on text level correlated lowest with standardized reading pre- and posttest compared to the word and sentence level. Especially for the first measurement time point, the correlations on text level stood out to be comparatively low. Across time points, they then tended to rise although staying below those on the word and sentence level. As the same pattern applies for the quop-L2 split-half reliabilities, this conspicuousness likely is a methodological artifact. Due to restricted reliability on text level, correlations were underestimated. Disattenuated correlations turned out to be more similar between the text and the remaining two levels (see Appendix H). Reliability probably was lowest on text level, as this level was designed to be the most difficult one. Descriptive mean scores seem to confirm that this aspect was successfully implemented (see Figure 7). With items being mostly difficult, reliability decreases along with the difficulty distribution's variance (Gulliksen, 1945). Especially for the first measurement time point, with students being confronted with a new test format, difficulty could have had the highest impact resulting in the lowest reliability. While students got more used to the test across time points, this connection might have faded leading to increasing reliability alongside validity coefficients.

As for correlations between quop-L2 scores and teacher judgments, although the absolute values lie lower than the correlations between quop-L2 scores and the standardized reading test, the result pattern was similar: Coefficients were lowest for the text level as compared to the other levels, most distinctively so for the first measurement time point. Specifically, while relations between quop-L2 scores and teacher judgments on the remaining two levels fit well to the mean correlation of between standardized student achievement and teacher judgments found by Südkamp et al. (2012) in their meta-analysis, correlations on text level deviated slightly. Here, coefficients were found to be slightly lower in general with a pronounced drop for the first measurement time point. In addition to reliability restrictions on quop-L2 text level, it might be the case that teacher judgments on text level show reliability issues, too. It is not possible to reveal this circumstance as a methodological artifact by disattenuating the correlations, because with the data at hand no reliability estimates for the teacher judgments could be obtained. However, Förster and Böhmer (2017) provide a suitable theory to support the claim of missing reliability in teacher judgments on text level. Based on the lens model (Brunswik, 1956; Hammond, 1996; Nestler & Back, 2013) they argue that teachers judge students' traits that cannot be directly observed by resorting to observable cues. For example, a students' reading comprehension could be judged based on behaviors that are related to reading comprehension, like how fluently they read aloud or how fast they talk. The more valid cues are used in this process while simultaneously ignoring invalid cues (in this example maybe the student's mental arithmetic speed), the more accurate the judgment will be. While for e.g. word comprehension, many valid cues like word fluency exist that are easily observable and distinguishable from invalid ones, the situation might be more complicated for text comprehension. Here, less observable factors like prior knowledge necessary for text comprehension come into play, making it harder to judge resulting in a lower reliability for this level. Summarizing, the result pattern for quop-L2 scores' correlations with teacher judgments could be the same as that for quop-L2 scores' correlations with the standardized reading test due to comparatively lower reliability of quop-L2 scores on text level as elaborated above. That with respect to teacher judgments, the absolute values scored lower than those for the standardized reading test could be explained by potentially more difficult judgments on text level resulting in lower reliability also for this measure thus further impeding validity coefficients.

Despite detailed contemplations like these, this study overall provides strong evidence for status validity of the quop-L2 assessment. Although suffering from potential reliability restrictions on text level with respect to early measurement points and teacher judgments, an overall image of strong status validity emerges: Both the comparison of convergent and discriminant correlations within this study and the integration of results into previous research - where applicable - clearly support quop-L2's status validity which has here been analyzed in its entirety. Therefore, using the differentiated diagnostic information provided by quop-L2 as a basis for instructional decisions is encouraged by these results. From a theoretical point of view, specifically previously scarcely used teacher judgments as well as discriminant validation measures in general have been employed. Thus, beyond proving quop-L2 status validity, this study contributes to research on validating learning progress assessments in general.

6.3 Slope Validity

Previous research on CBM slope validity often produced heterogeneous results. For this reason, I developed best-practice recommendations for validating slope in reading progress assessments which were at most implemented partially in studies conducted hitherto. Specifically, I firstly avoided modelling slope with OLSR as this could yield imprecise estimations (Christ, 2006; Hintze & Christ, 2004). Rather, a more flexible SEM approach allowing to estimate slope and relate it to other quantities within one model was applied here. Secondly, I controlled for intercept, i.e. level of performance, when modelling slope. This is necessary as intercept and slope are often correlated in reading (Baker et al., 2008). As past research showed inconsistent results when controlling for performance level at different time points of reading progress assessment (e.g. Schatschneider et al., 2008; Wanzek et al., 2010), I modeled performance level across all measurement time points. Thirdly, additionally to performance level, I controlled for beginning-of-term performance when relating slope to end-of-term performance to solidify potential results, as it has so far been done in single studies only (Baker et al., 2008; Tolar et al., 2014). Fourthly, as conceptualized by Yeo et al. (2011), not only relating slope - being an estimate of progress - to static scores, I also evaluated slope with respect to an estimate of change. Being the first study completely adhering to these best-practice recommen-

dations, methodologically sound and reliable results could be produced. In this, analyses were twofold: Quop-L2 slope was a) related to standardized reading pre-post change to verify quop-L2 slope represented reading progress and b) used as a predictor of standardized reading end-of-term performance while controlling for performance level and beginning-of-term performance. Apart from one exception, the results indicated slope validity. On sentence and text level, quop-L2 slope correlated with standardized reading pre-post change indicating slope indeed represented reading progress. This reading progress in turn predicted standardized reading end-of-term performance beyond performance level and beginning-of-term performance therefore probably conveying unique information. Results accordingly hint at the slope's robust predictive utility and thus at its instructional relevance. On word level, quop-L2 slope representing reading progress could not be verified as it did not correlate with standardized reading pre-post change. However, a substantial slope, so progress of some form, was captured that possessed predictive utility with respect to standardized reading end-of-term performance beyond performance level and beginning-of-term performance. Taken together, comparatively reliable and consistent results in favor of slope validity could be found.

Revisiting the unexpectedly missing relation on word level, no association could be found between quop-L2 word reading slope and standardized reading pre-post change. As discussed by previous studies (Cho et al., 2018; Tolar et al., 2014), the alignment of measures could influence the relation between quop-L2 slope and standardized reading test pre-post change on word level. Although the quop-L2 assessment was designed to target orthographic comparison as a component process of word reading comprehension, it uses word rareness as an item property to influence item difficulty, resulting in the use of infrequent words. However, orthographic comparison is viewed as especially important for visually recognizing frequent words (Andrews, 1982). Phonological recoding, as another component process on word level, on the other hand is thought to specialize in the recognition of infrequent words (Richter et al., 2012). It could therefore be the case that the standardized reading test, utilizing only frequent words, is better suited to measure orthographic comparison while the quop-L2 assessment rather assesses phonological recoding efficiency. Whereas this difference is neither reflected in status validity results nor in the slope's predictive utility with respect to end-of-term performance, it only shows when relating quop-L2 slope and standardized reading test pre-post change. It might be

that a good reader must efficiently handle both component processes, i.e. orthographic comparison and phonological recoding, but that these component processes, or else their efficiencies, develop independently from each other. This could explain why static quop-L2 scores were related to standardized reading test scores and quop-L2 slope predicted end-of-term performance in the standardized reading test while change on both measures (i.e. slope and pre-post difference) was not associated. In total, a low alignment between quop-L2 and standardized reading test on word level, caused by measurement of different component processes, might have led to them being unrelated in terms of change. Maybe removing word rareness would drive the quop-L2 reading progress assessment more into the direction of capturing orthographic comparison efficiency, thus increasing alignment with the standardized reading test. Pursuing this idea, I re-estimated the latent correlation between quop-L2 slope and standardized reading pre-post change on word level based on quop-L2 scores incorporating only items using frequent words. However, the correlation remained non-significant (see Appendix I). Consequently, word rareness alone cannot explain differences between quop-L2 slope and standardized reading pre-post change in word comprehension. This withholds a necessary slope validity indicator on word level, namely evidence for quop-L2 slope on word level indeed representing progress in reading comprehension. Nonetheless, the substantial slope captured here was related to standardized reading end-of-term performance.

Summarizing, in a design adhering to current best-practice recommendations, slope validity of the quop-L2 reading progress assessment could be established. Specifically, on sentence and text level, both criteria for slope validity were fulfilled: Quop-L2 slope a) was linked to reading progress due to being related to another estimate of change and b) predicted end-of-term performance while controlling for performance level and beginning-of-term performance. On word level, although being unable to link quop-L2 slope to reading progress, a substantial slope, so progress of some form, was captured by the quop-L2 assessment, successfully predicting end-of-term performance in reading. Thus, the test proves – albeit in absence of the expected relation to another progress estimate - its predictive utility also on word level. Consequently, slope validity as established here supports that quop-L2 is providing differentiated diagnostic information about reading progress that is highly suitable as a basis for instructional decisions. As a theoretical implication, the unexpected result on word level underlines the importance

of slope validation in progress assessments as status validity alone does not guarantee a valid progress measurement. Further, the design used in this study is the first to integrate the best-practice recommendations for slope validation developed above. Beyond generating reliable evidence for quop-L2 slope validity it thus constitutes an important theoretical contribution to the so far often inconsistent research on slope validity by providing a reference point for designing future studies.

6.4 Practical Implications

Quop-L2 validly captures progress in the efficiency of component processes of reading comprehension that in turn predicts end-of-term performance beyond level and prior performance on the criterion measure. This encourages using the assessment in its intended way, i.e. as a basis for instructional decisions. It seems that the quop-L2 assessment is a good choice to monitor and facilitate students individually in learning how to read by continuously providing differentiated feedback about their reading progress. Thus, it allows for DBDM at classroom level on a frequent basis. Results even point into the direction of the quop-L2 assessment prospectively enabling DBDM also at school level. For example, in the context of the Every Student Succeeds Act (2015) in the USA, schools are held accountable for all students to pass federal state achievement tests applied at the end of the school year to ensure adequate educational standards. Learning progress predicting performance on such tests could be used to identify students at risk of failing them (Yeo et al., 2011)[6] and thus help schools to meet accountability goals. Naturally, direct generalizability of this study's result to the English language area is probably limited by differences between the English and German language like varying grapheme-phoneme correspondence (Goswami et al., 2005). Nevertheless, the prospect to use quop-L2 in DBDM on more than one level emphasizes its potential.

In any case, with quop-L2, there is a short and thus economical assessment available that can be used flexibly for individual or group measurement directly at school to orientate instruction towards optimal student reading success. Running online requiring nothing more than a browser, it is deployable cross-platform without needing to go through an

[6]Note that this study refers to the No Child Left Behind Act (2001), which was replaced by the Every Student Succeeds Act in 2015. Yearly standardized tests applied to hold schools accountable for educational outcomes remain a core part in the Every Student Succeeds Act.

installation process or similar potential sources of trouble. Therefore, quop-L2 could, beyond application in school, without any modifications support digital homeschooling, which is especially common during the ongoing COVID-19 pandemic. In a situation where teachers have less direct contact with their students, it is particularly important to obtain data about students' learning progress. Especially in the light of 82% of teachers across all school forms (90% in elementary schools) naming balancing out learning differences as the major challenge after fully reopening schools in a representative *Forsa*-survey (Verband Bildung und Erziehung, 2020, p.8), this data becomes highly relevant. Testing procedures like the quop-L2 reading progress assessment could help to quantify the actual size of this concern while at the same time showing starting points for an accurate facilitation by providing differentiated diagnostic information about reading progress.

6.5 Limitations and Future Research

While finding satisfactory reliability coefficients on the word and sentence level, reliability of quop-L2 scores on text level was comparatively lower. Therefore, the quop-L2 assessment could be improved for future use. Potentially, decreasing average item difficulty could increase reliability estimates. Likewise, it might be that teacher judgments about student performance lack reliability especially regarding text comprehension (Förster & Böhmer, 2017). Südkamp et al. (2012) note various methodological characteristics necessary to prevent harming teachers' judgment accuracy, like providing a frame of reference for their judgments. Beyond these important aspects, boosting reliability might be taken even further. It could prove helpful to train teachers before collecting their judgments in upcoming studies, e.g. based on the considerations offered by Förster and Böhmer (2017), thereby increasing their diagnostic accuracy. These measures could increase quop-L2 as well as teacher judgment reliability and thus also validity on text level.

Additionally, future research is needed to replicate and refine the validity results presented here, as validity is not determined in a single study alone (Messick, 1989). For one thing, further validity indicators reinforcing the placement of the quop-L2 reading progress assessment into its nomological network - both in the sense of static scores and slope - would be desirable. For another thing, specifically word level slope should

be examined more closely. Here, it remains yet unclear why the change in scores was unrelated between quop-L2 and standardized reading test assessments although static scores turned out to be associated and quop-L2 slope predicted end-of-term performance in the standardized reading test. To fully generalize the interpretation of scores obtained from a specific task to the construct domain in a broader sense, these scores should be related to those from tasks targeting the same construct (Messick, 1995a). Therefore, the missing link between the change in scores obtained from the two measures under consideration impedes the generalizability of quop-L2 word level slope to the domain of word comprehension progress. It would be very interesting to integrate this apparent inconsistency between static scores and slope in future work. Independent of the unexpected results with respect to slope on word level, future research should, unlike this study, validate quop-L2 slope using multiple other progress measures of both convergent and discriminant types. In doing so, quop-L2 slope validity would be strengthened by further evidence, particularly by also considering discriminant progress measures. What is more, future studies could solidify the relevance of quop-L2 progress scores for instructional decisions. In this domain, it could be investigated whether early reading progress predicts later reading difficulties, which would build a strong case for the instructional relevance of the quop-L2 assessment (see Kuhn et al., 2019 for an example in mathematics). Successfully predicting reading failure could accelerate the use of quop-L2 progress for DBDM not only on classroom but also on school level as illustrated above: Explicitly identifying students at risk of failing achievement tests could directly impact school accountability (Yeo et al., 2011). In sum, confirmation and expansion of the results obtained in this study, especially focusing on word level slope, are needed in a first step to complete quop-L2 validation. Based on this, in a second step, instructional relevance of quop-L2 scores should shift into focus.

Beyond the quop-L2 reading progress assessment, this study is the first to provide a design comprising both status and slope validity in their entireties while adhering to best practice recommendations developed based on the current state of research with respect to reading progress assessments. Backed on the encouraging results obtained in this study, it could be that this design type can be used to systematically investigate and disentangle additional factors influencing slope validity. These might for example be grade level or the alignment of measures (Cho et al., 2018; Tolar et al., 2014). Consequently,

dissemination of this design type might help to shed light on the so far inconsistent results in the field regarding slope validity, thus being one component in the roadmap to gain full comprehension of reading progress.

6.6 Conclusion

Overall speaking, results of an extensive validity analysis indicate structural as well as both status and slope validity of the newly developed quop-L2 reading progress assessment. Specifically, the proposed structure of conceptualizing reading comprehension as efficient component processes on the word, sentence, and text level could be confirmed. Static quop-L2 scores across all levels and measurement time points showed higher correlations with convergent achievement tests and teacher judgments as opposed to discriminant measures. Quop-L2 slope was related to pre-post change in standardized reading performance (albeit not on word level) and predicted end-of-term performance incremental to performance level and prior performance on the criterion measure across all levels. Consequently, the quop-L2 reading progress assessment seems capable of delivering differentiated diagnostic information about reading comprehension progress that can be used to inform instructional decisions to maximize student learning. This application of classroom-level DBDM could easily be transferred to digital homeschooling contexts allowing teachers to keep track of student progress during the current COVID-19 pandemic and become especially helpful when needing to balance out potential learning differences after reopening schools. Obtaining these results, this study is the first to use a comprehensive design for validating progress assessments adhering to best practice recommendations developed based on the latest state of research. This work thus provides a theoretical contribution to the development of progress assessment validation studies beyond the quop-L2 testing procedure. Dissemination of this design might help future research to integrate inconsistent results uncovered in this field so far.

7 References

Andrews, S. (1982). Phonological recoding: Is the regularity effect consistent? *Memory & Cognition*, *10*(6), 565–575.

Baker, S. K., & Good, R. (1995). Curriculum-based measurement of english reading with bilingual hispanic students: A validation study with second-grade students. *School Psychology Review*, *24*(4), 561–578.

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.

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

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press.

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

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. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review*, *35*(1), 128–133.

Christ, T. J., & Desjardins, C. D. (2017). Curriculum-based measurement of reading: An evaluation of frequentist and bayesian methods to model progress monitoring data. *Journal of Psychoeducational Assessment*, *36*(1), 55–73. https://doi.org/10.1177/0734282917712174

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443–507). American Council on Education.

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

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, *37*(3), 184–192.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*(4), 407–433.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, *8*(1), 128–141.

Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J. D. (2010). Creating a progress-monitoring system in reading for middle-school students: Tracking progress toward meeting high-stakes standards. *Learning Disabilities Research & Practice*, *25*(2), 60–75.

Every Student Succeeds Act. (2015). 20 U.S.C. Paragraph 6301.

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

Förster, N., & Böhmer, I. (2017). Das Linsenmodell – Grundlagen und exemplarische Anwendungen in der pädagogisch-psychologischen Diagnostik. In Südkamp A. & Praetorius A.-K. (Eds.), *Diagnostische Kompetenz von Lehrkräften* (pp. 46–50). Waxmann.

Förster, N., Kuhn, J.-T., Munske, J. S., & Souvignier, E. (2017). Construction of a test series for learning progress assessment in reading using rule-based item design. Talk at the 16th biennial meeting of the European Association of Research on Learning and Instruction. Tampere, Finland.

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

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

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

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

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. (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., & Compton, D. L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children*, *71*(1), 7–21.

Fuchs, L. S., Fuchs, D., 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

Goswami, U., Ziegler, J. C., & Richardson, U. (2005). The effects of spelling consistency on phonological awareness: A comparison of english and german. *Journal of Experimental Child Psychology*, *92*(4), 345–365.

Gulliksen, H. (1945). The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika*, *10*(2), 79–91.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. Oxford University Press.

Hintze, J. M., & Christ, T. J. (2004). An examination of variability as a function of passage variance in cbm progress monitoring. *School Psychology Review*, *33*(2), 204–217.

Holling, H., Blank, H., Kuchenbacker, K., & Kuhn, J.-T. (2008). Rule-based item design of statistical word problems: A review and first implementation. *Psychology Science*, *50*(3), 363.

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children*, *59*(5), 421–432.

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

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

Kim, Y.-S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology*, *102*(3), 652.

Klein Entink, R. H., Kuhn, J., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14*(1), 54.

Krajewski, K., Küspert, P., & Schneider, W. (2002). *DEMAT 1+: Deutscher Mathematiktest für erste Klassen*. Beltz.

Kranzler, J. H., Brownell, M. T., & Miller, M. D. (1998). The construct validity of curriculum-based measurement of reading: An empirical test of a plausible rival hypothesis. *Journal of School Psychology*, *36*(4), 399–415.

Kuhn, J., Schwenk, C., Souvignier, E., & Holling, H. (2019). Arithmetische Kompetenz und Rechenschwäche am Ende der Grundschulzeit. Die Rolle statusdiagnostischer und lernverlaufsbezogener Prädiktoren. *Empirische Sonderpädagogik*, *11*(2), 95–117.

Lenhard, W., Lenhard, A., & Schneider, W. (2017). *ELFE II - Ein Leseverständnistest für Erst-bis Sechstklässler*. Hogrefe.

Lenhard, W., Schroeders, U., & Lenhard, A. (2017). Equivalence of screen versus print reading comprehension depends on task complexity and proficiency. *Discourse Processes*, *54*(5-6), 427–445. https://doi.org/10.1080/0163853X.2017.1319653

Levin, J. A., & Datnow, A. (2012). The principal role in data-driven decision making: Using case-study data to develop multi-mediator models of educational reform. *School Effectiveness and School Improvement*, *23*(2), 179–201. https://doi.org/10.1080/09243453.2011.599394

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, *47*(2), 71–85. https://doi.org/10.1080/00461520.2012.667064

Marston, D. B. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. R. Shinn (Ed.), *Curriculum-based measurement* (pp. 18–78). The Guilford Press.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114.

Meis, R., Poerschke, J., & Lehmann, R. H. (1997). *DVET: Duisburger Vorschul-und Einschulungstest: Beiheft mit Anleitung und Normentabellen*. Beltz.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Macmillan.

Messick, S. (1995a). Standards of validity and the validity of standards in performance asessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Messick, S. (1995b). 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

Nestler, S., & Back, M. D. (2013). Applications and extensions of the lens model to understand interpersonal judgments at zero acquaintance. *Current Directions in Psychological Science*, *22*(5), 374–379.

No Child Left Behind Act. (2001). 20 U.S.C. Paragraph 6319.

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

Petermann, F., & Petermann, U. (2008). HAWIK-iv. *Kindheit und Entwicklung*, *17*(2), 71–75.

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

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.

Richter, T., & Christmann, U. (2009). Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. In N. Groeben & B. Hurrelmann (Eds.), *Lesekompetenz* (pp. 25–58). Juventa Verlag.

Richter, T., Isberner, M.-B., Naumann, J., & Kutzner, Y. (2012). Prozessbezogene Diagnostik von Lesefähigkeiten bei Grundschulkindern. *Zeitschrift für Pädagogische Psychologie*.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

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

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

Schermelleh-Engel, K., Moosbrugger, H., Müller, H., et al. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612.

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

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education*, *34*(3), 164–172.

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

Souvignier, E., Förster, N., Hebbecker, K., & Schütze, B. (2019). Quop: An effective web-based approach to monitor student learning progress in reading and mathematics in entire classrooms. In S. Jornitz & Wilmers A. (Eds.), *International perspectives on school settings, education policy and digital strategies*. Barbara Budrich.

Souvignier, E., Förster, N., & Zeuch, N. (2016). Lernverlaufsdiagnostik. In K. Seifried, S. Drewes, & M. Hasselhorn (Eds.), *Handbuch Schulpsychologie* (pp. 140–149). Verlag W. Kohlhammer.

Speece, D. L., & Ritchey, K. D. (2005). A longitudinal study of the development of oral reading fluency in young children at risk for reading failure. *Journal of Learning Disabilities*, *38*(5), 387–399.

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

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*(8), 795–819. https://doi.org/10.1002/pits.20113

Suchań, B., Wallner-Paschon, C., Stöttinger, E., & Bergmüller, S. (2007). *PIRLS 2006: Internationaler Vergleich von Schülerleistungen*. Leykam.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762. https://doi.org/10.1037/a0027627

Tarelli, I., Lankes, E.-M., Drossel, K., & Gegenfurtner, A. (2012). Lehr-und Lernbedingungen an Grundschulen im internationalen Vergleich. In W. Bos, I. Tarelli, A. Bremerich-Vos, & K. Schwippert (Eds.), *IGLU 2011 Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 137–173). Waxmann Verlag.

Tichá, R., Espin, C. A., & Wayman, M. M. (2009). Reading progress monitoring for secondary-school students: Reliability, validity, and sensitivity to growth of reading-aloud

and maze-selection measures. *Learning Disabilities Research & Practice*, *24*(3), 132–142.

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

van der Maas, H. L. J., & Wagenmakers, E.-J. (2005). A psychometric analysis of chess expertise. *The American Journal of Psychology*, *118*(1), 29–60.

Verband Bildung und Erziehung. (2020). *Die Arbeitssituation von Lehrkräften nach den Schulöffnungen*. http://vbe-nrw.de/downloads/Pressemitteilungen/2020_06_09__forsa_VBE_Corona_Schuloeffnungsphase.pdf

Wagner, H.-J., & Born, C. (1994). *Diagnostikum: Basisfähigkeiten im Zahlenraum 0 bis 20: DBZ 1*. Beltz-Verlag.

Wanzek, J., Roberts, G., Linan-Thompson, S., Vaughn, S., Woodruff, A. L., & Murray, C. S. (2010). Differences in the relationship of oral reading fluency and high-stakes measures of reading comprehension. *Assessment for Effective Intervention*, *35*(2), 67–77.

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

Weiss, R. H., & Osterland, J. (2012). *CFT 1-R: Grundintelligenztest Skala 1 - Revision*. Hogrefe.

Wimmer, H., & Mayringer, H. (2014). *Salzburger Lese-Screening für die Schulstufen 2-9 (SLS 2-9)*. Verlag Hans Huber.

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

8 Appendices

Appendix A

Figure 8
Overview of Combination of Test Halves per Time Point in the Total Sample.

Item Pool

	Test 1	Test 2	Test 3	Test 4
First Half	Items 1-10	Items 1-10	Items 1-10	Items 1-10
Second Half	Items 11-20	Items 11-20	Items 11-20	Items 11-20

Testing Procedure

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8
Group 1	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10
	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20
Group 2	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10
	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20
Group 3	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10
	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20
Group 4	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10
	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20
Group 5	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10
	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20
Group 6	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10
	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20
Group 7	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10
	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20
Group 8	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10	Items 1-10
	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20	Items 11-20

Appendix B

Table 5
Number of Observations Available for Convergent Status Validity Measures per Variable Pair.

	Reading Pre			Reading Post			TJ Dimensional			TJ Criterial		
Quop-L2	Word	Sentence	Text	Word	Sentence	Text	Word	Sentence	Text	Word	Sentence	Text
Time point 1	305	284	303	298	298	297	328	326	327	328	326	327
Time point 2	311	292	310	308	309	306	334	334	332	334	334	332
Time point 3	302	285	301	302	302	300	324	324	323	324	324	323
Time point 4	273	256	269	290	292	289	295	294	293	295	294	293
Time point 5	289	274	288	305	304	304	309	308	309	309	308	309
Time point 6	312	294	309	312	313	313	334	333	333	334	333	333
Time point 7	289	274	287	293	293	293	311	310	311	311	310	311
Time point 8	312	294	308	314	315	313	334	334	332	334	334	332

Note. Reading Pre = ELFE II Pretest. Reading Post = ELFE II Posttest. TJ = Teacher Judgments. A variable pair consists of a quop-L2 score and a status validity score. Values depicted here are based on quop-L2 total scores. A total score being available for an observation presumes available scores on word, sentence, and text level. Therefore, in relating a status validity measure to a quop-L2 score, the values depicted here are the minimal number of observations available for this case. For example, a correlation between quop-L2 word level scores and standardized reading pretest performance is based on 305 obersvations (row one, column one).

Table 6
Number of Observations Available for Discriminant Status Validity Measures per Variable Pair.

Quop-L2	Intelligence	Mathematics
Time point 1	302	309
Time point 2	310	315
Time point 3	302	306
Time point 4	271	278
Time point 5	287	292
Time point 6	310	315
Time point 7	291	296
Time point 8	311	316

Note. Intelligence measured with CFT 1-R, Mathematical Competence measured with DEMAT 1+. A variable pair consists of a quop-L2 score and a status validity score. Values depicted here are based on quop-L2 total scores. For example, a correlation between quop-L2 total scores and CFT 1-R is based on 302 obersvations (row one, column one).

Appendix C

Table 7
Means and Standard Deviations of quop-L2 Scores per Level and Time Point.

	Word		Sentence		Text	
Quop-L2	Mean	SD	Mean	SD	Mean	SD
Time point 1	56.59	12.79	56.64	13.55	40.29	13.04
Time point 2	58.82	12.63	58.77	13.59	43.42	13.61
Time point 3	61.25	13.01	61.82	13.29	46.16	13.86
Time point 4	63.11	12.6	63.12	12.78	47.93	13.86
Time point 5	64.86	12.36	65.96	12.9	50.32	14.06
Time point 6	65.67	12.36	67.02	12.52	51.38	13.96
Time point 7	66.96	12.38	68.29	12.25	53.43	13.75
Time point 8	67.88	12.01	68.31	12.3	53.93	13.57

Note. SD = Standard Deviation. Quop-L2 scores are CISRT scores ranging from 0 to 100.

Table 8
Means and Standard Deviations of Convergent Validity Measures per Level.

	Reading Pre		Reading Post		TJ Dimensional		TJ Criterial	
Level	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Word	29.02	9.72	38.31	11.16	4.67	1.42	19.37	11.40
Sentence	9.58	5.63	14.36	6.41	4.34	1.53	9.12	5.20
Text	5.63	4.19	9.45	5.24	4.02	1.56	6.61	4.44

Note. SD = Standard Deviation. Reading Pre = ELFE II Pretest. Reading Post = ELFE II Posttest. TJ = Teacher Judgments.

Table 9
Means and Standard Deviations of Discriminant Validity Measures.

	Intelligence		Mathematics	
Level	Mean	SD	Mean	SD
Total	28.28	8.38	27.57	6.22

Note. SD = Standard Deviation. Intelligence: CFT 1-R, Mathematical Competence: DEMAT 1+.

Appendix D

Table 10
Intercorrelations of All Variables.

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Quop-L2 W	1																
2. Quop-L2 S	.70	1															
3. Quop-L2 T	.60	.65	1														
4. Reading Pre W	.64	.68	.56	1													
5. Reading Pre S	.72	.76	.70	.82	1												
6. Reaing Pre T	.54	.63	.59	.64	.78	1											
7. Reading Post W	.62	.62	.47	.79	.77	.60	1										
8. Reading Post S	.67	.69	.55	.75	.84	.68	.85	1									
9. Reading Post T	.58	.64	.54	.61	.76	.74	.66	.77	1								
10. Intelligence	.31	.38	.34	.35	.45	.33	.38	.42	.44	1							
11. Mathematics	.38	.48	.33	.48	.46	.44	.43	.43	.40	.46	1						
12. TJ Dim W	.58	.64	.51	.63	.68	.58	.66	.69	.63	.48	.54	1					
13. TJ Dim S	.59	.63	.51	.63	.69	.58	.67	.70	.63	.46	.54	.95	1				
14. TJ Dim T	.59	.62	.50	.65	.70	.58	.70	.72	.65	.48	.56	.91	0.94	1			
15. TJ Criterial W	.38	.39	.36	.44	.47	.39	.58	.58	.49	.35	.32	.48	0.47	0.46	1		
16. TJ Criterail S	.48	.51	.46	.53	.61	.51	.61	.63	.54	.41	.38	.61	0.63	0.63	0.79	1	
17. TJ Criterial T	.45	.48	.42	.48	.57	.47	.53	.55	.48	.39	.37	.61	.63	.64	.66	.92	1

Note. Exemplary, quop-L2 scores from time point two were used here. W = Word Level, S = Sentence Level, T = Text Level. Reading Pre/Post = ELFE II Pretest/Posttest. Intelligence measured with CFT 1-R, Mathematical Competence with DEMAT 1+. TJ = Teacher Judgments. Dim = Dimensional. All correlations are significantly different from zero with $p < .001$.

Appendix E

Table 11
Quop-L2 Split-Half Reliabilites per Level and Time Point.

Quop-L2	Word	Sentence	Text
Time point 1	.82	.86	.65
Time point 2	.86	.89	.74
Time point 3	.85	.87	.74
Time point 4	.88	.90	.77
Time point 5	.86	.90	.77
Time point 6	.84	.89	.75
Time point 7	.89	.92	.83
Time point 8	.88	.91	.81

Note. All correlations are significantly different from zero with $p < .001$.

Table 12
Quop-L2 Retest Reliabilites per Level and Time Point.

Quop-L2	Word	Sentence	Text
Time points 1 - 2	.69	.68	.68
Time points 2 - 3	.69	.66	.66
Time points 3 - 4	.71	.67	.67
Time points 4 - 5	.66	.65	.65
Time points 5 - 6	.63	.64	.64
Time points 6 - 7	.65	.66	.66
Time points 7 - 8	.63	.65	.65

Note. All correlations are significantly different from zero with $p < .001$.

Appendix F

Table 13
Goodness-of-Fit Indices for Structural Validity Models per Quop-L2 Time Point.

Quop-L2	RMSEA	SRMR	CFI	TLI
Time point 1	.04	.02	.99	0.98
Time point 2	.04	.02	.99	0.98
Time point 3	.03	.02	.99	0.99
Time point 4	.03	.02	.99	0.99
Time point 5	.03	.02	.99	0.99
Time point 6	.03	.02	.99	0.99
Time point 7	.04	.02	.99	0.98
Time point 8	.03	.02	.99	0.99

Appendix G

Table 14
Convergent Status Validity Correlations (Standardized Reading Performance).

	Reading Pre			Reading Post		
Quop-L2	W	S	T	W	S	T
Time point 1	.65	.81	.52	.65	.71	.53
Time point 2	.64	.76	.59	.62	.69	.54
Time point 3	.65	.77	.54	.66	.73	.57
Time point 4	.62	.76	.63	.62	.76	.61
Time point 5	.62	.74	.56	.64	.75	.62
Time point 6	.60	.68	.61	.65	.73	.64
Time point 7	.64	.72	.61	.64	.74	.65
Time point 8	.60	.65	.61	.62	.71	.66

Note. Levels: W = Word, S = Sentence, T = Text. All correlations are significantly different from zero with $p < .001$.

Table 15
Convergent Status Validity Correlations (Teacher Judgments).

	TJ Dimensional			TJ Criterial		
Quop-L2	W	S	T	W	S	T
Time point 1	.68	.65	.42	.68	.65	.41
Time point 2	.63	.67	.54	.59	.66	.53
Time point 3	.67	.65	.55	.63	.63	.50
Time point 4	.57	.65	.58	.56	.62	.56
Time point 5	.63	.67	.60	.64	.68	.58
Time point 6	.64	.66	.65	.62	.66	.62
Time point 7	.61	.61	.64	.55	.62	.60
Time point 8	.63	.62	.62	.58	.64	.61

Note. Levels: W = Word, S = Sentence, T = Text. TJ = Teacher Judgments. All correlations are significantly different from zero with $p < .001$.

Table 16
Discriminant Status Validity Correlations.

Quop-L2	Intelligence	Mathematics
Time point 1	.43	.40
Time point 2	.39	.45
Time point 3	.38	.44
Time point 4	.32	.40
Time point 5	.40	.43
Time point 6	.40	.47
Time point 7	.37	.43
Time point 8	.36	.42

Note. Intelligence measured with CFT 1-R, Mathematical Competence with DEMAT 1+. All correlations are significantly different from zero with $p < .001$.

Appendix H

Table 17
Disattenuated Convergent Status Validity Correlations.

	Reading Pre			Reading Post		
Quop-L2	W	S	T	W	S	T
Time point 1	.73	.90	.69	.73	.79	.70
Time point 2	.70	.83	.74	.68	.75	.67
Time point 3	.71	.85	.67	.72	.80	.71
Time point 4	.67	.82	.77	.67	.82	.75
Time point 5	.68	.80	.68	.70	.81	.76
Time point 6	.66	.74	.76	.72	.79	.79
Time point 7	.69	.77	.72	.69	.79	.76
Time point 8	.65	.70	.73	.67	.76	.79

Note. Correlations from Appendix G corrected for reliability constraints. Levels: W = Word, S = Sentence, T = Text.

Table 18
Disattenuated Discriminant Status Validity Correlations.

Quop-L2	Intelligence	Mathematics
Time point 1	.46	.45
Time point 2	.42	.50
Time point 3	.41	.49
Time point 4	.34	.44
Time point 5	.43	.47
Time point 6	.43	.52
Time point 7	.39	.47
Time point 8	.38	.46

Note. Correlations from Appendix G corrected for reliability constraints. Intelligence measured with CFT 1-R, Mathematical Competence with DEMAT 1+.

Appendix I

Post-Hoc-Analysis: Re-estimation of the latent correlation between quop-L2 slope and standardized reading pre-post change on word level based on quop-L2 scores incorporating only items using frequent words. For this, the same analysis codes and models as presented before were used. Results showed a good fit of the combined LGM/LCM model on word level (*RMSEA* = .04, *SRMR* = .04, *CFI* = .99, *TLI* = 0.99). However, the correlation between quop-L2 slope and standardized reading pre-post change remained non-significant (r = .13, *SE* = 0.11, p = .26).

9 Declaration of Academic Integrity

I hereby confirm that this thesis on the subject "Static Score and Slope: A Comprehensive Validity Analysis of the Quop-L2 Reading Progress Assessment" is solely my own work and that I have used no sources or aids other than the ones stated. All passages in my thesis for which other sources, including electronic media, have been used, be it direct quotes or content references, have been acknowledged as such and the sources cited.

_____

(date and signature of student)

I agree to have my thesis checked in order to rule out potential similarities with other works and to have my thesis stored in a database for this purpose.

_____

(date and signature of student)