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

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

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

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

-----Wie Reliabilität ins Spiel bringen?-----

**Abstract**

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

**Introduction**

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

**Research Questions**

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

**Method**

Participants

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

Participants and Design

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

Instruments

* LPA

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

* Standardized Tests

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

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

* Teacher Ratings

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

Analytic Strategy

* LPA Scores

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

* Construct Validity

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

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

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

* Criterion Validity

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

**Results**

**Discussion**

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

**Fragen**

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

**Archiv**

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

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

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

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

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