Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01.06
Volume 14 (2022), No. 1, 122–146

Sven Anderson, Daniel Sommerhoff, Michael Schurig, Stefan Ufer, & Markus Gebhardt

Developing Learning Progress Monitoring Tests Using Difficulty-Generating Item Characteristics: An Example for Basic Arithmetic Operations in Primary Schools

Abstract

This study investigates difficulty-generating item characteristics (DGICs) in the context of basic arithmetic operations for numbers up to 100 to illustrate their use in item-generating systems for learning progress monitoring (LPM). The fundament of the item-generating system is based on three theory-based DGICs: arithmetic operation, the necessity of crossing 10, and the number of second-term digits. The Rasch model (RM) and the linear logistic test model (LLTM) were used to estimate and predict the DGICs. The results indicate that under the LLTM approach all of the three hypothesized DGICs were significant predictors of item difficulty. Furthermore, the DGICs explain with 20% a solid part of the variance of the RM's item parameters. The identification and verification of the DGICs under the LLTM approach provide important insights into how to address the challenges in the development of future LPM tests in mathematics.

Sven Anderson, M. A. (corresponding author), ORCID: 0000-0002-2323-8543 · Dr. Michael Schurig, ORCID: 0000-0002-7708-0593, Research in Inclusive Education, Faculty of Rehabilitation Sciences, TU Dortmund University, Emil-Figge-Straße 50, 44227 Dortmund, Germany
email: sven.anderson@tu-dortmund.de
michael.schurig@tu-dortmund.de

Prof. Dr. Daniel Sommerhoff, Department of Mathematics Education, IPN – Leibniz Institute for Science and Mathematics Education, 24098 Kiel, Germany, ORCID: 0000-0002-0559-7120
email: sommerhoff@leibniz-ipn.de

Prof. Dr. Stefan Ufer, Chair of Mathematics Education,
Department of Mathematics, Faculty of Mathematics, Computer Science and Statistics, LMU Munich, Theresienstr. 39, 80333 München, Germany, ORCID: 0000-0002-3187-3459
email: ufer@math.lmu.de

Prof. Dr. Markus Gebhardt, Learning Disability Pedagogy including Inclusive Pedagogy, Faculty of Human Sciences, University of Regensburg, Sedanstraße 1, 93055 Regensburg, Germany, ORCID: 0000-0002-9122-0556
email: markus.gebhardt@paedagogik.uni-regensburg.de

Keywords
learning progress monitoring (LPM), Rasch model (RM), linear logistic test model (LLTM), item-generating rules, elementary arithmetic

Entwicklung eines Tests zur Lernverlaufsdiagnostik mit schwierigkeitsgenerierenden Merkmalen: Ein Beispiel für die Diagnose grundlegender arithmetischer Fertigkeiten in der Grundschule

Zusammenfassung
Diese Studie untersucht den Einfluss schwierigkeitsgenerierender Merkmale für die Gestaltung von Items zur Lernverlaufsdiagnostik arithmetischer Basiskompetenzen im Zahlenraum bis 100. Das System zur Itemkonstruktion basiert dabei auf drei theoriegeleiteten schwierigkeitsgenerierenden Merkmalen: der verwendeten arithmetischen Operation, der Notwendigkeit des Zehnerübergangs, und der Stellenanzahl des zweiten Terms. Zur Schätzung und Vorhersage der Itemparameter wurden das Rasch-Modell (RM) und das linear-logistische Testmodell (LLTM) verwendet. Die Ergebnisse des LLTM-Ansatzes deuten darauf hin, dass alle drei vermuteten schwierigkeitsgenerierenden Merkmale signifikante Prädiktoren für die Itemschwierigkeit sind. Basierend auf den drei schwierigkeitsgenerierenden Merkmalen konnten 20% der Varianz der Itemschwierigkeitsparameter des RM erklärt werden. Diese Studie verdeutlicht, dass die Identifikation und Prüfung schwierigkeitsgenerierender Merkmale wichtige Erkenntnisse liefern, wie Herausforderungen bei der Entwicklung zukünftiger Tests zur Lernverlaufsdiagnostik in Mathematik berücksichtigt werden können.

Schlagworte
Lernverlaufsdiagnostik, Rasch-Modell (RM), Linear-logistisches Testmodell (LLTM), itemgenerierende Regeln, elementare Arithmetik

1. Introduction

Learning progress monitoring (LPM) represents an increasingly popular approach for assessing, monitoring, and visualizing students' individual learning development with short, high-frequency and easy-to-handle tests (e.g., Fuchs et al., 2019). With the results of LPM, teachers obtain insights into students' learning processes, which facilitates the early identification of emerging learning problems. Furthermore, LPM can support teachers in evaluating the success of implemented learning programs (e.g., Deno, 2003). Recent research also concludes that LPM in the digital form, managed as a computer-based or web-based tool, offers a more economical usability in school practice (Mühling et al., 2019; Souvignier, 2018). Previous

research on the systematic use of LPM also showed positive effects on students' learning achievements (e.g., Förster & Souvignier, 2015; Stecker et al., 2005).

Despite the usefulness of LPM, the development of adequate LPM instruments is challenging, as they have to fulfil a number of specific psychometric criteria. In particular, the valid measurement of learning processes requires a large number of parallel tests that are comparable in difficulty and dimensionality (e.g., Wilbert, 2014; Wilbert & Linnemann, 2011). Parallel, but non-identical tests are needed because of memory effects as well as practice effects that may otherwise confound the probability of solving items. Accordingly, homogeneity of test difficulty is a crucial prerequisite for LPM. Learning development can only be reasonably measured if it is ensured that the parallel tests are of equal difficulty. To address this requirement, a large number of test items with the same difficulty is required, which can then be distributed systematically across parallel test versions to avoid memory or practice effects. Knowledge about the characteristics that influence the difficulty of a test item can significantly support this item generation system. Therefore, LPM instruments need not only to fulfil the classical test quality criteria (e.g., validity, reliability), but also one-dimensionality, homogeneous test difficulty, test fairness, and sensitivity to change (Wilbert, 2014). To ensure validity when selecting suitable test items, the construction of LPM instruments relies on two approaches: the robust indicator approach and the curriculum sampling approach (Fuchs, 2004). Finding robust indicators involves choosing tasks that best represent the various subskills of a specific domain or correlate strongly with them. Using the curriculum sampling approach, typical tasks are selected that represent curricular requirements over a school year. For each LPM test, students receive tasks based on the learning goals of an entire school year.

Regardless of the procedure used for developing the LPM, a theory-based systematical item design and an empirical validation of LPM using item response theory (IRT) approaches are often lacking (e.g., Wilbert & Linnemann, 2011). So far, a framework for the systematic item design of LPM measures that explicitly links information about cognitive operations needed to solve an item and their correspondence in item characteristics with LPM development has not yet been established. To generate and select items for LPM, Wilbert (2014) proposes the use of linear extensions of the Rasch model (Rasch, 1980) such as the linear logistic test model (LLTM; Fischer, 1973; Fischer & Molenaar, 1995).

For mathematics, a large part of the research in LPM has focused on computation in primary schools (e.g., Foegen et al., 2007; Hartmann & Müller, 2014; Hosp et al., 2016; Tindal, 2013). In a review of LPM in the field of mathematics computation, Christ et al. (2008) emphasize the need for a framework to achieve higher and more consistent reliability and validity of measurements. Currently, there is still a gap of implementation and evaluation of a theory-based item-generating system for LPM. For the domain of elementary arithmetic, characteristics that influence item parameters are already considered in item selection for parallel LPM tests (e.g., Hartmann & Müller, 2014; Sikora & Voß, 2017). However, there is still a lack of systematic item design based on statistical evaluations. At present, there is

little literature available on the investigation of *difficulty-generating item characteristics* (DGICs) in the area of basic mathematical competencies that can be used for systematic LPM item construction (e.g., Balt et al., 2020). The present study addresses this research gap by focusing on the design and evaluation of an item generation system, applying theory-based modeling of item complexity using LLTM to analyze items exemplarily for an LPM addition and subtraction test for numbers up to 100. The present paper is therefore structured as follows: Firstly, we give a brief introduction of rule-based item design using LLTM. Secondly, based on this, we provide an overview of the characteristics affecting the difficulty of basic arithmetic operations. Building on this, we describe a rule-based item design of arithmetic basic skills for numbers up to 100. Furthermore, we present a study in which this rule-based item design is utilized in order to assess the difficulty of three item characteristics. Finally, we discuss as to what extent rule-based item design using LLTM is useful for the construction and evaluation of LPM instruments.

2. Rule-Based Item Generation With LLTM

Rule-based item design is a method from the research area of automatic item generation (AIG; e.g., Gierl & Haladyna, 2013; Irvine & Kyllonen, 2002) and rests upon the combination of findings from cognitive psychology and psychometric theory. AIG addresses the increased need for large pools of construct valid test items due to the ongoing trend towards computer- or web-based assessments. In rule-based item generation, difficulty characteristics are identified in advance and tested empirically. Knowledge about difficulty characteristics has several advantages: the increase of construct validity, the identification of templates that can be used for time-economical item design, or the opportunity of a content valid interpretation of test results (e.g., Arendasy et al., 2006; Kubinger, 2008). Rule-based item design is also particularly well suited for computer-supported algorithmic item generation (e.g., Geerlings et al., 2011) and of particular interest for the use of LPM in teaching as the tests require many items to measure students' learning development over a longer period (e.g., Wilbert, 2014). In a first step, the cognitive operations required to solve a specific task are identified. Subsequently, item models are generated that reflect the identified cognitive operations. This procedure enables the identification and evaluation of the considered cognitive operations necessary to solve a specific task, which are used to determine DGICs and, ultimately, the predicted probability of solving an item.

The design of items and instruments for LPM is particularly difficult as the use of LPM with multiple parallel tests over a period of time places specific demands on the underlying test theory. For this, IRT models are recommended (e.g., Anderson et al., 2011; Wilbert, 2014; Wilbert & Linnemann, 2011). IRT is an approach that determines the probability of solution for each item in a test, taking into account the ability of test takers based on their response behavior (Reise et al.,

2005). The response behavior (solving or not solving a task) then depends on both the individual characteristics of the person (person parameter) and the difficulty of the task (item parameter). A frequently used IRT model is the Rasch model (RM; Rasch, 1980).

The model equation of the RM for dichotomous data is:

$$P(x_{vi} = 1|\theta_v, \sigma_i) = \frac{\exp(\theta_v - \sigma_i)}{1 + \exp(\theta_v - \sigma_i)}$$

It is assumed that the probability P that person v solves item i depends on the person's ability (θ_v) as well as the item's difficulty (σ_i). Item parameter and person parameter are estimated in the RM for dichotomous data based on manifest response behavior.

An IRT model suitable for analyzing the difficulty of cognitive operations in rule-based items is the LLTM (Fischer, 1973; Fischer & Molenaar, 1995). LLTM allows the estimation of multiple basic parameters (DGICs) that are, based on a theoretical model and prior research, assumed to drive an item's difficulty. Knowledge about DGICs enables the prediction of the difficulty for newly developed items and, by considering different DGICs, items of various difficulty (within a certain range) can be easily generated. The LLTM breaks down the item parameter into a linear combination of specific hypothesized DGICs. Each DGIC involved in an item changes its difficulty. The total difficulty is the weighted sum of the DGICs.

In the LLTM the item's difficulty can be expressed as:

$$\sigma_i = \sum_{j=1}^{m} q_{ij}\eta_j + c$$

Here q_{ij} defines the weight of DGIC j on item i, for example 0 if DGIC j is not present in item j . η_j represents the difficulty parameter corresponding to DGIC j, which is independent of the specific item I. m equals the number of DGICs included in the model. c denotes a normalization constant. The weights of the DGICs are determined on the basis of theoretical hypotheses and prior research, so that the difficulty of the individual DGICs can be evaluated (Poinstingl, 2009). Based on the estimated difficulties of the DGICs, new items with a pre-determined difficulty (based on the LLTM) can be constructed, which is why this information can well be used for automatic item generation (e.g., Embretson & Kingston, 2018).

3. Specification of an Item-Generating System for Basic Arithmetic Operations for Numbers up to 100

3.1 Research on Students' Strategies and Difficulties in Addition and Subtraction

Addition and subtraction are the two algebraic operations introduced first in primary school, often in the first two grades. Although they can be interpreted as very basic, research evidence indicates that multi-digit addition and subtraction, for example 76 + 15 or 97 – 48, is challenging for many children, especially for low-performers and children with learning disabilities (Hickendorff et al., 2019). There are not only performance problems, but also differences in learning development. In their meta-analysis, Reilly et al. (2015) showed that there have been small but stable mean gender differences in favor of boys in mathematics over the past two decades.

Generally, children starting school are already able to solve simple addition and subtraction problems, for example by using counting strategies (e.g., Baroody, 1987), and results indicate that first graders' capabilities to solve such tasks are often underestimated (see further Baroody et al., 2006; Clarke et al., 2006). However, when expanding addition and subtraction to numbers up to 20 and then 100, these initial strategies often cannot be applied anymore or are time-consuming and effortful. Thus, enabling students to solve multi-digit addition and subtraction problems and providing efficient solution strategies for such tasks are among the most important goals in initial primary school mathematics education (e.g., Karp et al., 2011; National Council of Teachers of Mathematics, 2000, 2006). However, research has underlined that corresponding tasks, as for instance 14 + 23 or 90 – 23, vary considerably in their difficulty. This is attributed to students' varying use of different solution strategies, for example, number-based strategies or digit-based strategies (see Hickendorff et al., 2019, for an elaborate framework of correspondent strategies) as well as to different task characteristics (see also Daroczy et al., 2015, for a more general review on factors contributing to word problems' difficulty). For the construction of LPM items, the task characteristics are of particular interest, as they can easily be used for a rule-based item construction. In contrast, it is difficult to create items based on different solution strategies as students decide on their individual solution strategy during the problem-solving process, and even if tasks trigger a certain strategy for some learners, for example 39 + 47, they may not equally trigger this strategy for other learners in the same way. In this regard, research has repeatedly underlined students' heterogeneity, regarding both their individual capabilities for solving addition and subtraction tasks and (obviously related) their use of various solution strategies (see Baroody et al., 2006; Benz, 2005; Cooper et al., 1996; Verschaffel et al., 2007).

There are multiple characteristics of multi-digit addition and subtraction tasks up to 100 that can be interpreted in the sense of DGICs. Often mentioned, funda-

mental DGICs are (a) the *arithmetic operation* itself, that is addition and subtraction (e.g., Beishuizen, 1993; Cooper et al., 1996; Selter, 2001), (b) the necessity of *crossing 10,* that is the necessity to coordinate between ones and 10s as different places in the place-value notation during the operation as the addition/subtraction of the ones is not within the range of 0 to 9 (e.g., Beishuizen et al., 1997; Cooper et al., 1996; Fiori & Zuccheri, 2005), as well as (c) the *number of second-term digits* in the operation, that is either a one- or two-digit addend or subtrahend (e.g., Cooper et al., 1996). Multiple example items are given below.

A study by Benz (2005) revealed great differences in students' solution rates between tasks with different DGICs. First of all, her data revealed significant differences between addition and subtraction tasks, especially towards the end of the second grade and when presented without contextualization (i.e., no word problems). Based on her data, she suspects that differences between the solution rates for tasks with either operation are lower when informal strategies are applied, again highlighting the impact of different solution strategies on item difficulty. Moreover, her data also underlines that tasks with the necessity of crossing 10, for example 27 + 35 that requires the calculation of 7 + 5 = 12 and thus leads to a "crossing of 10" based on the addition of the ones, were less likely to be solved than those without that necessity. Finally, tasks including a second term with two digits were generally less likely to be solved for both arithmetic operations in comparison to tasks with one-digit second terms.

Beyond these three DGICs, research gives (mostly theoretical or qualitative) evidence of multiple other DGICs, implying, for example, that the addition and subtraction of multiples of 10s, for example, 30 + 20 or 50 – 30, is generally easier than tasks as for instance 32 + 24 or 57 – 33, which are comparable regarding the three DGICs pointed out above (i.e., no necessity of crossing 10 and including second terms with two digits) or that the addition resulting in a multiple of 10, for example, 53 + 7 or 56 – 6, is easier than other addition tasks with the necessity of crossing 10 even further, for example, 18 + 3. Moreover, the task 51 – 3 would be classified as being easier than 50 – 30 based on the DGIC that tasks with second terms with only one digit are easier than tasks with two second-term digits but classified as more difficult than 50 – 30 due to the necessity of crossing 10. These examples underline that concentrating on few, rather general DGICs can help to structure the difficulty of tasks and may allow for an easy item classification – and in the context of LPM easily comprehensible feedback for teachers and students. However, to accurately determine the difficulty of items, for example in research contexts, (a) more DGICs would have to be used and (b) in particular not only additively, but also including interactions. This would lead to a higher explained variance and thus a better classification of the items' difficulties, which would be favorable from a research perspective. However, giving teachers feedback on students' skills based on (e.g.) 10 DGICs and higher-level interactions appears unrealistic. Thus, using few central DGICs to explain a relevant portion of variance in order to give teachers a good first indication on how to support students seems indicated.

Overall, although DGICs can be used to approximate the difficulty of tasks and are useful in task construction, empirical evidence of the exact magnitude of the DGICs of addition and subtraction tasks is still lacking, as there are mostly only general tendencies regarding the higher importance of some DGICs over others.

3.2 Item Design of the Current Study

Based on the findings from prior mathematics education research, DGICs of addition and subtraction problems for numbers up to 100 can be identified. For the current study, three important characteristics were used to model the difficulty of the items: the *arithmetic operation* (addition versus subtraction; DGIC 1), the *necessity of crossing 10* (no crossing versus with crossing; DGIC 2), and the *number of second-term digits* (one-digit number versus two-digit numbers; DGIC 3). Therefore, the items used in this study consisted of mixed addition and subtraction problems including addition tasks with two one-digit numbers (e.g., 7 + 4), a one-digit and a two-digit number (e.g., 24 + 5), or the addition of 2 two-digit numbers (e.g., 47 + 26), and subtraction tasks with a one-digit or a two-digit number (e.g., 16 – 7) or 2 two-digit numbers (e.g., 78 – 49). For the tasks, the value of the individual digits varied between 0 and 9. For all problems, a fill-in-the-blank format was used, with the correct answer always being missing. Furthermore, there were tasks with and without crossing the tens barrier as well as tasks to the next 10.

Table 1 exemplifies the design matrix (see also Appendix Table A1) for the computation test with three exemplary items and the three dichotomously scored DGICs arithmetic operation (DGIC 1), necessity of crossing 10 (DGIC 2), and number of second-term digits (DGIC 3). The item number refers to the position of the item in the test. The 0s and 1s are the specified weights of the DGICs for each item. For the DGIC 1 (arithmetic operation), a weight of 0 represents addition, a weight of 1 represents subtraction. A weight of 0 for the DGIC 2 (necessity of crossing 10) means that crossing 10 is not necessary for solving the item, whereas with a weight of 1 it is. The items with a weight of 0 for the DGIC 3 (number of second-term digits) have a one-digit second term, a weight of 1 shows that the second term has two digits. For example, Item 5 (56 + 3 = 59) neither includes a subtraction nor the necessity of crossing 10 nor a two-digit subtrahend. Thus, all weights for this item are 0. In contrast, Item 7 (43 + 9 = 52) includes the crossing of a 10. Consequently, the weight matrix includes a 1 for the crossing of a 10 for this item, otherwise 0s.

Table 1: Exemplary Design Matrix for Items 5, 7, and 8

Item number	DGIC 1	DGIC 2	DGIC 3	Item
5	0	0	0	56 + 3 = 59
7	0	1	0	43 + 9 = 52
8	1	1	1	32 – 17 = 15

Note. DGIC = difficulty-generating item characteristic.

Based to the three DGICs of the design matrix in Table 1, a pool of 80 items was generated. The three DGICs were varied within the item design process so that all possible combinations were adequately represented in the item pool (with three DGICs, eight different combinations were considered; see the design matrix for the first 41 items included in the further analyses in the Appendix Table A1).

4. Research Questions

To test the construct validity of the model, the characteristics of the items in terms of DGICs and the statistical properties of the items were associated. The hypothetically assumed rule-based difficulty-generating item characteristics have to explain the RM's item difficulty parameters. In addition, there is the assumption that differences in gender will become apparent. We followed two central research questions:

Research Question 1: Do the three difficulty-generating item characteristics (DGICs) that were used to create the test items influence their difficulty?

In reference to Question 1, we hypothesized that the three identified DGICs have significant influence on the difficulty of the items as indicated by prior mathematics education research (e.g., Benz, 2005).

Research Question 2: How much variance is explained by the LLTM model with the three DGICs?

With regard to Question 2, we hypothesized that the three identified DGICs have a significant impact in variance explanation of the RM's item difficulty parameters. However, prior research has underlined the existence of multiple other DGICs and the importance of their interrelation for an exact estimation of item difficulty. Moreover, research has underlined that item characteristics are only one aspect influencing item difficulty and that students' solving strategies are also important for item difficulty. This led us to hypothesize that the variance explanation might be somewhere between 10% (i.e., clearly above 0) and 50% (i.e., still leaving much room for other DGICs, strategies, and other variables for variance explanation).

5. Method

5.1 Participants

The sample consisted of N = 591 students (M_{age} = 8.80, SD = 0.76; boys: 52.28%, girls: 47.72%; 9.14% of the students with special educational needs) in Grade 2 (n = 205; 34.69%) and 3 (n = 386; 65.31%) from 13 German primary schools in North Rhine-Westphalia (28 classes). All students who did not reach at least 5 items were excluded from the analysis as reasonable participation in the test could not be guaranteed and students might have been unwilling to participate. Overall, n = 10 participants (1.69%) were excluded.

5.2 Measures

Prior to the study, a first test with fixed order was created based on a pool of 80 items, which included all eight possible combinations of the three DGICs. The fixed order was chosen to safeguard that participants worked on each combination regularly. The test was implemented on the online platform Levumi (www.levumi.de; Gebhardt et al., 2016; Mühling et al., 2019). Between October 2019 and January 2020, the students were tested in their classrooms in groups of 10. To perform the test, each participating student used a tablet device. The test time was 5 minutes. At the beginning, the students received a brief technical instruction, an example item was solved, and the students were given the opportunity to ask the test supervisor questions. Students could start the test themselves by clicking on the start button. After the test time had expired, the test ended automatically. This way the test cannot be considered a power test due to the restricted time but is rather a long speed test.

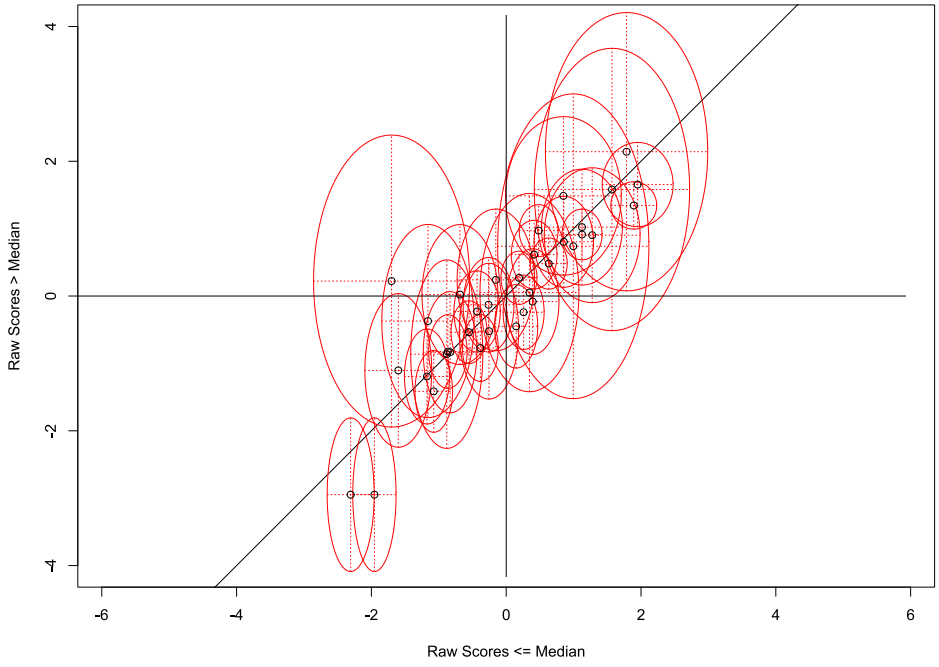
5.3 Statistical Analyses

The participants were not expected to solve nearly as many items as the test contained in the given 5-minute testing time as the high number of items was created to also account for exceptionally good students. Accordingly, most students did not answer all items. The mean number of items answered was 18.48 (SD = 9.59, min = 1, max = 67). For the analyses, only items that were answered by at least 20 students were used to ensure an adequate model estimation. Thus, 41 items were included in the analyses (see the descriptive statistics in the Appendix Table A2). RM is a prerequisite to conduct LLTM for item analysis and model comparisons. The following computations were all done with the package eRM (Mair & Hatzinger, 2007). Thus, a dichotomous RM was fitted. The occurrence of missings is calculated and pooled separately for each subgroup (by NA structure) in eRM

(Mair & Hatzinger, 2007). In our understanding, items that are answered more often should not be weighted more heavily that way. To evaluate the items of the test, the item fit was evaluated by infit and outfit statistics. Items with infit and outfit values below 0.5 and above 1.5 (Wright & Linacre, 1994) were excluded. This resulted in an additional item exclusion of five items, namely Items 32, 34, 36, 38, and 41 (see Appendix Table A2). Item 33 was also excluded due to inappropriate response patterns within subgroups, leaving 581 cases and 35 items. The reliability of the resulting weighted maximum likelihood estimation (WLE) person parameters reached .82.

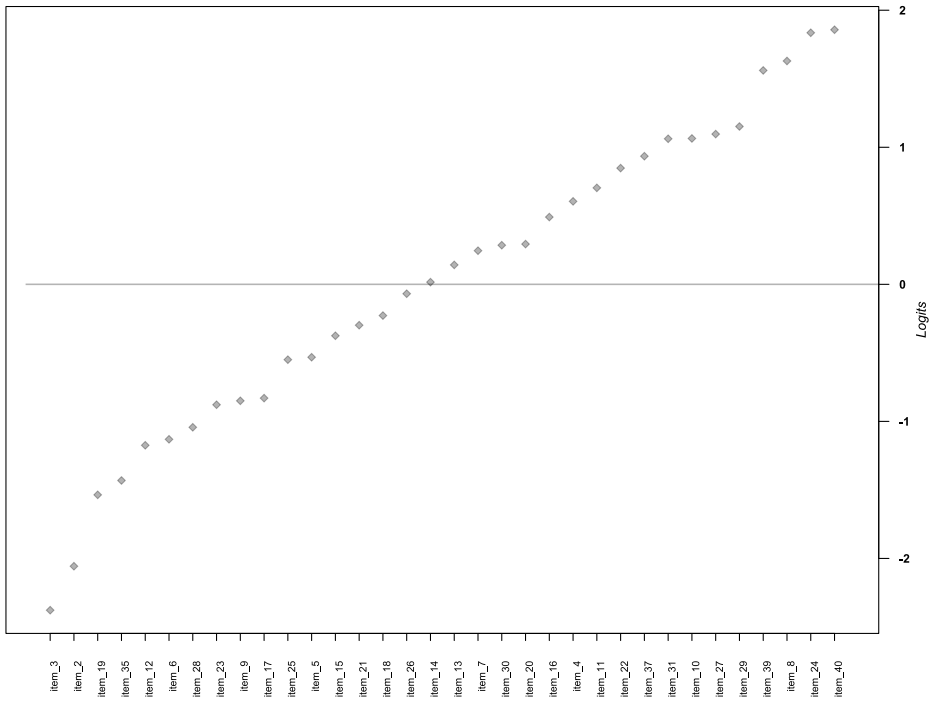
The package *eRM* employs a conditional maximum likelihood approach for parameter estimation. There are different ways of testing the appropriateness of an RM. In this paper, the likelihood ratio test (LRT; Andersen, 1973; Kubinger, 1989), the graphical model test (Rasch, 1980), and a Wald-type test (Glas & Verhelst, 1995) are applied. The LRT is a global test of model fit that checks the assumption of equality of the item parameters between subpopulations. A significant result were differences in the global item parameters between groups. An LRT with a median split resulted in a LRT-value of 32.24 (df = 34, p = .554), an LRT with a mean split in 44.80 (df = 34, p = .102), thus showing no significant differences in item parameters between these subgroups. In the graphical model test, the estimated item parameters of the median-split subgroups are compared against each other (Figure 1). The ellipses indicate the size of the confidence intervals.

Figure 1: Graphical Model Check of Items for Addition and Subtraction of Numbers up to 100



Though the confidence intervals vary in magnitude due to sparse data because of the item location in the test, nearly all item's ellipses are split by the bisecting angle, indicating equal item parameters between the groups. The Wald test specifies the analysis on the item level. Here, all items are analyzed individually. The Wald tests (split criteria: median; mean) showed significant differences in the difficulties of two items in *median*-split subgroups (Item 8 and Item 11) and two items (Item 3 and Item 14) in a *mean*-split subgroup. The items were kept because the infit and outfit values of the item fit statistics showed that they do not have to be deemed detrimental. The final item parameters are given in Figure 2 to illustrate the spread in difficulty. The item numbers in Figure 2 correspond to the sequence in the test. For example, Item 3 is the third item that the students worked on in the test. As additional analyses, differential item functioning (DIF; Holland & Wainer, 1993; see Figure A1 and Figure A2 in the Appendix) was tested for gender and grade. However, no gender effects (see Reilly et al., 2015) could be found in these items. DIF revealed only one item (Item 3) with a relatively high effect of gender, which however was still insignificant. DIF analyses for grade resulted in multiple non-linear significant effects, which, however, were expected based on prior results from mathematics education, which also showed non-linear relationships between second and third grade (e.g., Benz, 2005).

Figure 2: Item Difficulties of Items for Addition and Subtraction of Numbers up to 100



The LLTM was estimated stepwise. In a first step, a model (Model 1) was estimated with DGIC 1 (arithmetic operation). In a second step, another model (Model 2) was estimated that contained DGIC 2 (necessity of crossing 10) in addition to DGIC 1 (arithmetic operation). Finally, DGIC 3 (number of second-term digits) was used to estimate a third model (Model 3) that included all three DGICs identified in Section 3. Overall, the 35 items included in the analyses contained 17 items that showed the characteristic of DGIC 1 (arithmetic operation), 16 of DGIC 2 (necessity of crossing 10), and 17 of DGIC 3 (number of second-term digits). The stepwise estimation with three models was done to (a) illustrate the functionality of LLTMs, and (b) evaluate the additional benefit of adding the DGICs to the model. In the LLTM, the item parameters are regressed on the DGICs item parameters for each item. This contributes to unexplained variance based on the lower degrees of freedom (*df*; Poinstingl, 2009). These item parameters are not random but were fixed before the estimation. To determine the item difficulties, a design matrix with the weights of the three DGICs had to be established. Although the elements of such a design matrix can, in principle, be any positive number, also including fractions (Poinstingl, 2009), we assigned dichotomous values (see Table 1). As this paper sets out to illustrate the applicability of LLTM for LPMs, only additive components were addressed in this study for reasons of simplicity.

6. Results

6.1 Research Question 1: Influence of Identified DGICs on Item Difficulty

To determine the influence of the three identified DGICs and thus address Research Question 1, the DGICs are added stepwise. In a first run, DGIC 1 (arithmetic operation) is implemented as the only parameter (Model 1). In Model 2 and 3, the weight vectors of DGIC 2 (necessity of crossing 10) and DGIC 3 (number of second-term digits) are added to the design matrix. The additive procedure of the establishment of the item difficulties within the concurrent models is exemplified in Table 2.

As the Items 5 and 7 do not include the operation of subtraction, their item parameters are fixed to 0.00 in Model 1. Generally, items in LLTM that do not have a DGIC weight unequal 0 are fixed to difficulty 0.00 as only the presence of DGICs can affect an item difficulty. Here, the easiest combination of characteristics in terms of content is selected as the baseline so that all additions of characteristics that imply difficulty lead to item parameters > 0.

Table 2: Composition of Item Parameters in LLTM Models

Item number	Item	Design matrix assignment	Item difficulty			
			RM	Model 1	Model 2	Model 3
5	56 + 3	0,0,0	−0.53 [−0.77, −0.30]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
7	43 + 9	0,1,0	0.25 [0.02, 0.47]	0.00 [0.00, 0.00]	0.75 [0.65, 0.85]	0.94 [0.83, 1.05]
8	32 − 15	1,1,1	1.63 [1.40, 1.86]	0.52 [0.42, 0.61]	1.46 [1.30, 1.62]	2.54 [2.32, 2.75]

Note. Values in square brackets indicate the 95% confidence interval for each parameter estimation. The parameter estimates are given in logits.

As Item 8 is a subtraction item and thus holds characteristic DGIC 1, the item difficulty is fixed to the global parameter of Model 1, where only this parameter is estimated (see Table 3). In Model 3, when all difficulty parameters are introduced, the item difficulty of Item 5 is still fixed to 0.00, as its design matrix assignment is still 0,0,0. However, the difficulty of Item 8, which combines all three DGICs, is estimated to be 2.54, which reflects an additive combination of all three DGICs (Baghaei & Kubinger, 2015) given in Table 3, 0.75 + 0.94 + 0.84 ≈ 2.54. All items with the same hypothesized DGICs share the same difficulty, as difficulty is only estimated based on the DGICs. In comparison, the RM allows a free estimation of item difficulty for each item, thus allowing a better prediction of empirical data than the LLTM, however leading to less insights on why the items have a certain difficulty. The results of the comparison of the concurrent models (Table 3) show that the more of the three hypothesized DGICs are implemented in the models, the higher the quality of the model with regard to the likelihood, controlled by likelihood ratio tests, and the Akaike Information Criterion (AIC) as well as the Bayesian Information Criterion (BIC).

Table 3 highlights that the introduction of further DGICs causes a more desirable likelihood and better information criteria. Model 2 holds a significantly greater likelihood than Model 1 ($LR_{\text{Model 1 to Model 2}} = 219.19$, $df = 1$, $p < .001$) and Model 3 holds a significantly greater likelihood than Model 2 ($LR_{\text{Model 2 to Model 3}} = 260.5$, $df = 1$, $p < .001$). The RM still holds the most desirable likelihood ($LR_{\text{Model 3 to RM}} = 882.87$, $df = 31$, $p < .001$) and information criteria. This is, however, to be expected as each item's parameter is introduced as an independent random parameter in the model, while the parameters in the LLTMs are treated as fixed based on the linear combination of the DGICs. This means that the parameter estimation for the LLTMs is more parsimonious, which can be illustrated with an example: In this rather easy exemplary calculation, only 35 items are used, resulting in 34 parameters ($i - 1$) to be estimated in the RM. In the LLTMs, only 1 to 3 parameters are calculated (m).

Table 3: Log-Likelihood of the Models and Respective Difficulty Generating Item Characteristics

Model	CLL	Number of parameters	BIC	AIC	Estimate		
					DGIC 1 (η_1)	DGIC 2 (η_2)	DGIC 3 (η_3)
1	−4087.79	1	8181.94	8177.58	0.52 [0.42, 0.61]		
2	−3978.23	2	7969.19	7960.46	0.71 [0.60, 0.81]	0.75 [0.65, 0.85]	
3	−3847.98	3	7715.05	7701.96	0.75 [0.65, 0.86]	0.94 [0.83, 1.05]	0.84 [0.74, 0.95]
RM	−3406.54	34	7158.68	7023.38	–	–	–

Note. CLL = conditional log-likelihood; BIC = Bayesian Information Criterion; AIC = Akaike Information Criterion; DGIC = difficulty-generating item characteristic. Values in square brackets indicate the 95% confidence interval for each parameter estimation. The parameter estimates are given in logits.

Table 3 shows that all item parameters hold significant explanatory power, even when controlled for one another (identifiable for Model 3 by the confidence intervals of the DGICs in the square brackets). For DGIC 1 (arithmetic operation), subtraction is about three quarters of a logit more difficult than addition in Model 3 and for DGIC 2 (necessity of crossing 10), the crossing of a 10 is almost a full logit more difficult than not crossing a 10. For DGIC 3 (number of second-term digits), if there are two digits in the second term, it is about four fifth of a logit more difficult than if there is only a single digit in the second term.

6.2 Research Question 2: Impact of the Three DGICs in Variance Explanation of the RM's Item Difficulty Parameters

To answer Research Question 2 regarding the variance explanation of the three identified DGICs, correlations are calculated across the models. The appropriateness of the item parameters themselves can be assessed by the correlations of the item parameters between the models, especially when compared to the RM. The correlations are given in Table 4.

The item parameters of Model 1 and the item parameters of the RM reach an insignificant correlation of $r = .11$, which, however, is expected, since a model using only 1 *df* (thus allowing only two different values, 0.00 and 0.52, as item difficulty) is compared to a model with 34 *df*s. Yet, the correlation reaches a highly significant value of $r = .41$ for the item parameters of Model 3 and the RM. This can be interpreted as 20% explained variance in the item difficulties due to the introduced parameters. It underlines that the introduction of only three parameters (a) explains 20% of the item parameters in the Rasch model with 34 parameters, (b) can be deemed successful, and (c) confirms their importance for the general item difficul-

ty. However, 20% also shows that there is room for more DGICs, interactions, as well as other variables to account for unexplained variance, which would be valuable for further item generation.

Table 4: Means, Standard Deviations, and Correlations of the Item Parameters with Confidence Intervals

Model	*M*	*SD*	Model 1	Model 2	Model 3
Model 1	−0.25	0.26			
Model 2	−0.69	0.46	.59** [.35, .76]		
Model 3	−1.17	0.65	.44** [.15, .66]	.76** [.60, .87]	
RM	−0.00	1.08	.11 [−.21, .40]	.23 [−.08, .50]	.41** [.12, .64]

Note. Values in square brackets indicate the 95% confidence interval for each correlation.
*$p < .05$. **$p < .01$.

7. Discussion

In the current study, we developed an item-generating system for basic arithmetic operations for numbers up to 100 based on three DGICs. As expected, all three DGICs (arithmetic operation, necessity of crossing 10, number of second-term digits) significantly contributed to the prediction of item difficulty parameters in the LLTM. In addition, the three DGICs contribute substantially to the variance explanation of the RM's item difficulty parameters (Research Question 2). Previous assumptions about the difficulty characteristics of multi-digit addition and subtraction tasks up to 100 were thus confirmed. Gender-specific differences could not be identified. While we deliberately chose only three DGICs in the context of our research, as this relatively easy model allows easy-to-comprehend feedback for teachers, it would be highly valuable from a research perspective to consider additional DGICs and their interactions in order to achieve a higher variance explanation, thus allowing a better determination of item difficulty. Still, based on different solution strategies (e.g., counting strategies; see Hickendorff et al., 2019), the DGICs will anyhow only explain a certain share of the variance. Investigating students' currently used strategies has the potential to generate further information about the way the items are solved. This information can be used to derive additional variance explanation on the individual level. According to Wilbert (2014), the identification of DGICs is of interest in several aspects:

First, the information on the DGICs can contribute to the validation and further development of psychological theories. DGIC analyses can help to refine the initial theoretical assumptions about DGICs and can lead to a further development of appropriate theoretical models. In this context, findings about the relative diffi-

culties of DGICs can provide a basis for interpreting the importance of DGICs. Furthermore, more DIF analyses could show whether DGICs apply equally to different samples, making it possible to analyze which groups have particular difficulties in not only performing tasks generally, but also regarding specific hurdles in terms of DGICs. Especially in the field of inclusive education, this could be interesting for students with and without special educational needs or between groups of different special educational needs.

Second, knowledge about the DGICs is particularly useful for the construction of parallel test versions that can be used to monitor learning progress. For LPM, which requires frequent measurements over a long period of time, many different test items with known and comparable difficulty are needed. For this purpose, linear extensions of the RM such as LLTM are useful to identify DGICs, which can then be used as templates to generate items of the required difficulty without additional effort. With regard to previous research on LPM instruments in the field of computation, Christ et al. (2008) conclude that the use of more systematic sampling and item construction would improve the quality of LPM. To date, there has been little literature on the investigation on DGICs in the field of basic mathematical skills and for developing LPM (e.g., Balt et al., 2020; Ehlert et al., 2013). However, for the theory-based construction of test items for LPM, it is necessary to obtain valid information about which characteristics influence the difficulty of an item. This will support the development of knowledge about robust indicators. These enable the construction of LPM test procedures that can be used across classes and independently of graded curricula. Such curriculum-independent LPM tests are then also suitable for use in inclusive education, where often not all of the students are taught according to the same curriculum (Gebhardt et al., 2016).

Third, the identification of DGICs is also useful for the fine-grained analysis of students' learning development as it is possible to clearly highlight important aspects for further planning of individualized interventions. This allows the provision of formative feedback that has a concrete impact on the planning of future teacher interventions, rather than giving teachers only mean values that only allow for very general pedagogical conclusions. Previous research has shown that many teachers struggle with interpreting LPM results (e.g., Espin et al., 2017; Stecker, 2017). By identifying a small number of DGICs that serve as a framework for item construction, it is possible to make it easier for teachers to interpret results and thus establish a basis for designing appropriate interventions. This type of qualitative feedback can provide teachers with concise information about the domains in which a student is still struggling. For the domain of multi-digit addition and subtraction tasks up to 100 that we addressed in this study, more specific feedback could consist of informing the teacher that students have already confidently mastered addition tasks with two-digit summands but have not yet mastered crossing 10. With computer-based and web-based LPM tools, it is possible to quickly provide such qualitative feedback.

8. Conclusion

As noted at the beginning of this paper, the formulation and verification of a set of DGICs based on educational research results using LLTM enables the investigation of construct validity and the examination of the influence of the underlying DGICs on the difficulty of items. The results are therefore relevant for the establishment and further development of psychological theories, for test development, and educational practice. The explanation of roughly 20% of the variance in 34 item parameters by three DGICs alone is a solid result. Yet, of course, this also shows that more DGICs are needed to achieve stronger variance explanation of item difficulty. So far, we do not control for a certain type of strategy to solve the items but with the DGICs, we imply a particular type of strategy (calculating using structures of the decimal system). Differential profiles of item difficulties might exist when controlling for strategies.

In the present study, it is not possible to examine additional DGICs due to the systematically created item pool. In addition, it has not yet been possible to control adequately for a greater range of background characteristics (e.g., dyscalculia) that might have a potential influence on item difficulty. One important future question is the influence of sequence effects due to the limited processing time of LPM. The DGICs robustness against time was only tested in a limited fashion, too. By now, a fixed item order was evaluated. In the future, this has to be extended for randomized item orders. In future studies, the processing time of the items or the test has to be considered as an item characteristic or function of a difficulty generating characteristic. Rasch Poisson counts models seem to be a promising alternative (Baghaei & Doebler, 2019). Such models can also take into account the nested structure of the data which has not been considered in this study. The nested structure may also have an effect on the standard error (*SE*) of the item parameters which has not been accounted for now. Cognitive diagnostic modeling is another alternative for detailed modeling when deterministic information of the component matrices is available (Ravand & Robitzsch, 2015). In addition, the question should be investigated, whether the parameters causing difficulties primarily affect students in the lower grades and students with special educational needs.

In summary, the study emphasizes the usefulness of the LLTM for the identification and evaluation of difficulty-generating characteristics for an item-generating system for basic arithmetic operations. With the identified and verified three DGICs, it is possible to easily develop items of varying difficulty for parallel test versions, a basic prerequisite for the effective use of LTMs. The three DGICs provide a basis for computer-based LPM and thus enable an easy to use, practical, and easy to interpret application.

Acknowledgements

The current research is part of the project Dortmund Profile for Inclusion-Oriented Learning and Teacher Training – DoProfiL. DoProfiL is part of the *Qualitätsoffensive Lehrerbildung*, a joint initiative of the Federal Government and the Länder, which aims to improve the quality of teacher training. The program is funded by the Federal Ministry of Education and Research [Bundesministerium für Bildung und Forschung; Förderkennzeichen 01JA1930]. The authors are responsible for the content of this publication.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch Model. *Psychometrika*, *38*(1), 123–140. https://doi.org/10.1007/BF02291180

Anderson, D., Lai, C.-F., Alonzo, J., & Tindal, G. (2011). Examining a grade-level math CBM designed for persistently low-performing students. *Educational Assessment*, *16*(1), 15–34. https://doi.org/10.1080/10627197.2011.551084

Arendasy, M. E., Sommer, M., Gittler, G., & Hergovich, A. (2006). Automatic generation of quantitative reasoning items – A pilot study. *Journal of Individual Differences*, *27*(1), 2–14. https://doi.org/10.1027/1614-0001.27.1.2

Baghaei, P., & Doebler, P. (2019). Introduction to the Rasch Poisson counts model: An R tutorial. *Psychological Reports*, *122*(5), 1967–1994. https://doi.org/10.1177/0033294118797577

Baghaei, P., & Kubinger, K. D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research, and Evaluation, 20*, Article 1. https://doi.org/10.7275/8f33-hz58

Balt, M., Fritz, A., & Ehlert, A. (2020). Insights into first grade students' development of conceptual numerical understanding as drawn from progression-based assessments. *Frontiers in Education*, *5*, Article 80. https://doi.org/10.3389/feduc.2020.00080

Baroody, A. J. (1987). The development of counting strategies for single-digit addition. *Journal for Research in Mathematics Education*, *18*(2), 141–157. https://doi.org/10.2307/749248

Baroody, A. J., Iung Lai, M., & Mix, K. S. (2006). The development of young children's early number and operation sense and its implications for early childhood education. In B. Spodek & O. N. Saracho (Eds.), *Handbook of research on the education of young children* (2nd ed., pp. 187–221). Lawrence Erlbaum.

Beishuizen, M. (1993). Mental strategies and materials or models for addition and subtraction up to 100 in Dutch second grades. *Journal for Research in Mathematics Education*, *24*(4), 294–323. https://doi.org/10.2307/749464

Beishuizen, M., Van Putten, C. M., & Van Mulken, F. (1997). Mental arithmetic and strategy use with indirect number problems up to one hundred. *Learning and Instruction*, *7*(1), 87–106. https://doi.org/10.1016/S0959-4752(96)00012-6

Benz, C. (2005). *Erfolgsquoten, Rechenmethoden, Lösungswege und Fehler von Schülerinnen und Schülern bei Aufgaben zur Addition und Subtraktion im Zahlenraum bis 100* [Students' success rates, calculation methods, solutions and mistakes in addition and subtraction tasks in the range up to 100]. Franzbecker.

Christ, T. J., Scullin, S., Tolbize, A., & Jiban, C. L. (2008). Implications of recent research: Curriculum-based measurement of math computation. *Assessment for Effective Intervention*, *33*(4), 198–205. https://doi.org/10.1177/1534508407313480

Clarke, B., Cheeseman, J., & Clarke, D. (2006). The mathematical knowledge and understanding young children bring to school. *Mathematics Education Research Journal*, *18*(1), 78–102. https://doi.org/10.1007/BF03217430

Cooper, T. J., Heirdsfield, A., & Irons, C. J. (1996). Children's mental strategies for addition and subtraction word problems. In J. T. Mulligan & M. C. Mitchelmore (Eds.), *Children's number learning* (pp. 147–162). Australian Association of Mathematics Teachers and Mathematics Education Research Group of Australasia.

Daroczy, G., Wolska, M., Meurers, W. D., & Nuerk, H.-C. (2015). Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology*, *6*, Article 348. https://doi.org/10.3389/fpsyg.2015.00348

Deno, S. L. (2003). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention*, *28*(3–4), 3–12. https://doi.org/10.1177/073724770302800302

Ehlert, A., Fritz, A., Arndt, D., & Leutner, D. (2013). Arithmetische Basiskompetenzen von Schülerinnen und Schülern in den Klassen 5 bis 7 der Sekundarstufe [Basic arithmetic competencies of secondary school students from Grades 5 to 7]. *Journal für Mathematik-Didaktik*, *34*(2), 237–263. https://doi.org/10.1007/s13138-013-0055-0

Embretson, S. E., & Kingston, N. M. (2018). Automatic item generation: a more efficient process for developing mathematics achievement items? *Journal of Educational Measurement*, *55*(1), 112–131. https://doi.org/10.1111/jedm.12166

Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., & Rooij, M. de (2017). Data-based decision-making: Developing a method for capturing teachers' understanding of CBM graphs. *Learning Disabilities Research & Practice*, *32*(1), 8–21. https://doi.org/10.1111/ldrp.12123

Fiori, C., & Zuccheri, L. (2005). An experimental research on error patterns in written subtraction. *Educational Studies in Mathematics*, *60*(3), 323–331. https://doi.org/10.1007/s10649-005-7530-6

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models*. Springer. https://doi.org/10.1007/978-1-4612-4230-7

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

Förster, N., & Souvignier, E. (2015). Effects of providing teachers with information about their students' reading progress. *School Psychology Review*, *44*(1), 60–75. https://doi.org/10.17105/SPR44-1.60-75

Fuchs, L. S. (2004). The past, present and future of curriculum-based measurement research. *School Psychology Review*, *33*(2), 188–192. https://doi.org/10.1080/02796015.2004.12086241

Fuchs, L. S., Fuchs, D., Seethaler, P. M., & Zhu, N. (2019). Three frameworks for assessing responsiveness to instruction as a means of identifying mathematical learning disabilities. In A. Fritz, V. G. Haase, & P. Räsänen (Eds.), *International handbook of mathematical learning difficulties* (pp. 669–681). Springer. https://doi.org/10.1007/978-3-319-97148-3_39

Gebhardt, M., Diehl, K., & Mühling, A. (2016). Online-Lernverlaufsmessung für alle SchülerInnen in inklusiven Klassen. www.LEVUMI.de. [Online learning progress monitoring for all students in inclusive classes. www.LEVUMI.de]. *Zeitschrift für Heilpädagogik, 67*(10), 444–454.

Geerlings, H., Glas, C. A. W., & Van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, *76*(2), 337–359. https://doi.org/10.1007/S11336-011-9204-X

Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge. https://doi.org/10.4324/9780203803912

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 69–95). Springer. https://doi.org/10.1007/978-1-4612-4230-7_5

Hartmann, E., & Müller, C. M. (2014). *Lernfortschrittsdiagnostik Grundrechenarten. 120 Drei-Minuten-Tests für den inklusiven Mathematikunterricht – ZR bis 100: 1.-4. Klasse* [Learning progress monitoring: Basic arithmetic operations.120 three-minute-tests for inclusive mathematics teaching]. Persen.

Hickendorff, M., Torbeyns, J., & Verschaffel, L. (2019). Multi-digit addition, subtraction, multiplication, and division strategies. In A. Fritz, V. G. Haase, & P. Räsänen (Eds.), *International handbook of mathematical learning difficulties* (pp. 543–560). Springer. https://doi.org/10.1007/978-3-319-97148-3_32

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement* (2nd ed.). The Guilford Press.

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development*. Lawrence Erlbaum.

Karp, K., Caldwell, J., Zbiek, R. M., & Bay-Williams, J. (2011). *Developing essential understanding of addition and subtraction for teaching mathematics in Pre-K–Grade 2*. National Council of Teachers of Mathematics.

Kubinger, K. D. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie [Critical evaluation of latent trait theory]. In K. D. Kubinger (Ed.), *Moderne Testtheorie – Ein Abriß samt neuesten Beiträgen* (2nd ed., pp. 19–83). Psychologie Verlags Union.

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item-generating rules to measuring item administration effects. *Psychology Science*, *50*(3), 311–327.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1–20. https://doi.org/10.18637/jss.v020.i09

Mühling, A., Jungjohann, J., & Gebhardt, M. (2019). Progress monitoring in primary education using Levumi: A case study. In H. Lane, S. Zvacek, & J. Uhomoibhi (Eds.), *CSEDU 2019. Proceedings of the 11th International Conference on Computer Supported Education* (pp. 137–144). SCITEPRESS – Science and Technology Publications.

Müller, M. (2020). Item fit statistics for Rasch analysis: Can we trust them? *Journal of Statistical Distributions and Applications*, *7*, Article 5. https://doi.org/10.1186/s40488-020-00108-7

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*.

Poinstingl, H. (2009). The linear logistic test model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychology Science Quarterly*, *51*(2), 123–134.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research, and Evaluation, 20*(11), 1–12. https://doi.org/10.7275/5g6f-ak15

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement. A meta-analysis of National Assessment of Educational

Progress assessments. *Journal of Educational Psychology*, *107*(3), 645–662. https://doi.org/10.1037/edu0000012

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory. *Current Directions in Psychological Science*, *14*(2), 95–101. https://doi.org/10.1111/j.0963-7214.2005.00342.x

Selter, C. (2001). Addition and subtraction of three-digit numbers: German elementary children's success, methods and strategies. *Educational Studies in Mathematics*, *47*(2), 145–173. https://doi.org/10.1023/A:1014521221809

Sikora, S., & Voß, S. (2017). Konzeption und Güte curriculumbasierter Messverfahren zur Erfassung der arithmetischen Leistungsentwicklung in den Klassenstufen 3 und 4 [Conception and quality of curriculum-based measurements for the computation performance of primary school students in grade 3 and 4]. *Empirische Sonderpädagogik*, *9*(3), 236–257. https://doi.org/10.25656/01:15163

Souvignier, E. (2018). Computerbasierte Lernverlaufsdiagnostik [Computer-based learning progress assessment]. *Lernen und Lernstörungen*, *7*(4), 219–223. https://doi.org/10.1024/2235-0977/a000240

Stecker, P. M. (2017). Reflections on teachers' data-based decision making. *Learning Disabilities Research & Practice*, *32*(1), 71–72. https://doi.org/10.1111/ldrp.12128

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

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education,* Article 958530. https://doi.org/10.1155/2013/958530

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Verschaffel, L., Greer, B., & DeCorte, E. (2007). Whole number concepts and operations. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 557–628). Information Age Publishing.

Wilbert, J. (2014). Instrumente zur Lernverlaufsdiagnostik: Gütekriterien und Auswertungsherausforderungen [Tools for learning progress monitoring: Quality criteria and challenges with regard to interpretation]. In M. Hasselhorn, W. Schneider, & U. Trautwein (Eds.), *Tests und Trends: Vol. 12: Lernverlaufsdiagnostik* (pp. 281–308). Hogrefe.

Wilbert, J., & Linnemann, M. (2011). Kriterien zur Analyse eines Tests zur Lernverlaufsdiagnostik [Criteria for analyzing a test measuring learning progress]. *Empirische Sonderpädagogik*, *3*(3), 225–242.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370–371.

Appendix

Table A1: Design Matrix and Item Descriptives

Item number	DGIC 1	DGIC 2	DGIC 3	Item	*n*	*M*	*SD*
1	1	0	0	78 – 6 = 72	581	0.69	0.46
2	0	0	1	30 + 20 = 50	581	0.88	0.32
3	0	1	0	7 + 4 = 11	581	0.91	0.29
4	1	1	0	93 – 7 = 86	581	0.52	0.50
5	0	0	0	56 + 3 = 59	581	0.70	0.46
6	1	0	0	47 – 7 = 40	581	0.79	0.41
7	0	1	0	43 + 9 = 52	571	0.59	0.49
8	1	1	1	32 – 17 = 15	552	0.36	0.48
9	0	1	0	24 + 6 = 30	529	0.77	0.42
10	0	1	1	47 + 26 = 73	511	0.47	0.50
11	1	0	1	76 – 23 = 53	470	0.54	0.50
12	1	0	1	70 – 30 = 40	428	0.83	0.38
13	0	0	1	42 + 24 = 66	403	0.67	0.47
14	0	1	1	37 + 43 = 80	365	0.70	0.46
15	1	1	0	13 – 6 = 7	341	0.76	0.43
16	1	0	1	83 – 23 = 60	313	0.65	0.48
17	0	1	0	8 + 4 = 12	287	0.83	0.38
18	1	0	0	67 – 4 = 63	274	0.76	0.43
19	0	0	1	40 + 10 = 50	250	0.90	0.31
20	1	1	0	23 – 6 = 17	236	0.70	0.46
21	0	0	0	24 + 5 = 29	208	0.78	0.42
22	1	0	1	48 – 26 = 22	188	0.62	0.49
23	1	0	0	97 – 7 = 90	164	0.84	0.37
24	1	1	1	78 – 49 = 29	155	0.47	0.50
25	0	1	0	78 + 2 = 80	133	0.78	0.41
26	0	1	0	73 + 8 = 81	121	0.73	0.44
27	0	1	1	43 + 38 = 81	112	0.58	0.50
28	1	0	1	80 – 70 = 10	95	0.81	0.39
29	0	0	1	32 + 17 = 49	88	0.56	0.50
30	1	0	1	67 – 47 = 20	77	0.68	0.47
31	0	1	1	49 + 31 = 80	60	0.58	0.50
32	1	1	0	16 – 7 = 9	51	0.73	0.45
33	0	1	0	9 + 8 = 17	47	0.75	0.44
34	0	0	1	70 + 20 = 90	41	0.81	0.40
35	1	0	0	38 – 2 = 36	41	0.85	0.36
36	1	1	0	76 – 8 = 68	38	0.74	0.45
37	1	1	1	48 – 29 = 19	34	0.68	0.48
38	1	0	0	28 – 8 = 20	31	0.84	0.37
39	0	0	0	81 + 7 = 88	29	0.66	0.48
40	0	1	0	47 + 9 = 56	25	0.64	0.49
41	1	0	1	50 – 30 = 20	21	0.95	0.22

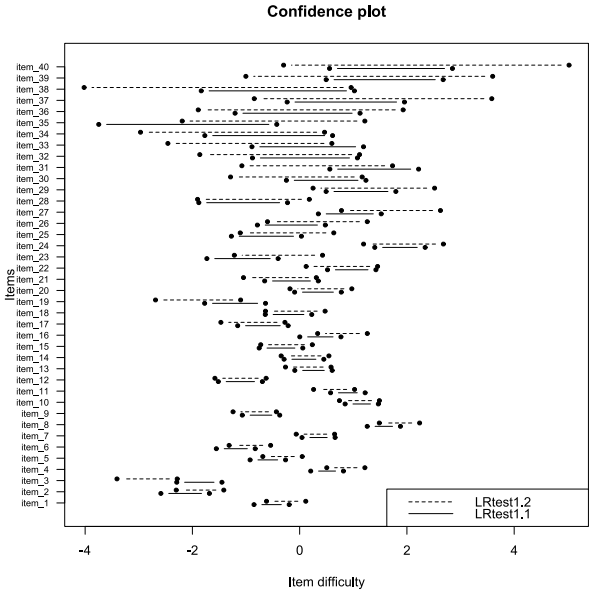
Note. DGIC = difficulty-generating item characteristic.

Table A2: Item Fit Statistics in the RM on Item Level

	Unconditional outfit	Unconditional infit	Conditional outfit		Conditional infit	
Item	MSQ	MSQ	MSQ	*SE*	MSQ	*SE*
1	0.80	0.87	0.74	0.09	0.85	0.05
2	0.95	0.90	0.81	0.19	0.95	0.09
3	1.29	1.08	0.88	0.22	0.98	0.11
4	1.04	1.03	0.89	0.06	0.93	0.04
5	1.12	1.13	0.92	0.09	1.00	0.05
6	0.80	0.89	0.69	0.12	0.85	0.06
7	0.98	1.03	0.91	0.06	0.93	0.04
8	0.83	0.91	0.79	0.07	0.86	0.04
9	0.89	1.02	0.80	0.11	0.91	0.05
10	0.91	0.93	0.89	0.06	0.91	0.04
11	1.06	1.08	0.84	0.06	0.89	0.04
12	1.12	1.04	0.85	0.12	0.90	0.06
13	0.83	0.91	0.76	0.07	0.82	0.04
14	0.61	0.74	0.59	0.07	0.68	0.04
15	0.93	1.00	0.76	0.09	0.88	0.05
16	0.95	0.92	0.71	0.06	0.77	0.04
17	0.94	1.17	0.85	0.13	0.98	0.06
18	0.94	0.85	0.751	0.07	0.81	0.04
19	1.15	1.17	0.71	0.16	0.89	0.07
20	0.78	0.86	0.81	0.07	0.88	0.04
21	0.85	1.05	0.79	0.08	0.89	0.04
22	1.09	1.01	0.92	0.06	0.91	0.04
23	0.87	0.79	0.71	0.11	0.86	0.05
24	0.81	0.93	0.91	0.06	0.95	0.04
25	1.21	1.16	0.85	0.07	0.91	0.04
26	1.13	0.92	0.90	0.09	0.91	0.05
27	0.87	0.91	0.80	0.06	0.87	0.04
28	0.82	0.97	0.97	0.10	1.01	0.91
29	0.74	0.90	1.07	0.06	1.11	0.00
30	0.89	0.84	0.93	0.06	0.96	0.34
31	1.35	1.10	1.20	0.06	1.18	0.04
32	1.84	1.20	0.93	0.07	0.95	0.17
33	1.19	1.18	1.38	0.08	1.32	0.04
34	2.04	1.07	1.47	0.08	1.34	0.04
35	0.66	0.94	0.80	0.07	0.90	0.04
36	0.43	0.71	1.26	0.06	1.15	0.04
37	0.88	1.13	1.39	0.08	1.28	0.04
38	4.71	1.34	2.08	0.06	1.60	0.04
39	0.96	0.90	2.17	0.06	1.66	0.04
40	1.47	1.08	1.56	0.06	1.45	0.04
41	0.05	0.27				

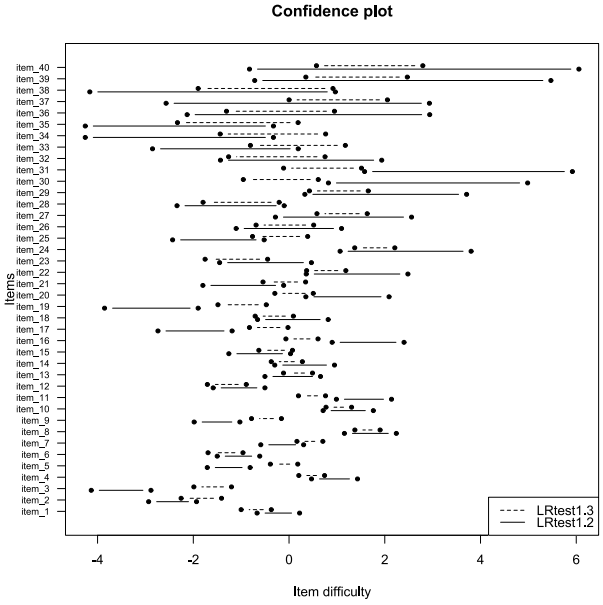
Note. MSQ = mean square residual; the unconditional fit statistics based upon the estimation of the item and person parameters; conditional fit statistics only rely on the item parameters (e.g., Müller, 2020). To estimate the conditional item fit statistics data without missingness is necessary. Therefore, single imputation was done with R package mice (Van Buuren & Groothuis-Oudshoorn, 2011).

Figure A1: Plot of the Confidence Intervals for the Item Parameters with Split Criterion Gender



Note. The solid line indicates the group of girls.

Figure A2: Plot of the Confidence Intervals for the Item Parameters with Split Criterion Grade



Note. The solid line indicates the group of second graders.