Pohle, Lara; Hosoya, Georg; Loftfield, Catharina; Jenßen, Lars

Indicators measuring preschool teachers' stimulation quality. Theoretical background and empirical testing

Zeitschrift für Pädagogik 65 (2019) 4, S. 525-541

Quellenangabe/ Reference:
Pohle, Lara; Hosoya, Georg; Loftfield, Catharina; Jenßen, Lars: Indicators measuring preschool teachers' stimulation quality. Theoretical background and empirical testing - In: Zeitschrift für Pädagogik 65 (2019) 4, S. 525-541 - URN: urn:nbn:de:0111-pedocs-239934 - DOI: 10.25656/01:23993

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-239934
https://doi.org/10.25656/01:23993

in Kooperation mit / in cooperation with:

BELTZ JUVENTA

http://www.juventa.de

D 7484

ZEITSCHRIFT FÜR PÄDAGOGIK

Heft 4 Juli/August 2019

■ *Thementeil*

Kompetenzerwerb zukünftiger LehrerInnen in der universitären Ausbildung

■ *Allgemeiner Teil*

Der Einfluss von Lehrkräfteverbänden in der Steuerung von Schulsystemen: Deutschland und Frankreich im Vergleich

Zur Situationsspezifität des pädagogischen Ethos: Eine empirische Studie im Bereich der betrieblichen Berufsbildung

Forschendes Lernen prüfen: Hochschuldidaktische Gedanken zu einer Theorie des Prüfens

BELTZ JUVENTA

ZEITSCHRIFT FÜR PÄDAGOGIK
Jahrgang 65 – Heft 4
Juli/August 2019

Inhaltsverzeichnis

Allgemeiner Teil

Dokumentation

Table of Contents

Topic: Prospective Teachers' Competence Development in Higher Education

Articles

Lara Pohle/Georg Hosoya/Catharina Loftfield/Lars Jenßen

Indicators Measuring Preschool Teachers’ Stimulation Quality

Theoretical background and empirical testing[1]

Abstract: This study focused on the theoretical background and empirical testing of quality indicators to assess preschool teachers’ stimulation quality in the domain *counting, magnitudes, numbers* in Germany. Theoretical assumptions, factor analysis based on polychoric correlation and internal consistency indicated that clarity of mathematical content, cognitive activation and constructive learner support allow for maximal distinction between preschool teachers’ stimulation quality. Results were drawn from the observation of 25 preschool teachers working in Berlin and Brandenburg. Existing instruments used to measure preschool teachers’ instructional quality mainly stem from the US-American context following a school-oriented approach.

Keywords: Process Quality, Stimulation Quality, Early Childhood Mathematics Education, Observational Assessment, Domain-Specificity

1. Introduction

Throughout the past few years, researchers have examined the relation between teaching quality and student achievement; the maximization of children’s learning outcome being a major concern of their studies (Blazar, 2015). Together with this, one focus of interest has been directed at teaching practices that are assumed to lead to children’s increase in learning. However, the identification of indicators promoting effective instruction remains a challenge in early childhood mathematics education (Connor et al., 2018). This might be the reason why some researchers “draw upon the fairly extensive research on mathematics teaching at older age levels to develop an understanding of what such teaching involves in the case of young children” (Ginsburg & Amit, 2008, p. 275). All aspects considered, the absence of indicators might explain the rather small number of studies that have examined early mathematics instructional quality (McGuire, Kinzie, Thunder & Berry, 2015). However, there are three observation instruments that have been applied in most of these studies: the Classroom Observation of Early Mathematics Environment and Teaching (COEMET; Sarama & Clements, 2007), the Classroom Assessment Scoring System (CLASS; Pianta, La Paro & Hamre, 2008)

1 This research was part of the Pro-KomMa project (Professionalization of Early Childhood Teacher Education: Convergent, Discriminant, and Prognostic Validation of the KomMa Models and Tests), funded by the German Federal Ministry of Education and Research (Grant No. 01PK15003A).

and the Early Childhood Environment Rating Scale (ECERS; Harms & Clifford, 1980). All of these instruments stem from the US-American context and follow a school-oriented approach. Therefore, their application in socio-pedagogically oriented German preschools is subject to some limitations. In Germany, early education is characterized by a variety of approaches such as play, situational learning and guided learning. These differences between these two countries make it at least necessary to reconsider the previous use of instruments because "factors that seem successful in some countries may not be appropriate elsewhere" (Panayiotou et al., 2014, p. 74).

In addition to the difficulty of country-specific teaching and learning philosophies, existing measures to assess teaching quality tend to differ regarding their underlying theories even if they are developed within the same educational context. While some instruments are characterized by normative aspects, others are guided by ideas of effective teaching. Additionally, some measures mainly consider subject-didactical aspects, while some relate to generic aspects of teaching, and others combine both approaches (Brunner, 2018). For example, the above-mentioned instruments to assess quality in preschools primarily consider global criteria regarding teacher-child interactions (CLASS) or even refer to structural aspects such as physical space and hygiene (ECERS). Here, the criticism concerning the lack of domain-specific factors (Lehrl, Kluczniok & Rossbach, 2016) becomes obvious. Although COEMET was designed to capture subject-specific aspects of teaching, it also includes general criteria such as environmental aspects. Not only the underlying theories, but also the scales of the instruments differ widely. For example, some instruments consist of single items while others comprise whole dimensions. Additionally, there are observation tools that are hierarchically organized, meaning that low scorings within one dimension can be compensated by higher ratings in another dimension (Brunner, 2018).

In their article, Kilday and Kinzie (2009) analyse a number of different observation instruments and conclude that there is a lack of measures to be applied in preschool. They also emphasize the measures' alignment with standards introduced by the National Council of Teachers of Mathematics (NCTM), which therefore reduces their relevance in educational systems other than the US-American or Canadian one. Likewise, a report published by the European Commission, the Education, Audiovisual and Culture Executive Agency (EACEA), Eurydice and Eurostat (2014) refers to the gap of reliable and valid measures. Nevertheless, there is fundamental interest in assessing preschool teachers' professional competence in mathematics.

The preceding remarks support the need to create new observation measures in order to assess teacher behavior, which represents one facet of professional competence and is expected to promote children's learning (Early et al., 2007). One attempt to develop an instrument capturing domain-specific process quality in German preschools has been made by Isele (2014). However, its application is not restricted to preschools; instead, it is used as "Transitionsinstrument" (Isele, 2014, p. 11) and is geared to the conditions found at primary schools, too. This gap highlights the need to develop measures, which, on the one hand, are domain-specific, and, on the other hand, meet the requirements of being applied in socio-pedagogically oriented preschools. Therefore, a scale of subject-specific single items has been developed and tested to assess preschool teachers' performance in mathematics as one specific facet of their professional competence. The process of item construction followed an inclusive approach (Brunner, 2018), hence subject-specific as well as subject-unspecific items were operationalized domain-specifically. This article highlights the theoretical grounding and empirical testing of items in order to set a basis for "Understanding the factors that make some teachers more effective than others [which] is vital to achieving and supporting high-quality instruction" (Guarino, Dieterle, Bargagliotti & Mason, 2013, p. 164).

2. Conceptualizing Stimulation Quality

Recently, attention has been placed on preschool teachers' professional competence and associated models, which, among other facets of competence, focus on teachers' performance (Blömeke, Gustafsson & Shavelson, 2015; Gasteiger & Benz, 2016; Nentwig-Gesemann, Fröhlich-Gildhoff & Pietsch, 2011; Weinert, 2001). According to Blömeke et al., performance can be defined as "observed behavior in a particular real-world situation" (Blömeke et al., 2015, p. 11) which is further narrowed down by adding the aspect of domain-specificity. With regard to preschool teachers, Gasteiger and Benz take a step beyond in specifying "pedagogical and didactical action" as the skill to "design opportunities to learn mathematics or (…) [to] use situations spontaneously for mathematical learning by asking adequate questions or choosing appropriate learning material" (Gasteiger & Benz, 2018, p. 112). By doing so, the authors respond to the differences that exist between formal and informal learning settings and, consequently, point to the different demands made on preschool teachers when compared to schoolteachers (Gasteiger & Benz, 2016). Resulting from this, a preschool teacher's performance can be identified as concrete practices of teaching in any learning situation. As indicated above, these learning situations can be characterized by guided instruction as well as play or the use of spontaneously occurring situations. Although the latter seem to be the main features of early childhood education in socio-pedagogically oriented countries like Germany, planned activities or direct instruction are also valued to provide learning opportunities for young children (Gasteiger & Benz, 2018).

Along with this, Anders, Roßbach and Kuger (2016) report a vast number of factors that affect children's development and are therefore subject to empirical examination. One prominent factor within the scope of early childhood education deals with the effects of quality as it is considered to be a key towards children's mathematics learning (McGuire et al., 2015). Against this background, particularly process quality (Anders et al., 2012) and instruction (Blazar, 2015; Mashburn et al., 2008) are regarded to be important.

Drawing on different findings from research, Anders, Grosse, Rossbach, Ebert and Weinert define process quality as "refer[ing] to global characteristics such as warm climate […] as well as domain-specific stimulation in educational areas" (Anders, Grosse, Rossbach, Ebert & Weinert, 2013, p. 195). On a general level, Pianta et al. add "proximal-level interactions and transactions among teachers, children, and materials" (Pianta et al., 2005, p. 145) to be decisive characteristics of process quality.

We generally share these understandings of process quality. However, we focus on teachers and their actions because they are recognized as a strong determinant for children's learning (Blazar, 2015; Mashburn et al., 2008) and can be regarded as a key contributor to achieve educational stimulation. Following well-received theories (Baumert et al., 2010; Pianta & Hamre, 2009), we add aspects of domain-specificity as its absence is often criticized (Lehrl et al, 2016; Schlesinger & Jentsch, 2016) and because general indicators are not sufficient to grasp quality (Anders et al., 2012). For example, especially the aspect of mathematical correctness has been previously neglected (Brunner, 2018); however, it is vital to teaching math in preschool (Gasteiger & Benz, 2018). In addition, we consider consequences arising from the special demands made on preschool teachers when they have to incorporate formal and informal aspects of learning (Gasteiger & Benz, 2018). A brief description of all dimensions describing the construct that focuses on preschool teachers' actions, *stimulation quality,* is given below:

Clarity of mathematical content: This dimension encompasses domain-specific aspects and math-related peculiarities that are more or less exclusive to the central idea of numbers, magnitudes and counting. Following Helmke's (2009) concept *clarity of content,* the focus here lies on the correct depiction of the subject matter (Gasteiger & Benz, 2018; Ufer, Heinze & Lipowsky, 2015), which is central to develop mathematical concepts (Kinzie et al., 2014), and on maintaining content-related coherence (Brunner, 2018). It is assumed that subject-didactic aspects like math-specific language (Purpura & Logan, 2015), accuracy of content (Blazar, 2015), domain-specific strategies (Gasteiger & Benz, 2018) and the relation between magnitudes and numbers (Krajewski & Schneider, 2009) play an important role in the context of fostering number skills in young children. All in all, this category is characterized by subject-didactic aspects and content-related coherence.

Cognitive activation: Being concerned with higher-order thinking, this facet represents the integration of knowledge and aims to involve children in "demanding processes of problem-solving and understanding" (Hugener et al., 2009, p. 68). Studies have shown that cognitive activation and children's learning outcome in math are related (Lipowsky et al., 2009). Here, teaching behaviors that potentially engage learners into higher-order thinking are put into focus. Drawing from research, indicators like challenging questions, linkage to existing knowledge (Lipowsky et al., 2009) or references to everyday life in order to build on mathematical literacy (cf. Brunner, 2018), as well as using different modes of representation (Bruner, 1966) are operationalized by providing definitions and examples related to the domain of numbers, magnitudes, and counting.

Constructive learner support: In general, this dimension focuses on practices used to establish a climate that promotes children's learning. In addition to other aspects, this category includes constructive feedback and a positive relationship between teachers and learners (Praetorius, Pauli, Reusser, Rakoczy & Klieme, 2014), which has effects on children's learning (Lipowsky et al., 2009). Additionally, we subsume aspects of classroom organization into this category, as classroom management counts as "prerequisite for […] an orderly atmosphere in which to engage in content-related activities" (Lipowsky et al., 2009, p. 530). Although it might be regarded as a more general category, knowledge about the subject matter is needed in order to be able to provide learner support and, for example, adequately react to possible misconceptions (Gasteiger & Benz, 2018). In talking about numbers, magnitudes, and counting it would be useful to be familiar with the development of preschool quantity-number competencies (Krajewski & Schneider, 2009). Furthermore, the remaining items cannot be rated without specifically considering the mathematical content because they are operationalized with respect to the idea of numbers, magnitudes, and counting and include respective descriptions and examples. This specification also applies to the category cognitive activation.

Since process quality is mainly examined with the help of observational measures (Kuger, Kluczniok, Kaplan & Rossbach, 2015), our study draws on observational assessment, too. Despite all restrictions linked to classroom observations (Schlesinger & Jentsch, 2016), researchers reached consensus proclaiming this method as a valid source of information (Lerkkanen et al., 2012).

3. Method

After conceptualizing the construct from a theoretical point of view, further steps were undertaken to develop the scale of the *Messinstrument zur Erfassung der mathematischen Anregungsqualität im Bereich numerischer Basiskompetenzen* (MiA-Num) to be applied in early childhood mathematics education. This multi-step approach covered item generation, expert review and testing.

3.1 Item Development

Drawing on the conceptualization of stimulation quality, indicators capturing the sub-dimensions were formulated and further operationalized into observable items in order to distinguish between teachers and assess their performance. All items were constructed in compliance with the guidelines of item formulation (Rost, 2004). Due to economic reasons and for the purpose of gaining differentiated information about the subjects, a three-point ordinal scale *(does not apply, does partly apply, does fully apply)* was chosen. In order to avoid rater bias arising from subjective influences, each category is composed of a set of descriptions and examples that may be found in the teachers' performance. A rating manual contains all detailed information. After discussing each item in our team, several informal field observations served to identify rating problems resulting in first item adaptions. Table 1 shows an excerpt of the rating manual and gives

Item	The teacher links different modes of representation to approach mathematical ideas (enactive, iconic, symbolic).
○	In his or her actions, the teacher considers only one mode of representation in order to approach mathematical issues. For example, the teacher exclusively draws on the symbolic (or iconic, or enactive) mode of representation by asking questions like: "What is more? Five or seven?" (symbolic).
◑	In his or her actions, the teacher considers at least two modes of representation without linking them. For example, at the beginning of the learning sequence, the teacher says a number from one to ten and asks the children to point at the corresponding number on the wall (symbolic). Later, the children are asked to organise picture cards starting with the one showing the smallest number of sweets (iconic). (Here, two modes of representation are used. However, they are not linked and do not refer to the same mathematical idea.)
●	In his or her actions, the teacher considers at least two modes of representation by linking them in order to approach a mathematical issue. For example, the children are asked to determine the number of marbles in glasses in order to find out which glass contains *more (less, the most)* marbles (enactive). This action is repeated several times with different amounts. At the end of the learning sequence, the teacher shows pictures of animals and asks questions like: "Are there more elephants or more monkeys?" (iconic).

Tab. 1: Item Example

an example of the category *cognitive activation.* Similar to this, all remaining items are related to the idea of numbers, magnitudes, and counting and include corresponding examples. All items can be assessed through direct observation.

3.2 Expert Review

As validity regarding content constitutes a major facet of test construction (AERA, APA & NCME, 2014; Messick, 1989), an expert review was carried out in order to check how the items represent the construct. Following the procedure introduced by Jenßen, Dunekacke, and Blömeke (2015), nine external experts with research or practical experience in early mathematics education were presented with 24 items. The experts had to review the items regarding content, the items' potential to differentiate as well as the construct's overall representationon a four-point scale (ranging from 1 = not at all, to 4 = entirely). In order to answer all questions reliably, the experts were provided with theoretical background information about the construct *stimulation quality.*

In general, the experts' reviews indicate valid conclusions regarding content: The total set of items was rated to be a good representation of the construct ($M = 3.0$) and most indicators were said to explain differences in stimulation quality ($M = 3.3$). Only a minority of items had to be eliminated, whilst some were revised and the majority was accepted (see Tab. 2).

Dimensions	n	Items eliminated	Items added	Items revised	Items accepted
Clarity of content	5	0	0	3	2
Cognitive activation	11	4	1	1	6
Constructive learner support	8	0	0	1	7
Aggregate	24	4	1	5	15

Note. n = number of items

Tab. 2: Results of the Expert Review

3.3 Empirical Testing

The final scale was tested on *n* = 25 preschool teachers working in Berlin and Brandenburg who finished their job training within the past five years and were between 22 and 54 years old (*M* = 33.45; *SD* = 9.36). The majority of participants completed their training in a vocational school (72.7%) and only a minority (18.2%) graduated from university of applied sciences. All teachers were female.

Procedures, observation training, and administration
Two months before the observations took place, all teachers were given instructions to conduct a learning sequence, lasting 10 to 15 minutes in length dealing with *counting, magnitudes, numbers.* Content standardization allows for higher comparability and domain-specific insights (Schlesinger & Jentsch, 2016).

To ensure interrater reliability, all raters (*n* = 10) participated in a systematic observation training. Training videos were shown and ratings were discussed. At the end of the training, all raters passed a final test rating three videos independently. The intraclass correlation coefficient (ICC) was used to calculate interrater reliability, which showed good results (ICC = .966, two-way mixed, consistency).

To collect data, one or two trained raters conducted a non-participatory, standardized observation in all preschool settings. In order to adjust to the complex demands of direct observation, the raters took field notes serving as the basis for their ratings. If there were two raters in one setting, a consensus rating was carried out: Diverging ratings were discussed and a final rating was agreed on. While the consensus ratings (n = 19) were used for empirical testing, the independent ratings (*n* = 19) were the basis to recheck interrater reliability. The rest of the ratings (*n* = 6) were conducted by a single rater and used for empirical testing only.

Data analysis
First, interrater reliability was checked using the sum score that was assigned to the subjects by the individual raters. Due to different pairings of raters, we calculated the ICC on the basis of a random intercept only model.

Second, exploratory factor analysis (EFA) was conducted to assess the internal structure of the item scale using Maximum Likelihood (ML). Here, polychoric correlation was taken as the basis. Polychoric correlation is a suitable method to estimate correlation between two continuous, normally distributed variables if categorized versions of those variables exist (Gadermann, Guhn & Zumbo, 2012; Wirtz & Caspar, 2002). One can also say that polychoric correlation links metric variables and its manifest versions through thresholds: When the metric value crosses a threshold, the corresponding categorical value is assigned (Olsson, 1979). Apart from the aspect of the above-mentioned data structure, Holgado-Tello, Chacón-Moscoso, Barbero-García and Vila-Abad (2010) name further reasons why to choose polychoric correlation over Pearson correlation when carrying out factor analysis for ordinal data. Assuming that stimulation quality is an empirical distinct skill reflected in three manifestation categories, a one-factor model was applied. Although theory suggested this one-factor structure, EFA was preferred over CFA in order to identify unknown correlations.

Third, in order to make assumptions about the reliability of the item set, we calculated the ordinal alpha coefficient (Gadermann et al., 2012).

All calculations were conducted using the statistical programme R and its packages psych, lme4 and GPArotation.

4. Results

4.1 Descriptive Statistics

The length of observation varied between 10 and 45 minutes, while the average length was 19.76 minutes (SD = 9.13). The number of children taking part in the learning sequence varied between two and 14 averaging at 7.88 (SD = 2.99). Table 3 shows the descriptive statistics of the item scale. As can be seen, most of the ratings revealed acceptable variability with scores occupying the complete range of the scale; however, this did not apply for all items. The mean ratings for the majority of items were above average (between 1.0 and 2.0). Nevertheless, the sum score for each preschool teacher was calculated across all items, also showing satisfying variability with scores between 14 and 35 points and a mean of 26 points (SD = 4.33). Yet, the maximum score (42 points) was not reached.

Variables	Min	Max	*M*	*SD*
Content-specific correctness	0	2	1.76	.60
Mathematical language	0	2	1.44	.58
Clarity of content	0	2	1.60	.58
Linking numbers and magnitudes	1	2	1.28	.46
Acquisition of strategies	0	2	1.20	.58
Developmental-psychological aspects	0	2	1.60	.71
Linking modes of representation	0	2	1.48	.77
Building on children's mathematical actions and comments	0	2	.91	.70
Independent problem-solving	0	2	1.36	.76
Mathematical questions	1	2	1.16	.37
Instructive approach	0	2	.40	.82
Linking mathematics to everyday life	0	2	.44	.65
Stabilization of learning content	0	2	1.16	.69
Eliciting co-construction	0	2	.40	.58
Use of differentiation strategies	0	2	.76	.93
Positive climate	0	2	1.76	.60
Reinforcement	0	2	1.36	.70
Time to think	1	2	1.60	.50
Constructive feedback	0	2	1.21	.88
Effective use of learning time	0	2	1.48	.71
Mental under-/overload	0	2	1.84	.55

Tab. 3: Item Descripitive Statistics

4.2 Interrater Reliability

The ICC based on the random intercept only model shows that 60.26% of the variance go back to differences between preschool teachers and only 39.74% are ascribed to within-group differences.

4.3 Polychoric Correlation, Factor Analysis, and Ordinal Alpha

Computing polychoric correlation revealed estimation problems for five items, because not all categories were used and therefore either the two thresholds of the respective items did not differ from each other or no value could be assigned to the first threshold. Due to the occurrence of these estimation problems that might be traced back to the small sample size, the items in question were excluded from further analyses. Additionally, recoded items were excluded, as there might have been rating problems, which lead to estimation problems or implausible correlations.

Finally, the remaining items were used to conduct a series of factor analyses. Factor analysis revealed items with negative factor loadings meaning that preschool teachers who score high on the latent construct, more specifically, who show high stimulation quality in their performance, score low in respect of the items “linking mathematics to everyday life” and “eliciting content-specific co-construction”. As the overall assumption was made that teachers showing high stimulation quality are supposed to score high in all items, items where this theoretical assumption did not apply were excluded for further analyses. Moreover, items with a low factor loading (< .3) were excluded from interpretation (Kline, 1994).

Subsequently, another factor analysis was carried out showing satisfying factor loadings for all remaining items (see Tab. 4). Just like the assumption drawn from theory, the analysis suggested a one-factor solution, ultimately leading to seven remaining variables empirically defining the underlying construct that provides maximally distinct information about the sample at hand, accounting for 25% of the variance (eigenvalue = 1.78). Checking internal consistency for the final set of items revealed good results with ordinal Alpha = .70 (Gadermann et al., 2012). However, the findings should be checked on another sample.

Considering each preschool teacher’s sum score across the remaining items, again, shows satisfying variability with scores between 2 and 14 points and a mean of 9.52 points ($SD = 2.74$). Here, 20% of teachers scored 7 points or below, while 56% scored between 8 and 11 points and 24% scored 12 points or above. The maximum score, (14 points), was reached by two preschool teachers. Both scales, the original and the revised one, show significant correlation with $r = .90$ ($p < .05$).

Variables	Loadings	Communalities	Error variance
Mathematical language	0.50	0.25	0.75
Clarity of content	0.56	0.31	0.69
Acquisition of strategies	0.59	0.35	0.65
Linking modes of representation	0.62	0.38	0.62
Stabilization of learning content	0.38	0.14	0.86
Constructive feedback	0.39	0.15	0.85
Effective use of learning time	0.44	0.19	0.81

Tab. 4: Factor Analysis

5. Discussion

Following the request to develop and test observational measures in order to depict further details about preschool quality, the current research provides information about the instrument as such, and about the teachers of the sample. After finalizing the first phase of empirical testing, our research gives rise to undertake further steps in order to make valid, reliable and differentiated assumptions about preschool teachers' stimulation quality and to ensure quality of the instrument.

All items were developed in ways that facilitate their application in socio-pedagogically oriented countries. Moreover, their operationalization concentrates on the domain of numbers and includes clear examples as well as definitions that allow for easier scoring procedures. All in all, the aim to create an instrument that is easily applicable by trained observers was a decisive principle leading the construction process. Referring to this, interrater reliability indicated good results. The ICC pointed to a smaller amount of variance resulting from differences within teachers compared to differences between teachers. This shows that the raters recognized differences in performance and mainly agreed in their ratings. However, it must be said that the random intercept only model is not the proposed way to compute interrater reliability. Therefore, future studies should ensure fixed rater pairings in order to be able to apply techniques as suggested by literature (Wirtz & Caspar, 2002).

Due to economic reasons, a small number of ratings, ($n = 6$), were conducted by only one rater. This circumstance might have affected the reliability of the single ratings, which gives another reason to consider rater pairings as indicated above. Nevertheless, all raters were provided with a rating manual and participated in a systematic observation training where they had to pass a final test. This form of standardisation can be regarded as the basis for reliable assessments (AERA et al., 2014).

Polychoric correlation revealed estimation problems occurring from the omission of certain rating categories. The reason for this could be ascribed to the non-occurrence of certain situations meaning that the raters did not have the chance to observe a particular behavior. An explanation for this might be related to the instruction that was given to the teachers in the weeks before the observation, which might be responsible for the one-sidedness of the ratings because some of the items rather correspond to an instructional approach. Although the teachers were asked to create a setting that pictures their daily routine, it might have lead to the consequence that the teachers themselves chose a rather instructional approach instead of considering aspects of play or a situational approach. For example, the lowest category of the item "mathematical questions" could only be selected if the teachers asked no question at all. Considering the instruction as well as the artificiality of the situation, this seems to be unlikely. However, in another situation or when following another approach, circumstances could be different, which makes this kind of operationalization still reasonable. By collecting further data in an extended sample and in settings without previous instruction, the reasonability of operationalization should be examined.

To further ensure the quality of the item set, EFA was conducted to get a glimpse of the underlying structure and of the relationship between items, well aware of the fact that it is usually applied with larger sample sizes and thus criticisable. A one-factor structure was applied and items that did not meet this expectation in terms of negative or low factor loadings were excluded. Although two items indicate rather low factor loadings (< .4), they at least tend to load on the construct and were preserved due to theoretical reasons. Existing studies show similar difficulties, which is why further research examining the factorial structure is suggested (Soukakou, 2012). From a content-related point of view, the factorial structure supports the idea that items theoretically linked to clarity of content, cognitive activation and constructive learning support best describe and discriminate between preschool teachers' stimulation quality, which further fosters the idea of an empirical distinct construct reflected in three manifestation categories. This finding is supported by the internal consistency of all remaining items (ordinal Alpha = .70), although the content of the underlying construct may be underrepresented. However, correlation with the original scale proved reasonable results ($r = .90$) leading to the assumption that also a reduced scale allows to draw conclusions from the construct. Nevertheless, the original set of items should be used in replication studies to examine operationalization problems and factorial structure with larger samples. From a content-related point of view, the question arises whether there might be different factors representing the construct of stimulation quality. Due to the possibility of selecting different teaching approaches, it might be conceivable that one factor depicts stimulation quality in rather situational settings while another factor depicts it in instructional settings. This would also explain why some items show positive loadings while others indicate negative loadings in the one-factor model applied in our empirical examination. For example, the items "Mathematical questions" and "Stabilisation of learning content" might suggest a rather instructional approach while items like "Building on children's mathematical actions and comments", "Linking mathematics to everyday life", and "Eliciting co-construction" might be more characteristic of a situational approach. However, due to the above-mentioned constraints imposed by providing the teachers with an instruction, the latter items are less likely to receive high ratings, which, in the case of our one-factor model, might have led to their elimination. There might even be a third factor taking account of those teachers focusing on general aspects of learning which is expressed in items like "Use of differentiation strategies" and "Mental under-/overload". Future studies have to consider that different approaches demand different actions; otherwise, the results may be falsified. Referring to the study at hand, the factor structure and, based on that, the elimination of items should be reconsidered. Otherwise, the remaining items could possibly only point to stimulation quality in terms of instructional settings.

Talking about the challenges it must be said that a great deal of the sub-dimensions and indicators that are assumed to form stimulation quality are geared to school research. Although it might be a chance to profit from well-established ideas originating from other contexts, theory might not be fully transferable as, for example, requirements imposed on preschool teachers differ from those imposed on schoolteachers (Gasteiger & Benz, 2016). Therefore, the different teaching approaches and their consequences for the teachers' actions should be explicitly taken into account, especially when analyzing the results of empirical examination. Furthermore, operationalization of theory is not consistent across studies which is a common phenomenon found in research practice (Schlesinger & Jentsch, 2016), and, additionally, a consensus regarding quality factors does not even exist (Stipek & Chiatovich, 2017). Consequently, a comparison with results from other studies might be difficult. Nevertheless, the attempt made can be a good starting point to approach the construct *stimulation quality* in contexts that differ from school-oriented early childhood education systems. From a structural point of view, the sample size is criticisable and does not allow for drawing substantial conclusions. Replication of the results would be necessary and convergent validity should be tested. This might also contribute to approximate the problem of inconsistency concerning quality factors. Alongside of construct validity, concurrent validity should play a role in future studies. Here, the item set is correlated with other validated measures like tests capturing professional knowledge in the field of mathematics.

6. Conclusion

The goal of this study was to develop items that assess performance as one specific facet of preschool teachers' professional competence. Theoretical and well-received assumptions were the basis for item development. A systematic expert review that was carried out confirmed the theoretical grounding of stimulation quality together with its related items. Empirical testing subsequently revealed estimation problems that might go back to structural aspects such as sample size or content-related aspects. Therefore, replication with a larger sample is advisable and essential to check further aspects of validity. Nevertheless, this study gives valuable hints for future or replication studies and contributes to closing research gaps concerned with effective teaching and quality in preschool. Due to the lack of math-specific early childhood education studies throughout Germany, the development of refined measures to adequately capture domain-specific quality in socio-pedagogically oriented early education systems was the driving force of this research. By addressing the request to close this research gap, important implications for future studies can be drawn.

References

AERA, APA & NCME = American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing.* Washington D.C.: AERA.

Anders, Y., Grosse, C., Rossbach, H.-G., Ebert, S., & Weinert, S. (2013). Preschool and Primary School Influences on the Development of Children's Early Numeracy Skills between the Ages of 3 and 7 Years in Germany. *School Effectiveness and School Improvement, 24*(2), 195–211.

Anders, Y., Rossbach, H.-G., Weinert, S., Ebert, S., Kuger, S., Lehrl, S., & Von Maurice, J. (2012). Home and Preschool Learning Environments and their Relations to the Development of Early Numeracy Skills. *Early Childhood Research Quarterly, 27*(2), 231–244.

Anders, Y., Roßbach, H.-G., & Kuger, S. (2016). Early Childhood Learning Experiences. In S. Kuger, E. Klieme, N. Jude & D. Kaplan (Eds.), *Assessing Contexts of Learning. An international perspective* (pp. 179–208). Cham: Springer.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., & Tsai, Y.-M. (2010). Teachers' Mathematical Knowledge, Cognitive Activation in the Classroom, and Student Progress. *American Educational Research Journal, 47*(1), 133–180.

Blazar, D. (2015). Effective Teaching in Elementary Mathematics. Identifying classroom practices that support student achievement. *Economics of Education Review,* 48(2015), 16–29.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. (2015). Beyond Dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie 223*(1), 3–13.

Bruner, J. S. (1966). *Toward a Theory of Instruction.* Cambridge, MA: Bilknap.

Brunner, E. (2018). Qualität von Mathematikunterricht: Eine Frage der Perspektive. *JMD* 39(2), 257–284.

Connor, C. M., Mazzocco, M. M. M., Kurz, T., Crowe, E. C., Tighe, E. L., Wood, T. S., & Morrison, F. J. (2018): Using Assessment to Individualize Early Mathematics Instruction. *Journal of School Psychology,* 66(2018), 97–113.

Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R. H., Bryant, D., Cai, K., Clifford, R. M., Ebanks, C., & Griffin, J. A. (2007). Teachers' Education, Classroom Quality, and Young Children's Academic Skills: Results from seven studies of preschool programs. *Child Development, 78*(2), 558–580.

European Commission, EACEA, Eurydice, & Eurostat (2014). *Key Data on Early Childhood Education and Care in Europe. 2014 Edition.* Eurydice and Eurostat Report. Luxembourg: Publications Office of the European Union.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating Ordinal Reliability for Likert-type and Ordinal Item Response Data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation, 17*(3), 1–13.

Gasteiger, H., & Benz, C. (2016). Mathematikdidaktische Kompetenz von Fachkräften im Elementarbereich – ein theoriebasiertes Kompetenzmodell. *JMD,* 37(2), 263–287.

Gasteiger, H., & Benz, C. (2018). Enhancing and Analyzing Kindergarten Teachers' Professional Knowledge for Early Mathematics Education. *Journal of Mathematical Behavior,* 51, 109–117.

Ginsburg, H. P., & Amit, M. (2008): What is Teaching Mathematics to Young Children? A theoretical perspective and case study. *Journal of Applied Developmental Psychology, 29*(4), 274–285.

Guarino, C., Dieterle, S. G., Bargagliotti, A. E., & Mason, W. M. (2013). What Can We Learn About Effective Early Mathematics Teaching? A framework for estimating causal effects using longitudinal survey data. *Journal of Research on Educational Effectiveness, 6*(2), 164–198.

Harms, T., & Clifford, R. (1980). *Early Childhood Environment Rating Scale.* New York: Teachers College Press.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität.* Seelze: Kallmeyer.

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus Pearson Correlations in Exploratory and Confirmatory Factor Analysis of Ordinal Variables. *Quality & Quantity,* 44(1), 153–166.

Hugener, I., Pauli, C., Reusser, K., Lipowsky, F., Rakoczy, K., & Klieme, E. (2009). Teaching Patterns and Learning Quality in Swiss and German Mathematics Lessons. *Learning and Instruction 19*(1), 66–78.

Isele, P. (2014). *Entwicklung und Validierung eines Beobachtungsverfahrens zur Erfassung von förderrelevanter Prozessqualität beim Übergang vom Elementar- in den Primarbereich* (Doctoral dissertation). Dortmund: Technische Universität Dortmund.

Jenßen, L., Dunekacke, S., & Blömeke, S. (2015). Qualitätssicherung in der Kompetenzforschung. Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis. *Zeitschrift für Pädagogik,* Beiheft 61, 11–31.

Kilday, C., & Kinzie, M. (2009). An Analysis of Instruments that Measure the Quality of Mathematics Teaching in Early Childhood. *Early Childhood Education Journal, 36*(4), 365–372.

Kinzie, M.B., Whittaker, J.V., Williford, A.P., DeCoster, J., McGuire, P., Lee, Y., & Kilday, C.R. (2014). MyTeachingPartner-Math/Science pre-kindergarten Curricula and Teacher Supports. Associations with children's mathematics and science learning. *Early Childhood Research Quarterly, 29*(4), 586–599.

Kline, P. (1994). *An Easy Guide to Factor Analysis.* London: Routledge.

Krajewski, K., & Schneider, W. (2009). Early Development of Quantity to Number-word Linkage as Precursor of Mathematical School Achievement and Mathematical Difficulties: Findings from a four-year longitudinal study. *Learning and Instruction, 19*(6), 513–526.

Kuger, S., Kluczniok, K., Kaplan, D., & Rossbach, H.-G. (2015). Stability and Patterns of Classroom Quality in German Early Childhood Education and Care. *School Effectiveness and School Improvement, 27*(3), 418–440.

Lehrl, S., Kluczniok, K., & Rossbach, H.-G. (2016). Longer-term Associations of Preschool Education. The predictive role of preschool quality for the development of mathematical skills through elementary school. *Early Childhood Research Quarterly,* 36(3), 475–488.

Lerkkanen, M.-K., Kiuru, N., Pakarinen, E., Viljaranta, J., Poikkeus, A.-M., & Rasku-Puttonen, H., Siekkinen, M., & Nurmi, J. (2012). The Role of Teaching Practices in the Development of Children's Interest in Reading and Mathematics in Kindergarten. *Contemporary Educational Psychology, 37*(4), 266–279.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of Geometry Instruction and its Short-term Impact on Students' Understanding of the Pythagorean Theorem. *Learning and Instruction, 19*(6), 527–537.

Mashburn, A.J., Pianta, R.C., Hamre, B.K., Downer, J.T., Barbarin, O.A., Bryant, D., Burchinal, M., Early, D.M., & Howes, C. (2008). Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills. *Child Development, 79*(3), 732–749.

McGuire, P.R., Kinzie, M., Thunder, K., & Berry, R. (2015). Methods of Analysis and Overall Mathematics Teaching Quality in At-risk Prekindergarten Classrooms. *Early education and development, 27*(1), 89–109.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–104). New York: American Council on Education/Macmillan.

Nentwig-Gesemann, I., Fröhlich-Gildhoff, K., & Pietsch, S. (2011). Kompetenzentwicklung von FrühpädagogInnen in Aus- und Weiterbildung. *Frühe Bildung,* (2011), 22–30.

Olsson, U. (1979). Maximum Likelihood Estimation of the Polychoric Correlation Coefficient. *Psychometrika, 44*(4), 443–460.

Panayiotou, A., Kyriakids, L., Creemers, B.P.M., McMahon, L., Vanlaar, G., Pfeifer, M. Rekalidou, G., & Bren, M. (2014). Teacher Behavior and Student Outcomes: Results of a European study. *Educational Assessment, Evaluation and Accountability, 26*(1), 73–93.

Pianta, R.C., & Hamre, B.K. (2009). Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity. *Educational Researcher, 38*(2), 109–119.

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of Pre-Kindergarten Programs, Classrooms, and Teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science, 9*(3), 144–159.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *The Classroom Assessment Scoring System (CLASS).* Baltimore: Pail H. Brookes Publishing.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One Lesson is All You Need? Stability of instructional quality across lessons. *Learning and Instruction, 31*(2014), 2–12.

Purpura, D. J., & Logan, J. A. R. (2015). The Nonlinear Relations of the Approximate Number System and Mathematical Language to Early Mathematics Development. *Developmental psychology, 51*(12), 1717–1724.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2nd ed.). Bern: Huber.

Sarama, J., & Clements, D. H. (2007). *Manual for Classroom Observation (COEMET).* Version 3. Unpublished version.

Schlesinger, L., & Jentsch, A. (2016). Theoretical and Methodological Challenges in Measuring Instructional Quality in Mathematics Education Using Classroom Oberservations. *The International Journal on Mathematics Education,* 48(1), 29–40.

Soukakou, E. P. (2012). Measuring Quality in Inclusive Preschool Classrooms: Development and validation of the Inclusive Classroom Profile (ICP). *Early Childhood Research Quarterly, 27*(3), 478–488.

Stipek, D., & Chiatovich, T. (2017). The Effect of Instructional Quality on Low- and High-Performing Students. *Psychology in the Schools, 54*(8), 773–791.

Ufer, S., Heinze, A., & Lipowsky, F. (2015). Unterrichtsmethoden und Instruktionsstrategien. In R. Bruder, L. Hefendahl-Hebeker, B. Schmidt-Thieme & H.-G. Weigand (Eds.), *Handbuch der Mathematikdidaktik* (pp. 411–434). Berlin: Springer Spektrum.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – Eine umstrittene Selbstverständlichkeit. In Franz E. Weinert (Ed.), *Leistungsmessungen in Schulen* (pp. 17–31). Weinheim et al.: Beltz.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* Göttingen: Hogrefe.

Zusammenfassung: Diese Studie analysiert die theoretische Herleitung und empirische Testung von Qualitätsindikatoren zur Erfassung der Anregungsqualität frühpädagogischer Fachkräfte im Bereich ‚Zählen, Mengen, Zahlen' in Deutschland. Theoretische Annahmen, Faktorenanalyse basierend auf polychorischer Korrelation und interne Konsistenz zeigen, dass mathematische Klarheit, kognitive Aktivierung und konstruktive Lernunterstützung maximal zwischen der Anregungsqualität frühpädagogischer Fachkräfte unterscheiden. Die Ergebnisse beruhen auf der Beobachtung von 25 in Berlin und Brandenburg arbeitenden Fachkräften. Existierende Instrumente stammen vorrangig aus dem US-amerikanischen Raum und folgen einem schulorientierten Ansatz.

Schlagworte: Prozessqualität, Anregungsqualität, mathematische Frühförderung, Beobachtungsverfahren, Domänenspezifizität

Address of the Authors

M. Ed. Lara Pohle, Humboldt-Universität zu Berlin,
Institut für Erziehungswissenschaften,
Unter den Linden 6, Office Location: Friedrichstraße 194–199,
10117 Berlin, Deutschland
E-Mail: lara.pohle@hu-berlin.de

Dr. Georg Hosoya, Freie Universität Berlin,
FB Erziehungswissenschaften und Psychologie,
Kaiserswerther Str. 16–18, Office Location: Habelschwerdter Allee 45,
14195 Berlin, Deutschland
Mail: georg.hosoya@fu-berlin.de

M. A. Catharina Loftfield, Humboldt-Universität zu Berlin,
Institut für Erziehungswissenschaften,
Unter den Linden 6, Office Location: Geschwister-Scholl-Str. 7,
10117 Berlin, Deutschland
Mail: catharina.ma@hu-berlin.de

Dr. Lars Jenßen, Humboldt-Universität zu Berlin,
Institut für Erziehungswissenschaften,
Unter den Linden 6, Office Location: Friedrichstraße 194–199,
10117 Berlin, Deutschland
Mail: lars.jenssen@hu-berlin.de