Annemarie Fritz, Lars Balzer, Roelien Herholdt, Lara Ragpot & Antje Ehlert

A mathematics competence test for Grade 1 children migrates from Germany to South Africa

Abstract

This article presents the translation and adaptation process of a mathematics test for the acquisition of key mathematical (arithmetic) concepts by children from four to eight years of age. The origin of this test was in Germany, whence it was sourced by researchers at the University of Johannesburg. A conceptual model of hierarchical mathematics competence development forms the theoretical foundation of the test. This notion of hierarchical competence was tested in a one-dimensional Rasch analysis, which confirmed the hierarchical structure of the test with five levels of ability. In the translation process, it was imperative to ascertain whether the items of the translation had retained the conceptual content of the original test and had been allocated to the same conceptual levels as in the original test. In a number of pilot studies with a total of 1 600 South African children, we focused on the items that had been allocated to a different level, aiming to find out whether this was the result of translation errors. In analyses of different samples, discussing and reflecting on the model fit, and especially on items that did not fit well, 'misfitting' items could mostly be attributed to translation difficulties and differences in the children's strategies, and not to a generally altered model. The final model was established after the rephrasing of critical items. This model has already been tested with 500 additional South African children. Results are presented and discussed, with the focus on the Sesotho test results.

Keywords: MARKO-D test, Rasch model; conceptual development; mathematical concepts; assessment; South Africa; kindergarten, first grade; numeracy; arithmetic

Annemarie Fritz, University of Duisburg-Essen. Visiting Distinguished Professor, University of Johannesburg. Email address: fritz-stratmann@uni-due.de.
Lars Balzer, Swiss Federal Institute for Vocational Education and Training. Senior research associate, University of Johannesburg.
Antje Ehlert, University of Potsdam. Senior research associate, University of Johannesburg.
Roelien Herholdt, JET Education Services. Postgraduate student, University of Johannesburg.
Lara Ragpot, Department of Childhood Education, University of Johannesburg.

South African Journal of Childhood Education | 2014 4(2): 114-133 | ISSN: 2223-7674 |© UJ

UNIVERSITY OF JOHANNESBURG

SUN MEDIA METRO

Introduction: Conceptual development and principles of test construction

Longitudinal studies on the relationship between the mathematical competence of children at preschool age and their success at primary school have shown that such competence during the early years is an important indicator of children's later school performance (Aunola, Leskinen, Lerkkanen & Nurmi 2004; Weißhaupt, Peucker & Wirtz 2006). This applies particularly to children who have little prior knowledge when they arrive at school and who run a high risk of developing arithmetic learning disabilities. Ideally, the mathematical competence of children should therefore be recognised and described as early as possible – at preschool or during the first year of primary school – in order to effectively counteract arithmetical[1] difficulties.

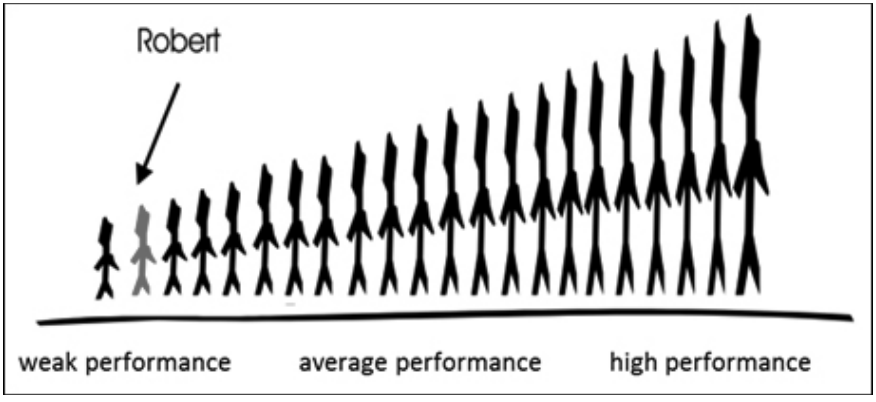
In this article we report on studies with a range of children from different parts of the densely populated, multilingual Gauteng Province of South Africa. We begin the article with a discussion of the theoretical origins of the test, arguing that there are no tests in South African languages with which to assess young children's arithmetical competence when they enter school and which have a robust theoretical base. Our argument is that practitioners need an educational test, *developmental* in its design, with which conceptual obstacles may be identified in individual children. Ideally, such tests are based on the principles of psychometric instrumentation, such as reliability of the instrument and validity of its results, which are interpreted from within a clear theoretical framework.

Currently, most teachers and school district psychology practitioners do not have access to such usable tests in different languages. Those teachers that have had training in test development usually have not been exposed to *contemporary* cognitive developmental psychology and have not learned to use tests that can yield results which will assist them in identifying Grade R and Grade 1 children's specific barriers to building mathematical concepts.[2] This void poses a real problem for teachers and remedial therapists.

After the introduction, the discussion moves to a description of a theoretical model for a test that captures children's mathematics competence on five levels of performance. This is followed by the narrative of the translation of a test, coupled with references to some theories of number development in children. Mention is also made of how language serves as integrating tool for number concept forming in young children, and of how this points to possible difficulties one may encounter in translation with regard to the *communication* of and in the test (both on the part of the assessor asking questions in oral language, and the child who communicates her or his understanding by oral responses to test items). Subsequent to this discussion there is a description of the study of one of three cohorts, in which the Rasch model is presented for one language group, namely Sesotho. The article concludes with a summary and the implications for the final standardisation, differential item function (DIF) analysis, and norming of the test for use in South Africa.

In this country, where there is only one locally standardised mathematics (curriculum-based) test for children of this age, the question about criteria for reliable assessment of mathematics competence arises, specifically with regard to the early years. In general, there are two options for capturing children´s mathematical competence. The first option is curriculum-based tests (such as the VASSI [Vassiliou 2003]), which assess children's performance in relation to the objectives of the school curriculum and in comparison with the performance of classmates. The disadvantage of this kind of test is that, based on the results (the sum of items solved), it only provides quantitative information about the norming score (the range) of the child (see Figure 1), but little information about an individual child´s knowledge and understanding.

Figure 1: Social norming of test results



(Source: See Kretschmann 2009)

A second option for the construction of tests is to do it on a developmentally oriented basis to assess the child´s conceptual knowledge in a certain domain, such as number sense and calculation ability in mathematics. Such a test enables one to get qualitative information about the concepts the child already knows; the concepts he or she is actually developing at the time of testing (in a child's *zone of proximal development*, according to Vygotsky 1978); and the concepts that have not yet been formed. Without a conceptual model of some sort, this would remain pedagogical guesswork, with no specifics.

In a search for ways to capture foundation phase children`s competence in mathematics, South African researchers (Henning 2013) learned of a German mathematics test known as the *MARKO-D* in 2011, when it was standardised and normed in that country. The test was published in Germany in 2013 (Ricken, Fritz & Balzer 2013). It is based on a developmental model of the mathematics concepts of number and arithmetical calculation in the four to eight years age group (Fritz, Ehlert & Balzer 2013). The starting point for the composition of this model was the assumption that key arithmetical concepts develop hierarchically and that children develop more sophisticated cognitive structures in a step-by-step manner. Each step is marked by

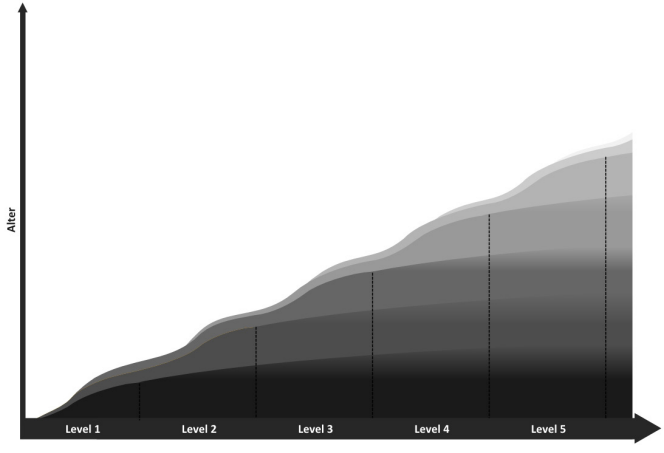
the forming of specific concepts that build on one another in an accumulative way. The literature on, for example, the development of number concepts, is quite firm about the fact that there is some sequence in the development of small numbers (Feigenson, Dehaene & Spelke 2004). However, once children have grasped, for example, cardinality of a number, or once cardinality induction has occurred, they can learn larger numbers faster and even simultaneously (Sarnecka & Carey 2008).

The *MARKO-D* test of early number development and calculation, which, according to the literature and the empirical development of the conceptual model on which the test was founded, comprises items that show a sequence in the development of number sense at this age, has not been used in South Africa before. The test was designed as an individual test for children, with items in the original German test being embedded in a little narrative of two squirrels. The story and pictures serve only as illustrations to motivate the children.

After testing a child on the *MARKO-D*, one not only has the test score, but one is also able to allocate the performance of the child to a specific conceptual level of the conceptual model (Fritz et al 2013). This means that, for instance, a child 'allocated' to Level III has developed the concepts of the first two levels, but is still developing an understanding of the concept of Level III and has little idea yet of the concepts of the higher levels. Since the development of concepts is not a totally linear, sequential process, newly developed concepts do not immediately replace earlier, 'quasi' or 'emergent' concepts. The new concepts and the earlier ones exist in parallel for some time during conceptual development. Only gradually (step by step), with continuous effective use, the new concept will begin to dominate the strategies for solving problems. This is described in Siegler's *overlapping waves theory* (Chen & Siegler 2000; Siegler 1996) (See Figure 2).

The *MARKO-D* testing instrument was developed from this perspective to capture children's concept development in number and calculation.

Figure 2: Levels of the *MARKO-D* test



The levels of conceptual development (also the level of difficulty of the test items) are described next, with some indication of how these are operationalised in the test items.

Levels of child concept development mapped onto the *MARKO-D* test

These levels have been fully described by Fritz et al (2013); what follows is only a summary with some indication of how the items operationalise the constructs on the levels.

Level I – Counting: The first concept in acquiring precise arithmetical (and thus mathematical) understanding is the knowledge of the sequence of number words, with the understanding that each number word has a specific meaning and that number words can be used for counting and enumerating sets. Sets, comprising elements, can be counted by putting single objects in a one-to-one correspondence with number words. More challenging than the counting of a set, is the task to count out a specific number of elements within a larger quantity. The child's ability to do this can be ascertained by means of Wynn´s 'Give a number' tasks, such as, 'Give me five counters!' from a group of many (Wynn 1990).

Examples of item types on Level I:

- Reciting the counting list up to 10
- Matching sets with tangible objects in a one-to-one correspondence
- 'How many are there'? tasks
- Wynn´s 'Give a number' tasks

Level II – The mental number line: At Level II, the representation of numbers changes from a qualitative representation to a linearly increasing mental representation in a 'line' (which need not be straight, because there is no evidence of the mental line being straight in children's own mental representation) of increasing or decreasing numbers. Children construct (their own) mental image of a *number line* that spans the counting list (Dehaene 2011). In this image or representation, the numbers gradually increase and 'later' or 'further in the list' – probably indicating *time*, *space,* and *number* representation (Dehaene & Brannon 2012) – implies 'greater number'. This representation allows children to determine preceding and succeeding numbers. Addition and subtraction problems become solvable by moving forward or backward along the mental number line. For this, all quantities have to be counted out individually, meaning that children apply the 'counting-all' strategy, or count from the first number onwards (See Carpenter & Moser 1983).

Examples of item types on Level II:

- Finding preceding and succeeding numbers
- Addition and subtraction in story problems

Level III – Cardinality and decomposability: Real, cardinal understanding of number requires the mental integration of the elements of the counted quantity, while

simultaneously grasping the concept of the whole that they constitute. In this way, the number becomes a composite unit (Steffe, Cobb & von Glasersfeld 1988) in which the distinct, individual objects of a quantity are combined into one quantity, independent of *where* (the order irrelevance principle) and *what* (representation) is counted out (Piaget 1965). The mental number line, expressed in language (number words), corresponds with the sequence of ascending cardinal units, where the quantities follow a fixed order (quantity seriation). At this level, children have now experienced cardinality induction (Feigenson 2012).

Examples of item types on Level III:

- Organising sets and identifying the elements with ease
- Applying the 'order-irrelevance' principle

Level IV – Class inclusion and embeddedness: Based on the understanding of or full induction into cardinality, the child understands that each number represents a specific quantity, which is composed of a specific number of elements. As a result, each number is understood as a composition of any kind of combination of smaller numbers, so that each number can be decomposed into partial quantities, which, together, are equivalent to the total quantity. The relationship between *part-and-part with the whole* is determined, so that the following task is solvable: 'Give me five counters. Three of them must be red!'

Examples of item types on Level IV:

- Determining subsets
- Determining subsets with only one set

Level V – Relationality: On this level, the number sequence (expressed in language with number words) is understood as a sequence of cardinal units in which each successive number word represents a cardinal number that is one larger than the cardinal number represented by the previous number word. Therefore, the intervals between successive numbers are *congruent*. Based on this knowledge, children now have a kind of scale, which enables them to compare quantities and to precisely determine the differences between numbers on this scale. They realise that numbers do not only represent concrete quantities, but also *counting acts* that can themselves be counted. Therefore, the following tasks become solvable:

Examples of item types on Level V:

- Determining number relationships
- Recognising differences between sets

Translation into four South African languages

During the translation process, the question of the role of language and the communication of concepts consistently came up: How does language feature in early mathematical concept development? Apart from the view that language acquisition itself

plays a crucial role in early number concept development, for the purposes of interview-based testing, this is also important: the child solves problems in an oral communication activity with the test administrator and relies on language for much of the conversation. Henning and Ragpot (2014) refer to Spelke (2012) and Carey (2009) to elaborate on this interplay between language and the development of number concepts:

> From the literature that we have encountered, our inclination has been to follow Susan Carey and Elizabeth Spelke, who say, in different ways, that a child's core knowledge and her or his experience/interaction build up to a point where they come together, or are "productively combined" (Spelke 2012:305), with language as the combinatory agent. Spelke observes that language is likely to be the only connecting feature between OTS and ANS[3] with which to further mathematical concept development.

Spelke explains this further, pointing to the combinatory and integrating role that language plays:

> I believe the role played by language is small (consistent with the intuitions of mathematicians), but crucial. All of the information supporting our numerical intuitions derives from the two core number systems and these systems are fully independent of language. Nevertheless, the language of number words and quantified expressions may serve to link the information together. Absent language, human infants and other animals may have all the information they need to represent the natural numbers, but they may lack the means to assemble that information into a set of workable concepts.
>
> (Spelke 2012:305)

In administering the test over the past three years, it has become evident that language – including vocabulary, but also syntax and morphology – plays a substantial role in the way in which children communicate their understanding of number.

On the basis of the German *MARKO-D*, the test was translated in order to use it locally, taking cognisance of the need to not only be precise, but also to give credence to linguistic variability. The next step was thus to translate and adapt the test for children in this country into at least four languages. The main focus here was to make sure that the instrument continues to assess what it does in the test's original language – in other words, to ensure that the translated items retain the conceptual content of the original test. Until some empirical work can shed light on the tool's efficacy and the validity of its results in the target language, the implications for exported, translated tests remain unknown, especially when the language of origin and the target language come from vastly different linguistic families and have different linguistic-cultural ambience and communication requirements (Slobin 1996; Lai, Garrido Rodriguez & Narasimhan 2014). In addition, research on language as a "temporary strategy in mental processes" (Lai et al 2014:139) suggests that language serves as a strategic cognitive option when seeking ways to address a problem (for example, an item on a test).

The translation proceeded through various stages. Firstly, the items were translated into English by a researcher who knows both languages and who studies childhood education. The translation was also back-translated to German by a primary school teacher in a German school in South Africa. Each item was translated and analysed in small trials with education students, practising teachers and children. It

was evident from this process that colloquial South African school English was not always grammatically correct and that a correct formal translation would not yield items that children would grasp. An example of this is an item containing the German term 'weniger als', which was translated by the grammatically correct 'fewer than'. However, in pilot tests with thirty children, it was found that 'less than', although grammatically incorrect, would serve the content better. Other translation problems had to do with the use of prepositions and adverbs, such as 'behind', 'in front of', 'after', 'before', 'next', and so forth.[4]

The English version of the test formed the basis for the translations into isiZulu and Sesotho, and eventually also Afrikaans (De Villiers, in progress). African language linguists from two South African universities were consulted in an effort to ensure usable and effective versions of the isiZulu and the Sesotho translations. However, at the time of the pilots, multilingual applied linguists alerted us to the overly formal qualities of the items as translated by the academics, some of which would not be understood by urban children, whose language use is characterised by dialecticalisation (RSA DBE 2013) and linguistic code-mixing (Henning 2012). This means that the geographical area where the language is used will determine the choice of words of the users, especially in forming conjunctions. So, for instance, in the folk terminology one finds references to 'deep' isiZulu, urban isiZulu, Johannesburg isiZulu and Durban isiZulu, with *isicamtho* being the term that describes a type of code-switching 'urban brew'. This is evident from a national investigation into schools (RSA DBE 2013) as well as classroom research (Henning 2012).

After three more pilot runs of the isiZulu and Sesotho versions, we settled on a translation that seemed appropriate for the children of the Gauteng Province, which is the industrial heartland of the country and arguably the most multilingual province. Here, many children live in multilingual homes, while also learning English. The Afrikaans version of the test was piloted in the same way, but with certain advantages, namely that word order is similar in German and Afrikaans and many of the terms were also easily translatable from word roots and similar morphemic structures (De Villiers, in progress). The Afrikaans version was back-translated to German and English and a panel of test administrators also reviewed the translations with a view to commenting on their experiences with children in the pilots.

In total, the process of translating the German test into four South African languages comprised nine iterations and revisions, with back-translation by different panels of language users. We adapted the narrative of the two squirrels in the German test into a localised story about two meerkats, named Jobo and Lona, and their friends, the rabbits Bongani and Naledi.

Research questions for the test in four South African languages

After having translated the test into four South African languages, we had to ensure that we could empirically validate the model in each language and that the translated items retained the conceptual content of the original test. We addressed three research questions:

1. Do all the items in each language form a one-dimensional hierarchical scale similar to the one in the German test?
2. Is the hierarchical scale of the different concepts identifiable and do the sequence of the concepts correspond to the sequence in the German test?
3. How do 'misfitting' items, which are allocated to another conceptual level, show up on examination?

The 'misfitting' items provide the researcher with the most revealing information and have to be considered in relation to the research questions. Are they allocated to another level because of translation errors or a culturally different understanding, or because the model does not fit this language use?

Method: Rasch modelling

When it comes to developing a test based on the idea of competence levels of cognition, it makes methodological sense to use item response theory (IRT) for empirical validation. If the construct in question is assumed to be unidimensional (in this instance, mathematical/arithmetical competence), the one-dimensional (1PL) Rasch model is appropriate (Wilson 2005).

The Rasch model was developed by the Danish mathematician Georg Rasch[5] with a view to bringing measurement in the social sciences closer to the standards of measurement in the natural and physical sciences. This resulted in a test construction process of linear measures for one or more unidimensional constructs. Thus, one requirement for this type of analysis is that the construct that is to be measured by the test has to be operationalised by a set of items. It is then necessary that all items in the test measure the construct in question, but nothing else, such as, for example, the language of communication in the test. One important characteristic of the Rasch model is that, in case of adequate *item fit*, it allows the creation of an interval scale representing both item difficulties and person abilities on the same scale. This means that only sum scores (solving items correctly or not) are needed and used to locate persons, with their ability, on this scale. In order to allow the creation of such a scale, it is also necessary that items have different levels of difficulty, ideally covering the whole difficulty range of the construct of interest, so that persons with a high ability will solve more items than persons with a lower ability (Bond & Fox 2007).

Since the Rasch model allows the creation and testing of an interval scale, representing both item difficulty and personal ability on the same scale, individual scores (solving items correctly or not) are sufficient to locate persons, with their individual ability, on this scale. However, as a perfectly unidimensional test is not possible in practice, a statistic is necessary to verify whether such a scale can be constructed appropriately, from an empirical perspective. Various model fit statistics are available for this purpose. Most popular are the so-called *infit* and *outfit* measures of model adequacy (Bond & Fox 2007). The infit statistic is used to compare the actual, observed relative solution frequency (the item difficulty) with the solution *probabilities* predicted, based on the model. The outfit statistic, on the other hand, is

sensitive to individual persons showing an answer pattern that is incompatible with the model – that is, persons who unexpectedly answer a rather difficult item correctly (by guessing, for example), or inversely, persons who unexpectedly answer an easy item incorrectly (for example, because of careless mistakes). The precise criteria for model compatibility by means of infit and outfit are handled differently. According to Linacre (2002), 'bad' outfit values are less important than inappropriate infit values. Therefore, only infit values were considered for item selection; for example, in the early PISA studies (Adams & Wu 2002). Furthermore, critical values have to be determined for the *goodness-of-fit* measurements. Standardised infit or outfit values (MNSQ) close to 1 indicate good model fit. Higher MNSQ values point to too low selectivity; while too low MNSQ values indicate too high selectivity and thus redundant items in the test. Wright and Stone (1999) suggest the range of 1 ± 0.5 as limiting values for MNSQ for tests in the high-stakes region, and Wright and Linacre (1994) recommend 1 ± 0.2, and for less demanding settings 1 ± 0.3 for the identification of well-fitting items.

First results of empirical testing of items in four South African languages in 2013

We tested over 1 600 children with the *MARKO-D* in the four South African languages (English, Sesotho, IsiZulu and Afrikaans) in several pilot studies from 2011 to 2013. We were able to, more or less, validate the model (addressing research question 1) in each language, with most of the items allocated to the appropriate level (addressing research question 2). However, there were some items in each language that did not fit as expected. Interestingly, these items were different in each language. Hence, the next step was to find explanations for the 'misfitting' items and to figure out whether they were due to the translations, coupled with cultural differences, or another sequence in acquiring the math concepts (addressing research question 3). If the last cause proved true, this would mean that the model would not hold in this language.

'Misfitting' items in the South African language versions of the test

We planned the exploration of possible causes for the divergent allocation of the misfitting items in three steps: The first step was to analyse the translations again; the second to reflect on the problem-solving strategies of the children; and the third was to question the model. It should be noted that these analytic processes provided insight into the content in different cultural groups and languages, as well as the structure of the hierarchical order. We concluded that the misfitting items in the translated versions of the test could mostly be attributed to translation difficulties and strategy differences in the problem-solving of the children, but not to a generally altered model (addressing research question 3).

Possible translation difficulties as cause for a different allocation of items

Example from the Sesotho version

One such item was Item 5 in the Sesotho version. According to the theoretical model, this is supposed to be a Level II item. Surprisingly, the Sesotho-speaking children were

tested on Level IV for this item in the Rasch analysis. The change in level of this item is likely to be related to the translation of the item. In English, the item reads: "What number is between 5 and 7?" In Sesotho, the item reads, "Ke palo efe e tlang mahareng a hlano (5) le supa (7)?" – the word 'palo' being used for the word 'number'. The problem seems to be that, depending on the context in which it is used, 'palo' could mean 'counting', 'reading', or 'number'. A pastor in a church could, for instance, use the word 'palo' in the context of reading a numbered verse from the Bible. The use of this word in the test item could therefore be confusing. Another aspect of the word 'palo' that could be problematic for urban Sesotho speakers is that it is not commonly used in everyday Sesotho in metropolitan areas. It is formal and would generally be used more by speakers of the language in rural (and thus culturally and linguistically more traditional) areas. In the mixed version of Sesotho, which is spoken in most urban Sesotho households, the word 'palo' would not be used when referring to number; the word 'nomoro', which is derived from the Afrikaans 'nommer', would be used instead. (This may be compared to the isiZulu translation, where the word 'inombolo', which is not confusing and is commonly used by urban isiZulu speakers, was used to refer to 'number': there was no apparent difficulty with this item and it tested on Level II, where it was supposed to.) The wording was changed to "Ke e feng e tlang mahareng ha (5) le (7)?" in the final Sesotho translation.

Examples from the IsiZulu version

Item 29 in the isiZulu test proved problematic. It is a Level IV item, but in the isiZulu translation it tested on Level II/III. Thus the item reverts to testing ordinality, which is related to the order of numbers on the number line and not to concepts on higher levels of the model. The relegation of the item to this level is likely due to the isiZulu wording. In English, the item reads: "What number is one smaller than 5?" Directly translated into isiZulu, the item reads: "Which number is under 5 by one?" The numbers 5 and 1 are thus given and the child is cued to simply count backwards by one, thus going forward or backward along the mental number line (Level II). The test administrators reported that the children often held up their hands, counting five fingers and then merely counting backwards.

Another problematic item in the isiZulu version was Item 30. This item is a Level III item, but it tested on Level II in the isiZulu translation, suggesting that it is easier. In English, the item reads: "What number is one bigger than 7?" The item in isiZulu reads: "Lyiphi inombolo engaphe isiZulu ngo-1 ku-7?" Directly translated, this reads in English as "Which number is up with one to 7?" Once again, the item is now relegated to Level II, which tests understanding of succession and sequence. On the mental number line, 'up one' from 7 signifies 'the next one'. The test administrators reported that the children would once again show seven fingers and just count on one more. In the final translation, the item reads: "Lyiphi enkulu ngokukodwa kuno-7?"

Children's problem-solving strategies on different levels as cause for a different allocation of items?

Examples from IsiZulu and Sesotho items

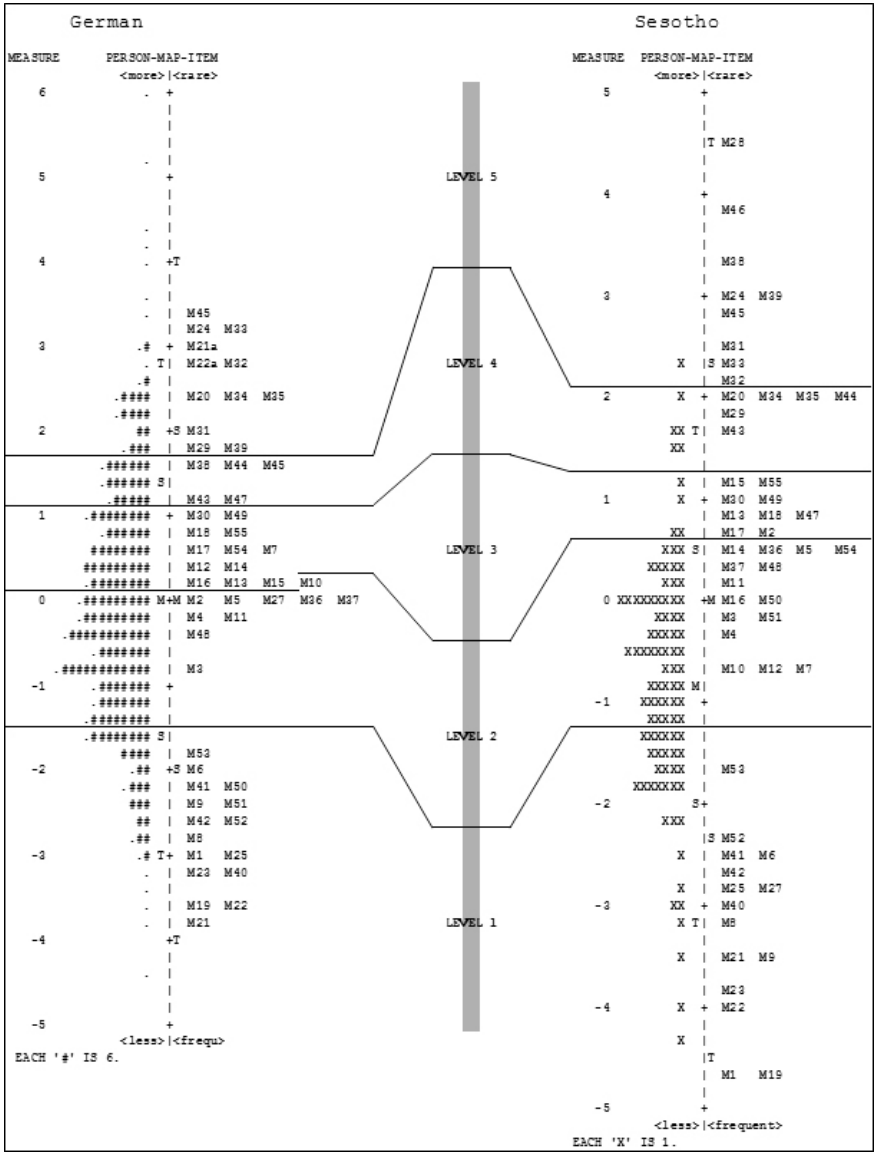
In order to solve the German equivalent of the question "Can you divide the apples for Bongani and Naledi so that both have the same? Put the fruit by Bongani and Naledi", German children use the ten counters that form part of the test pack. On the basis of a one-to-one assignment, they alternately give a counter to each character until all counters are evenly distributed among the animals. They therefore perform this task using a Level I strategy, with no arithmetical calculation. By contrast, Sesotho-speaking children worked on the task without using the counters provided – despite being requested to do so. They tried to mentally perform the task, so that they did not process this task with one-to-one assignment, but rather used arithmetic operations. If such observed strategies are considered in the categorisation of the items per level, then, in fact, the 'not too difficult' or 'easy' item in the German version again arranges itself, in line with the theory, into the model of the mathematical concepts on another level.

Second empirical testing of the items of the South African tests in 2013: Rasch analysis

After carefully changing the wording in the instructions of some items, we conducted another study with a new sample of 496 participants in the various language groups: 100 Afrikaans speakers; 197 English speakers (of which 99 were home language speakers and 98 were speakers of English as a second or additional language); 100 isiZulu speakers; and 99 Sesotho speakers. The participants were aged between 51 and 107 months (M=78.2 months; SD=6.4). There were 269 girls and 227 boys, all of whom were drawn from a population of children at the beginning of their school journey, in the second month of Grade 1, and from schools across a spectrum of socio-economic strata.[6]

Figure 3 shows item-person maps of the comparative results of two separate 1PL Rasch models for the German and Sesotho versions of the *MARKO-D* test. For each language version, we assessed 1) whether all the items form a cumulative scale; 2) whether segments on the scale that include items from a particular level can be identified; and 3) whether the sequence of these segments on the scale follows the sequence of levels in the theoretical model.

Figure 3: German and Sesotho item-person maps



On the left, [#] represents 6 persons and [.] represents 1-2 persons; on the right, [X] represents one person, and [M], [S] and [T] (left of the line for subjects and right of the line for items) represent the mean, as well as 1 or 2 standard deviations from the mean of the respective distribution. The mean of the item-difficulty scale is separately fixed at 0 logit for both item-person maps.

The unit of measurement on an item-person map, which measures and represents both the item difficulty and person ability, is defined as a *logit* ('log odds unit'). Measures are expressed on the logit scale with the average item measure or person ability arbitrarily set to 0 for every item-person map. On the common interval-scaled ability scale – with a range of -5 to +6 (German) and -5 to +5 (Sesotho) logit in the case of our study – the score distribution of the children is indicated on the left and the position of the individual items on the scale on the right. The higher a child's position, the higher is her or his ability, and the higher the position of an item, the more difficult it is. The higher the position of a child compared to that of a given item, the higher is the probability that this child will solve the item correctly.

The German *MARKO-D* (on the left side in Figure 3) was standardised with 1 095 children aged four to six years (M=64.6 months; SD=7.2). The items of the one-dimensional dichotomous Rasch model show satisfactory values (weighted infit MNSQ 1 ± 0.2 for 53 out of 55 items; weighted infit MNSQ 1 ± 0.3 for 2 items) and the person reliability is at .91. The validity requirements for a one-dimensional Rasch model can therefore be considered as fulfilled: the items form a one-dimensional cumulative scale. The German item-person map also indicates that the items administered cover the entire range of mathematical concepts described in the developmental model. Children's range of mathematical abilities is covered appropriately as well. The horizontal boundary lines between levels were added by allocating the items to the respective levels for which they had been constructed based on the model. The grouping of items according to model levels was successful. Our interpretation of the results is that the empirical allocation of all items corresponds to the theoretically predicted proficiency levels, with no exception. Thus, segments on the scale can be identified that include items which can be solved, based on one of the five numerical and arithmetical concepts proposed in the model, and the sequence of these segments on the scale follows the sequence of levels in the model.

Results of the Sesotho *MARKO-D* are still preliminary, as the current data for the Sesotho group includes scores from only 99 children aged five to nine years (M=76.8 months; SD=6.9), which is not enough to draw final conclusions. (More data will be collected in 2015.) However, the majority of Sesotho items of the preliminary one-dimensional dichotomous Rasch model show satisfactory values: a weighted infit MNSQ 1 ± 0.2 can be found for 43 out of 55 items, and a weighted infit MNSQ 1 ± 0.3 for 7 items. Poorer MNSQ infit values are found for 4 items, with 3 out of these 4 showing very low infit values (less than 0.7), which are termed 'overfit'. The responses are too predictable from the Rasch model perspective, but they do not destroy the whole picture – from a statistical point of view one could consider omitting these 4 items from the test without losing too much diagnostic information. One item cannot be estimated by the model, hence no fit statistic is computed for this item. The person reliability is at .87. For the moment, the requirements for the validity of a one-dimensional Rasch model can therefore be considered as nearly fulfilled, with some need of improvement.

The preliminary results of the data for the Sesotho group also indicate that the test may have been a little too difficult for the sample, as only a small number of children

were able to correctly solve items at higher levels and the average difficulty level of the items is higher than the average ability level of the children. However, the results also show that most of the *MARKO-D* items in Sesotho arrange in a similar sequence as the German model. This indicates that it is possible to operationalise the theoretically assumed mathematical concepts using tasks for Sesotho-speaking children too, and to model them in one Rasch model. We have therefore successfully responded to research questions 1 and 2 for the Sesotho group, with minor need for modification.[7]

Discussion: towards standardising and norming the test in South Africa

The aim of the present study was to discuss how the German *MARKO-D* math test was translated and piloted in South Africa. An important component of the work is the use of a conceptual model of mathematics concept development. The model postulates five hypothetical levels that build upon each other hierarchically, each distinguishing a specific arithmetical mathematics concept. In order to make the test usable in the target languages, we had to ensure that it measures the same concepts in each of the four South African languages as it does in the original German. The main obstacle in this process was to find appropriate item translations for each arithmetic concept, so as to be able to reconstruct the developmental levels from item solutions and ensure that each level of the model is represented in this achievement (and diagnostic) test by specific test items that map onto exactly those arithmetic concepts which are needed to solve the items of the specific level.

In order to do this, we had to complete an iterative procedure:

1. We started by translating and back-translating the items in each language;
2. We then verified the model and the appropriate allocation of the items in each language; and
3. Finally, we discussed and reflected on the model fit, and especially misfitting items.

This procedure was repeated several times with different samples and critical items were revised until the final model was arrived at in a test round in 2013. In that study, we found evidence that most of the items formed the same hierarchical (cumulative) scale in the Sesotho, isiZulu and Afrikaans language versions, and that the segments on the scale followed the sequence of levels in the original German model. This means that we were able to demonstrate that the model holds true in different languages and that items belonged to the levels for which they had been designed.

Throughout the pilot study phase and during the entire translation and verification process, we were constantly aware of how language features in the development of concepts. However, we have some evidence from the research reported in this article that it is possible to minimise (if not completely neutralise) the effect of language in the translation of test items. It is also possible to gain optimal unidimensionality of the instrument used to test mathematics competence as proposed by the underlying theory. Our argument is that if it is true that language serves as cognitive *combinatory*

agent in assembling knowledge to form concepts once children learn through language and other symbols (Carey 2009; Spelke 2012), then it would mean that they lodge much of their early understanding of maths (in the forming of concepts) in their use of language. If different language versions of the same test remain usable, it not only means that the children have been able to use their own language to develop concepts, but also that the route is likely to be similar across languages, even though languages may differ with regard to morphology, syntax, and so forth. Tests that assess children's knowledge in oral interview format (such as is done in the administering of the *MARKO-D*) are thus not undeniably free of the effects of language; however, after much iteration, one could ultimately come close to a reliable translation of such tests.

Based on these results, we can now go on to norming the tests in the four languages. These will be socially-oriented norms, which will allow for age cohort comparison and thus help in ascertaining whether, based on South African norms, children are accelerated, normal, or delayed in their development of mathematics concepts. In addition to this, researchers obtain *substantial qualitative information* during the administering process, which may be of help to therapists who assist children with learning support and remediation. Based on the test results, a child's competence can be allocated to a specific conceptual level and a plan can be devised to assist with the specific area where there seems to be a problem. For example, when a child's performance on the test is assigned to Level II, one not only knows, based on age cohort comparisons, whether the child is developing in an accelerated or average manner, but also that the child has already acquired the knowledge of Level I and is actually dealing with the understanding of the Level II concept. The child does not yet know the Level II concept, but it is on the child's "proximal development" level (Vygotsky 1978; Kozulin 1990).

Though the data is most satisfactory, some questions remain. As mentioned before, the item-person map (Figure 3) for the Sesotho version of the test shows that there are only a few children who are able to solve items on higher levels. The same is true for the isiZulu map. That means that we will have to take care to assess Sesotho- and isiZulu-speaking children from different educational backgrounds and include more children with a higher performance level in our norming sample.

The sample of English-speaking children is another sensitive issue. We assigned these children to a single group in our present calculations. However, it is evident that children who are home language users of English performed differently to children who use English as a second or additional language and who go to schools where they learn through the medium of English, although they do not use it at home as a primary language. The introduction of English as medium of instruction in the foundation phase for children who do not use it as a home language is one of the vexing issues in South African education. It is, therefore, important, in terms of future research, to split the English group into 'home language as medium of instruction' and 'not home language as medium of instruction'.

Lastly, but not least, the next empirical step will be to show that the final versions of the translated test show measurement invariance. This will mean that

items function in the same way across linguistically different groups. Item response theory-based studies of measurement invariance generally use methods of Differential Item Functioning (DIF) (see Raju, Laffitte & Byrne 2002), but this will be the subject of future work. Normally this type of analysis is not possible without first having established that the theoretical model holds empirically.[8]

Acknowledgements

Funding for the research programme in which this article was developed was obtained from the South African National Research Foundation (NRF Grant no.78827) and from the Zenex Foundation in South Africa. The views expressed in this article are the authors' only.

The leader of the research programme of mathematical cognition in primary school is Elizabeth Henning, the director of the Centre for Education Practice Research at the University of Johannesburg. The centre is the home of the MARKO-D research. We wish to thank her for guidance in the planning of this article. We also thank reviewers.

References

Adams RJ & Wu R (eds). 2002. *PISA 2000 Technical Report*. Paris: OECD.

Aunola K, Leskinen E, Lerkkanen M-K & Nurmi J-E. 2004. Developmental Dynamics of Math Performance from Preschool to Grade 2. *Journal of Educational Psychology*, 96(4):699-713.

Bond TG & Fox CM. 2007. *Applying the Rasch Model. Fundamental Measurement in the Human Sciences*. 2nd Edition. Mahwah, NJ: Erlbaum.

Borsboom D. 2006. The attack of the Psychometricians. *Psychometrika*, 71(3):425-440.

Carey S. 2009. *The Origin of Concepts*. Oxford: Oxford University Press.

Carpenter TP & Moser J. 1983. The development of addition and subtraction problem-solving skills. In: TP Carpenter, J Moser & T Romberg (eds). *Addition and subtraction: A cognitive perspective*. Hillsdale: Lawrence Erlbaum. 9-24.

Chen Z & Siegler R. 2000. Overlapping waves theory. *Monographs of the Society for Research in Child Development*, 65(2):7-12.

De Villiers H (in progress). *Die bruikbaarheid van die Afrikaanse weergawe van die MARKO-D toets* ('The usefulness of the Afrikaans version of the MARKO-D test'). MEd study. Johannesburg: University of Johannesburg.

Dehaene S 2011. The number sense. *How the mind creates mathematics*. Oxford: Oxford University Press.

Dehaene S & Brannon EM (eds). 2012. *Space, time and number in the brain. Searching for the foundations of mathematical thought.* Amsterdam: Elsevier.

Feigenson L, Dehaene S & Spelke ES. 2004. Core systems of number. *Trends in Cognitive Sciences*, 8(10):307-314.

Feigenson L. 2012. Objects, sets and ensembles. In: S Dehaene & EM Brannon (eds). *Space, time and number in the brain. Searching for the foundations of mathematical thought*. Amsterdam: Elsevier. 12-22.

Fritz A, Ehlert A & Balzer L. 2013. Development of mathematical concepts as basis for an elaborated mathematical understanding. *South African Journal for Childhood Education*, 3(1):38-67.

Henning E. 2012. Learning concepts, language, and literacy in hybrid linguistic codes: The multilingual maze of urban Grade 1 classrooms in South Africa. *Perspectives in Education*, 30(3):69-77.

Henning E. 2013. South African research in mathematical cognition and language in childhood: towards an expanded theoretical framework. *South African Journal of Childhood Education*, 3(2):56-76.

Henning E & Ragpot L. 2014. Pre-school children's bridge to symbolic knowledge: First framework for a cognition lab at a South African university. *South African Journal of Psychology. DOI:10.1177:/00812463145199 sap.sagepub.com*.

Kozulin A. 1990. *Vygotsky's Psychology: A biography of ideas*. Cambridge, MA: Harvard University Press.

Kretschmann R. 2009. Pädagogische Diagnostik als Grundlage für die Begleitung von Lernprozessen ('Pedagogical diagnostics as basis for learning support'). Presentation at the Pädagogische Hochschule Bern, Forum Unterrichtsentwicklung, Bern, 14 November.

Lai VT, Garrido Rodriguez G & Narasimhan B. 2014. Thinking-for-speaking in early and late bilinguals. *Bilingualism: Language and Cognition*, 17:139-152.

Linacre JM. 2002. What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions*, 16(2):878.

Piaget J. 1965. *The child's conception of number*. New York: Norton.

Raju NS, Laffitte LJ & Byrne BM. 2002. Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87(3):517-529.

Ricken G, Fritz A & Balzer L. 2013. *MARKO-D. Mathematik- und Rechenkonzepte im Vorschulalter – Diagnose* ('MARKO-D. Mathematics and arithmetic concepts for the pre-school age-group – Diagnosis'). Göttingen: Hogrefe Verlag.

RSA DBE (Republic of South Africa. Department of Basic Education). 2013. *ANA Report 2012: Diagnostic Report*. [Retrieved 2 July 2013] www.education.gov.za/LinkClick.aspx?fileticket...tabid=424&mid.

Sarnecka BW & Carey S. 2008. How counting represents number; how children must learn and when they learn it. *Cognition*, 108:662-674.

Siegler RS. 1996. *Emerging minds: The process of change in children´s thinking.* Oxford: Oxford University Press.

Slobin DI. 1996. From "thought and language" to "thinking for speaking". In: JJ Gumperz & SC Levinson (eds). *Rethinking linguistic relativity*. Cambridge: Cambridge University Press. 70-96.

Spelke E. 2012. Natural number and natural geometry. In: S Dehaene & EM Brannon (eds). *Space, time and number in the brain. Searching for the foundations of mathematical thought*. Amsterdam: Elsevier. 287-317.

Spelke ES. 2000. Core Knowledge. *American Psychologist*, 55(11):1233-1243.

Steffe LP, Cobb P & Von Glasersfeld E. 1988. *Construction of arithmetical meanings and strategies.* New York: Springer Verlag.

Vassiliou CP. 2003. *VASSI Mathematics Proficiency Test Foundation Phase*. Pretoria: MindMuzik.

Vygotsky L. 1978. *Mind in society.The development of higher psychological processes*. M Cole, V John Steiner, S Scribner & E Souberman (eds and transl). Cambridge, MA: Harvard University Press.

Weißhaupt S, Peucker S & Wirtz M. 2006. Diagnose mathematischen Vorwissens im Vorschulalter und Vorhersage von Rechenleistungen und Rechenschwierigkeiten in der Grundschule ('Diagnosis of mathematical pre-knowledge and arithmetic performance of children with barriers to mathemtical learning'). *Psychologie in Erziehung und Unterricht*, 53(4):236-245.

Wilson M. 2005. *Constructing measures. An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum.

Wright BD & Linacre JM. 1994. Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3):370.

Wright BD & Stone M. 1999. *Measurement Essentials.* 2nd Edition. Wilmington, DE: Wide Range Inc.

Wynn K. 1990. Children's understanding of counting. *Cognition*, 36:155-193.

Endnotes

1. The core terms "mathematics", "arithmetic", "calculation" and "numeracy" are used freely, and often interchangeably, to include broad semantic categories, that may subsume others.
2. The VASSI Mathematics Proficiency Test (Vassiliou 2003) is standardised and claims to be diagnostic. It is available in Afrikaans, English and Sesotho. The manual states that "only learners were included in a language group if their home language corresponded to the test language". This means that learners in the English norm group were English home language speakers. This excludes the majority of learners who go to schools where they learn through the medium of English as additional language. A further problem is that nowhere in the test manual is there mention of a cognitive developmental model that guided the author to include specific items in the test, or which should guide the test user in interpreting the test scores. Borsboom (2006) argues that the lack of a theoretical model with which to explain how attributes are linked to indicators is a threat to contemporary psychometric practices.
3. The object tracking system (OTS) and the approximate number system (ANS) are two instances of 'core' (Spelke 2000) or 'innate' knowledge.

4. It is worthwhile to note that neuroscientists such as Dehaene and Brannon (2012) now show that concepts of *time*, *space* and *num*ber are neurologically proximal. This could explain the regular use of prepositions or adverbs related to space and time in conjunction with references to number in everyday discourse.
5. See www.rasch.org.
6. More data will be collected in all languages during 2015, for final norming of the test.
7. Item-person maps for the other language groups look rather similar.
8. The recently published article by Graham Dampier in the *South African Journal of Education* 34(2) was based on very preliminary, pilot-testing data, collected in 2011 and 2012 and assembled from results of different versions of the test. Therefore the conclusions drawn can be considered preliminary.