Operationalizing levels of academic mastery based on Vygotsky's theory: The study of mathematical knowledge

Peter Nezhnov[a], Elena Kardanova[b], Marina Vasilyeva[c], &Larry Ludlow[c]

[a]Center for International Cooperation in Education Development (Russian Federation),[b]National Research University Higher School of Economics (Russian Federation),

and [c]Boston College (USA)

Author Note

Thanks are due to the schools and children who participated in this study; and to L. Ryabinina for her assistance with organizing and conducting the study and with data analysis.

Correspondence concerning this article should be addressed to Marina Vasilyeva, Lynch School of Education, Boston College, Chestnut Hill, MA 02467. E-mail: mvasilyeva@bc.edu.

Abstract

The present study tested the possibility of operationalizing levels of knowledge acquisition based on Vygotsky's theory of cognitive growth. An assessment tool (SAM-Math) was developed to capture a hypothesized hierarchical structure of mathematical knowledge consisting of procedural, conceptual and functional levels. In Study 1, SAM-Math was administered to 4th-grade students (N = 2216). The results of Rasch analysis showed that the testprovided an operational definition of the construct of mathematical competence that corresponded to the theoretically-based hierarchy of knowledge. In Study 2, SAM-Math was administered to students in 4th, 6th, 8th and 10th grades (N = 396) to examine developmental changes in the levels of mathematics mastery. The results showed that the mastery of mathematical concepts presented in elementary school continued to deepen beyond elementary school, as evidenced by a significant growth in conceptual and functional levels of knowledge. The findings are discussed in terms of their implications for psychological theory, test design and educational practice.

Operationalizing levels of academic mastery based on Vygotsky's theory: The study of mathematical knowledge

In the area of developmental psychology and educational research, Vygotsky's sociocultural theory of development occupies a prominent position (Albert, Corea, &Macadino, 2012; Cole, 1996; Gredler& Shields, 2008). This theory haswidely influenced both psychological research and educational practice, including the design of assessment tools and instructional approaches. For example, the concept of Zone of Proximal Development has led to the idea of dynamic assessment and the notion of scaffolding as an instructional tool that can be used to facilitate children's learning (Karlström&Lundin, 2013; Kozulin,1995; Poehner, 2009; Scrimsher&Tudge, 2003; Wertch&Tulviste, 1992). Other concepts, however, have attracted attention and generated theoretical discussions but have not yet been empirically tested. In particular, while Vygotsky (1978) wrote broadly about different levels of knowledge, there is still a need to operationalize his construct of knowledge in the developmental context.

According to Vygotsky's theory, cognitive growth can be described as a process of internalizingculturally transmitted knowledge, which involves acquisition of generalized schemas of thinking and symbolic systems (Vygotsky, 1978; 1994a).Exposure to cultural models stimulates a gradual internal process of knowledge development. At the early stages of this process, individuals master specific procedures and associative links. At this level, their problem-solving very much relies on external characteristics of the problem; their ability to solve problems depends on how similar they are to the ones that had been directly taught. From this level, knowledge continues to develop to a more deep-level understanding of conceptual relations underlying learned procedures and finally, to the highest level of understanding that allows a person to see the boundaries of the knowledge acquired and to be able to consider a multitude of possible relations within these boundaries. These theoretical

ideas have been expanded in the work of Galperin (1998),Davydov (1972, 1996), Elkonin(1989) and Nezhnov (2007)who outlined criteria for defining the mastery of successive levels of knowledge.

Up to now, there have been no direct ways of assessing the different levels of mastery within the outlined theoretical framework. There has been some work, in which levels of knowledge acquisition have been discussed in thecontext of the sociocultural theory, yet the empirical basis of these discussions includedprimarily qualitative classroom observations (refs). There are no currently available quantitative instruments that demonstrate the existence of atheoreticallypredicted hierarchy in children's mathematical knowledge.Yet, operationalizing the levels of academic mastery hypothesized by Vygotsky and his followers is not only important for testing the theory, but also for applied – educational – reasons. It is critical to provide educators with measurement tools that will allow them to obtain not just a quantitative picture of the relative rankings of different students, but a substantive characterization of students' knowledge, which should lead to identifying more precise targets for intervention. Such characterization can be conceptualized in terms of a hierarchical model reflecting different levels of mastery in an academic domain.

Thus,in the current studywe present an approach to assessing mathematical knowledge that is rooted in the theoretical foundation created by Vygotsky. By doing so, we address a current gap between one of the most powerful developmental theories and the practice of educational testing. Specifically, we propose an assessment tool designed to capture a hierarchical structure of mathematical knowledge. We then determine whether empirical data supports the theoretical construct that provides the basis for this new assessment tool, followed by an analysis of the application of this tool to the study of developmental changes in the nature of mathematical skills.

Levels of mathematical knowledge

According to Vygotsky's theory, learning plays a key role in the process of cognitive development, extending children's skills by creating a Zone of Proximal Development (ZPD). It is important to point out that the act of transferring information during formal math instruction is viewed as an impetus for subsequent internal development, but the process of internalization does not end when instruction ends (Vygotsky, 1978). In fact, based on this view, knowledge continues to develop well past the point when it was formally taught, gradually proceeding through different levels of depth.In recent work, researchershave attempted to provide more concrete psychological interpretation of this view, which required focusing on specific academic domains and defining criteria for distinguishing knowledge levels within a particular domain (Nezhnov, Kardanova, &Ryabinina, 2014).In this work, three increasingly complex levels of mastery have been proposed: (1) procedural knowledge, (2) conceptual understanding, and (3) functional competence.Below we provide the characteristics of the three levels in the context of mathematical learning, illustrated by two sets of examples. One set of problems represents the domain of numeric reasoning and the other represents the domain of geometric reasoning. In both cases, the examples reflect the mathematical content taught in elementary school and thus can be used to assess the depth of knowledge at the end of elementary school.

INSERT FIGURE 1 ABOUT HERE

Level 1: Procedural. The extent of understanding is relatively narrow, mostly encompassing knowledge of specific algorithms and standard procedures that have been directly taught. In problem solving at this stage, students are mostly oriented towards external (descriptive) features of the problem, which allow them to identify it as belonging to a particular category and invoke an algorithm used for this category of problems. In this case, the description of a problem can be associatively linked to a learned procedure.In other

words, the student does not have to make an effort to extract the underlying conceptual meaning.

An example of a fourth grade math problem requiring numerical reasoning at this level is presented in Figure 1 (problem 1A). In this case, the formulation of the problem contains direct clues about the required operation, namely division. Thus, to solve the problem, the student needs to divide a 5-digit number by a 2-digit number, which requires recalling a learned algorithm of long division and executing it correctly. The problem is presented in a standard way that should be familiar to a student who has learned the corresponding topic. The student still has to know how the algorithm works and some students, especially at the early stages of learning, may find it challenging to follow the steps of the procedure. However, once the procedure is mastered, there is no need to discover new ways of applying this knowledge outside of a standard context. A parallel example of a fourth-grade problem tapping the procedural knowledge of spatial measurement can be seen in Figure 1 (problem 1B). In this case, students are presented with a standard problem – calculating the area of a rectangle – and the presentation format is familiar to students who have been taught about areas of simple figures. The problem can be solved by iterating the square unit across the target rectangle. Applying this procedure in a straightforward way and using careful and systematic unit counting should bring the student to a correct response.

Level 2: **Conceptual**. Students begin to understand not only how to solve problems whose formulation easily leads them to the use of known algorithms, but also how to solve a whole range of problems related to the same concept, regardless of whether they are formulated in a standard or novel way. Solving problems at this stage generally requires understanding a mathematical principle, or a fundamental relation underlying a particular concept. In contrast to the previous level, the problems of this level are often formulated in a way that makes it difficult to map their description onto a given algorithm. The student needs

to analyze the meaning of the problem, which may require transforming its description in order to understand how to approach its solution.

An example of a fourth grade numeric problem at this level is presented in Figure 1 (problem 2A). To solve this problem, the child needs to reconstruct the computationalsteps carried out by someone else and to correct the erroneous part. What makes this problem a level 2 problem is that the child most likely cannot rely on a known algorithm to solve it but rather has to create an algorithm by taking into account the unique situation captured in the problem. In other words, the child needs to combine their logical understanding of the situation described in this problem with knowing how that situation can be modeled through mathematical means. A parallel example of a measurement problem is presented in Figure 1 (problem 2B). Here, the students have to measure the area of a target figure with a given square unit, but in contrast to the problem in row 1, they cannot simply move the unit across the figure. In this case, students need to have a deeper conceptual understanding of the relation between a unit of area and the area of the measured object, and in particular appreciate the fact that unit is a relative notion and that both the target object and the unit itself can be divided into smaller units, which can be later recombined. Whereas students at the procedural level often feel confused when presented with non-standard shapes and especially when the unit cannot be easily mapped onto the target shape, the students at the conceptual level realize that the shape similarity between the unit and the target is not a requirement for determining the area.

Level 3: Functional.Students must developthe depth of understanding and conceptual flexibility that will allow them to see a full range of possiblemental “moves” within the problem space and identify the sequence of moves that leads to a solution.Just like in level 2 problems, students master the concept in a generalizable way, so that they can solve novel problems. However, in addition to this, the problems at level 3 necessarily require that the

child compares multiple ways of approaching the problem. In a way, the child should carry out a series of mental experiments and compare their results.

An example of a fourth grade problem at this level is presented in Figure 1 (problem 3A). In solving this problem, the child may realize that they should use the largest digits for the highest place value. To assign specific digits to specific letters, the child may need to generate possible versions of a solution. However, unlike a simple trial-and-error (blind search), this exploration should be guided by the understanding of the place value and the need to maximize hundreds, then tens, then ones. An example of a measurement problem is presented in Figure 1 (Problem 3B). Just like a numerical problem at this level, this area problem requires the student to carry out some mental experimenting. In this case, students need to use both their conceptual understanding of area and their spatial reasoning skills (in particular, mental manipulation of objects) to determine the correspondence between the size of the unit and the size of the target figure. One way to solve this problem is to construct an imaginary rectangle that is comprised of two triangles congruent to the target figure. By determining how many unit triangles can fit into this rectangle and dividing the resulting number in half, one can arrive at the correct response.

Capturing the distinction between levels of knowledge

The instrument developed for the present study (Student Achievement Monitoring in Mathematics, or SAM-Math) was designed to assess the mastery of basic mathematical concepts typically taught in elementary school by capturing the three-level hierarchy described above.By the end of elementary school, students are typically introduced to a variety of mathematical content that serves as a basis for subsequent learning in mathematics. Thus, a team of experts in developmental psychology and mathematical education generated a pool of math problems covering the range of mathematical content that is expected to be learned in elementary school. Since SAM-Math was developed in Russia, the selection of

specific content areas was based on the Russian educational standardsin mathematics (http://www.school.edu.ru/dok_edu.asp?ob_no=19815). It should be noted though that the selected content areas correspond to those listed in the guidelines issued by the National Council of Teachers of Mathematics in the United States (NCTM, Principles and Standards for School Mathematics, http://www.nctm.org/standards/content.aspx?id=16909). These content areas include: (1) *Numbers and Operations;* (2)*Relations and Functions,* (3) *Patterns,* (4) *Measurement,*and(5) *Geometry.*

The critical feature of SAM-Math is thatthe test items within each contentarea vary systematically with respect to the depth ofknowledge required. That is, each of the five areas of mathematics included in the testis represented by problems tapping the three levels of mastery. Test developers outlined specific criteria within each content area for the types of problems that require a particular level of mastery (Nezhnov&Kardanova, 2011). Table 1 provides examples of concepts tested in two of the content areas(e.g., the concept of place value in the domain of*Numbers and Operations* andthe concept of unit in the domain of *Measurement*). As shown in Figure 2, all the test items are blocked into groups of three – while all the items within one block target the same content, they differ in the required depth of knowledge. Having a set of three problems representing the same content area allows us to determine the level of mastery in that area based on the highest-level problem solved correctly.

INSERT FIGURE 2 ABOUT HERE

Research Questions and Hypotheses

The specific aims of the present investigation were twofold. First, we sought to determine whether the test items constructed to representthe three levels of mastery indeed demonstrated an empirically-based hierarchy of difficulty corresponding to the theoretically-based hierarchy of knowledge levels. To address this question, we conducted Study 1, in

which the SAM-Math test was administered to a large sample of students, and submitted the data to item analysis and theory confirmation under the Rasch model (Rasch, 1980). We hypothesized that the structure of the test would reflect the three successive hierarchical levels of mastery described above: procedural knowledge, conceptual understanding, and functional competence,

Second, we explored changes in the distribution of levels of mathematical mastery across different grade levels. The problems included in the SAM assessment cover the mathematical concepts taught at the elementary school level. Yet, according to Vygotsky's theory, learning leads development – that is, the highest level of mastery of the elementary math concepts, indicative of the full internalization of this culturally acquired knowledge, can be expected to extend beyond elementary school (Vygotsky, 1994b). In this view, the functional assimilation of the learned concepts – the ability to use them flexibly across a variety of contexts – occurs primarily after the formal instruction is completed (Vygotsky, 1986). To address the developmental aspect of mathematical learning, we conducted Study 2, in which the SAM-Math test was administered to students in 4, 6, 8, and 10th grades. We hypothesized that by 4th grade, most students will master presented mathematical content at a procedural level. While a large number of 4th graders will master this content at a deeper, conceptual, level, we can expect a further development within this level in middle school. With respect to functional competence, we predicted that for the most part its development will occur beyond elementary school.

Study 1

Method

The development ofSAM-Mathtook two years, from the springof 2010 to the fall of 2012.This work involved outlining criteria for the three levels of mastery within each of the five content areasand designing test items representing each level, as well as securing reliability and

validityinformation for the assessment.Study 1 was conducted to examine whether the empirically-based hierarchy of students' performance on test items confirmed the theoretically-based hierarchy of math knowledge acquisition.

Participants. The participants included 2216 fourth-grade studentsrecruited from 192 elementary schools (293 classrooms) in the Russian Federation. Fourth grade was chosen because it is the last year inRussian elementary schools; children enter first grade around the age of 7 years, therefore by the end of fourth grade their age range is 10-11 years. The sample was approximately evenly divided by gender: 47% boys, 53% girls.

All participatingschools were located in one of the central regions of Russia.This region was selected because its socio-economic characteristics (e.g., average salary, unemployment, educational level, urban-to-rural ratio) were similar to those in the entire country, based on census results(Social and Demographic Portrait of Russia, 2010). For example, the distribution of the region's population by educational level (62% college and above, 30% high school, 8% below high school) was parallel to that in the country (65% college and above, 29% high school, 6% below high school).Also, the ratio of urban to rural students in the region(72% urban, 28% rural) was similar to that in the country (71% urban and 29% rural).The regional department of education, in consultation with school principals, provided permission to conduct the study. Thus, the region'swhole population of fourth grade students took part in the study. There was no selection at the school or classroom level.

Instrument. The SAM-Mathtest includeda total of 45 items, divided into 15 blocks in accordance with the test structure presented in Figure 1. As indicated earlier, the items represented five different content areas. The number of items varied across content areas,reflecting differential focus on particular topics in the elementary curriculum. For example, the area of *Numbers and Operations*(12 items) takes up a much larger part of elementary instruction and covers a larger number of topics than the area of *Geometry* (6 items).It should be noted that SAM-Math is not intended to provide a separate assessment of

students' knowledge within each specific content area, but rather to provide information about the level of a student's mastery of mathematics across the different areas taught in elementary school.

Items in thearea of *Numbers and Operations*assessed children's understanding of number systems, relations among numbers, meanings of operations, as well as their ability to perform calculations. The content area of *Relations and Functions*included problems that required analyzing quantitative relations (particularly those captured in word problems) and representing these relations using mathematical symbols. Another content area introduced in the elementary school that is critical for the development of algebraic thinking is *Patterns*. Items representing this area included either numeric or spatial patterns that required children to determine the rule governing relations between different elements and predict subsequent elements based on their position in the pattern. *Measurement* problems required the knowledge of measurement algorithms and procedures as well as conceptual understanding of measurable attributes of objects, the notion of unit, and the relation between the unit and the quantity to be measured. *Geometry* problems required the analysis of properties of geometric shapes, understanding spatial relations among objects and determining locations.

The majority of test items (37 out of 45, or 82%) hadan open-ended format.They requiredeitherprovidinga brief numeric response or a simple drawing in the test booklet (for example, completing a shape pattern or placing a dot in a certain location within a figure).The remaining items (8 or 18%) had a multiple-choice format with a choice of one or more correct answers from 4-5 options.The multiple-choice items were evenly distributed across the test – they were not concentrated in any particular content area or knowledge level. All 45 items were assembled in a booklet; three items comprising each block were presented consecutively in the same order: levels 1, 2, and 3.

Procedure. In each participating classroom, the test was administered to the whole class by the teacher. The teachers helped children complete participant information on the front page of the test booklet, provided instructions and kept track of the time. The testing was conducted on the same day,during two 45-minute testing sessions with a 15-minute break between them.This amount of time and the format of test administration, which were determined based on prior pilot testing, were sufficient for 4^{th} graders to complete test. The data collection for the whole sample was completed within a two-week period. When the data were collected, all items were scored dichotomously: a student received 1 point for a correct response and 0 for an incorrect or missing response (with a maximum total of 45 points).Students' scores were subjected to a series of analyses described in the next section.

Results

We begin with a classical analysis of psychometric properties of the test, followed by the findings of Raschanalysis.

Psychometric Properties of the Test.The results of the classical analysis of test characteristics are presented in Table 1. The item difficulty levelrefers to the proportion of students solving test problems correctly. A level of 0.61 indicates thatthe test was moderately easy. The range of item difficulty is wide, with the most difficult item solved correctly by only 16% of participants and the easiest item solved correctly by 98% of participants. As a measure ofitem discrimination level (i.e., the item's ability to discriminate between high- and low-performing individuals), we use the point bi-serial correlation, measured as a correlation between performance on a single item and the whole test. Generally, values of the point bi-serial correlation of 0.2 and above are considered as acceptable indicators of the item's ability to differentiate among test takers. In our data, the point bi-serial correlation for 2 of the 45 items was .15 and .16 (these were the easiest items that were solved correctly by most

students, including high- and low-performers). Yet, for the majority of items (43 out 45)the values of this index were between 0.25 and 0.55.

INSERT TABLE 1 ABOUT HERE

Rasch Analysis. The main portion of our analysis included application of theRaschdichotomous measurement model (Rasch, 1960; Wright & Stone,1979) as a confirmatory test of the extent to which our assessment scale was successful in capturing the three-level hierarchy of mathematical knowledge acquisition. It should be noted that in addition to this analysis, in which the test is treated as unidimensional, we examined an alternative model - the Multidimensional Random Coefficients Multinomial Logit Model (Adams, Wilson, & Wang, 1997). We found that the statistical parameters of the unidimensional model were either equally appropriate or better than the multidimensional model in describing the data. In particular, the Akeike's Information Criterion (AIC) and the Bayesian Information Criteion (BIC) had better values in the unidimensional model. Furthermore, the correlations among the three dimensions of the multidimensional model were very high, indicating that they measured the same variable. This is consistent with our conceptualization that the three levels of mastery (procedural, conceptual and functional) do not represent different latent variables but rather capture the differences in the levels of acquisition of the same latent construct: mathematical knowledge. Thus, we proceeded with the unidimensional model.

We begin reporting results of the Rasch analysis by presenting the "variable map" addressing the main question of the study – whether the difficulty structure of the items corresponds to the *a priori* theoretically-specified hierarchy. We then examine the fit between our data and the Rasch model. Next, we perform a series of principal components analyses (PCA) on the standardized residuals. These analyses serve as checks on the presence of multidimensional effects, thus testing one of the measurement criteria of Rasch models –

namely, that items operationally define a continuum along a unidimensional variable. Winsteps software (Linacre, 2011) was used for parameter estimation and data analysis.

*Variable map.*Figure 3 presents the Raschvariable map (also referred to as a construct map or Wright map, depending on the literature source), which shows the relative distribution of items and test takers in a common metric. Specifically, the variable map depictsthe joint distributionof items operationally defining the mathematics variable and the locations of students, based on their total correct scores, along this variable.The left column is the "logit"unit of measurement scale(Ludlow &Haley, 1995; Wright &Stone, 1979). An item logit is the log odds difficulty associated with a task, i.e. more difficult items have higher positive valued logits. A person logit is the log odds ability associated with a person, i.e. more able students have higher positive valued logits. On the map students are represented on the left side and the items are on the right. More difficult items and higher-performing students are located in the upper part of the map (positive logits), while easier items and lower-performing students are placed in the lower part of the map (negative logits).

INSERT FIGURE 3 ABOUT HERE

The studentsample is located relatively high on the mathematics variables, which means that the test wasrelatively easy for this student group as a whole—the average SAM-Math score, transformed into logits, is located at the position indicated by the "A". This result is consistent with the finding reported earlier regarding the mean percent correct value of .61. The mean item difficulty location is indicated by the "M". The distribution of students is wide and represents, for measurement purposes, excellent differentiation between higher and lower scoring students. The distribution of item locations, too, is excellent because the span includes very easy items appropriate for less able students and very difficult items appropriate for advanced students. Furthermore, the progression of items from easier-to-more difficult represents a smooth, uniform,progressive continuum of increasing difficulty.

Wright and Masters (1982) recommend the examination of clusters of items on the logit scale as a basis for the interpretation of a latent variable. SAM-Math items, as realizations of the latent Vygotsky-based mathematics variable, define three easy to identify clusters. Specifically, all the Level 1: Proceduralitems are located in the lower part of the map up to approximately -1 logit, further up the continuumare the Level 2: Conceptualitems ending at about +1 logitand, still further up the mathematics variable arethe Level 3: Functionalitems. Clearly, this ordering of items from one level to the next is consistent with the hypothesized structure of the Vygotskian-based assessment intended by the test developers.

The finding that the items form three clear and meaningful clusters consistent with theoretical expectations is complemented by the analysis of student groupings, which was done by calculating the so-called person separation index (Wright & Stone, 1979; Stone, 2004). This index compares the distribution of person measures (estimates of ability) with their measurement errors and indicates the spread of person measures in standard error units. The index is used to estimate the number of distinct levels, or strata (separated by at least three errors of measurement), in the distributions(Wright & Stone,1979; Smith, 2001). The number of strata are calculated as: Strata=(4G+1)/3, where G is the separation index. Our analysis produced a person separation index of 2.66, indicating four statistically distinct groups of students along the SAM-Math continuum. To follow up on this finding, we later present an analysis of proficiency levels, with four distinct groups formed on the basis of students' SAM-Math scores.

*Model fit.*Rasch goodness-of-fit analyses rely principally upon standardized residuals – the difference between the observed response and the response expected under the model (Wright &Stone, 1979). The residuals are squared and summarized in the form of unweightedand weightedmean squares (in terms of Winsteps output: OUTFIT MNSQ and INFIT MNSQ, respectively). It should be pointed out thatOUTFIT MNSQ is known to be very sensitive to

unexpected responses (Smith, 1991). Thus, INFIT MNSQ statistics that are weighted by the information function and take into account the variance of expected responses are more useful for the present goodness-of-fit analysis, namely, for the question of how consistent with model expectations the overall response patternsare. Generally, a criterion of +1.2 for item INFIT MNSQ statistics is used to flag potential problems. Our analysis showed that the value of INFIT MNSQitem statistics in the present sample varied from 0.84 to 1.12 with a mean 1.00 and SD = 0.07. This result indicates that all items in our 45-item test fit the model in accordance with the chosen criteria.

A similar approach was used to analyze person fit. Utilizing the same misfit criteria as above (+1.2), we identified students (n=317, or 14% of the total sample)with responses to test items whichwere unexpected. Note that since all the items were scored dichotomously, even one or two unexpected responses (e.g., a high ability student who failed to complete an easy item) could generate large value of person fit statistics. To put our findings regarding person fit into context, we simulated three data sets that fit the model perfectly. We calculated a number of students with INFIT MNSQ statistics out of range (above 1.2) for the three simulated data sets; they were 298, 311 and 294. Across the three sets, the average number of misfitting students was 301 students (or 14%). Thus, the percentage of students that can be considered as misfitting in our data was the same as that obtained with simulated data sets that had a perfect model fit.

When we analyzed the response profiles of students with fit statistics out of range, we identified two categories of misfitting students. The first category included high ability students who answered incorrectly 1-3 easy items. For example, the most misfitting student (INFIT MNSQ= 1.00 while OUTFIT MNSQ=9.9) had a very high ability level yet two of this student's five incorrect responses were on two of the easiest items on the test – one of them was in the beginning of the test while another one in the end. This pattern is consistent with "start-up" and "fatigue" effects frequently reported in the literature (Ludlow, Costa, Johnsen, Brown, Bessan, James, 2014). The second category includedlow ability students who

answered 1-3 difficult items correctly. These responses were found at all three levels and in different blocks; they appeared to be due to chance. Furthermore, there was no systematic connection among these students based on demographic or school characteristics.

*Dimensionality.*We examined the dimensionality of the SAM-Math by conducting a principal component analysis (PCA) of the standardized residuals(Linacre, 1998;Ludlow, 1985; Smith, 2002). The PCA serves as a check on the unidimensionality and local independence assumptions of the Rasch model. Theoretically, if the assumptions hold, then correlations between item-level residuals should be near zero.If there is no second dimension remaining in the residual variation, then the principal component analysis should generate eigenvalues all near one and the percentage of variance across the components should be uniform. The eigenvalues of the SAM-Math residual correlation matrix for 44 of 45 components ranged roughly from 1.5 to 0.7 and the eigenvalue for the last component is 0.174. In addition, the variance accounted for in the distribution was roughly evenly split across components.

To obtain further evidence of the test's unidimensionality through the randomness of the SAM-Mathresiduals we performed "tailored" simulations on random data with the same student ability and item difficulty structure (Linacre, 2011; Ludlow, 1985).The result of those steps revealedeigenvalues, percent of variance accounted for estimates, and plots of the first two components of the residualsthat were strongly consistent with the observed residual results. Moreover we performed a parallel analysis (Henson & Roberts, 2006) wherein we simulated 100 sets of eigenvalue analyses and the magnitudes of the first five eigenvalues of these random data were consistent with the magnitude of the eigenvalues from our residuals. Based on these results, there is no evidence of either multidimensionality of content or of a violation of local independence.

Benchmarking

Given the technical strength of the SAM-Math, benchmarks were established to helpseparate participants into groups according to the level of their achievement. The benchmarks reflected the three levels of our theoretical model, resulting in four distinct groups shown in Figure 4 (with the lowest-achieving group not having acquired Level 1 skills and the highest-achieving group having acquired Level 3 proficiency). Specific methods of developing benchmarks are described in detail elsewhere (Kardanova&Nezhnov, 2011). It should be pointed out that a benchmark indicates the lower limit of a corresponding proficiency level. If a student's test score exceeds the benchmark, it means that there is a 50% probability that this participant will be able to complete more than 50% of items of this level. All students whose test results are under a given value are considered as someone who did not acquire this proficiency level (as well as all the subsequent ones).

INSERT FIGURE 4 ABOUT HERE

Figure 5 shows the distribution of test participants on proficiency levels for the sample included in Study 1. By fourth grade, most students presented with problems that tap mathematical content introduced in elementary school were able to solve problems not only at a procedural, but also at a conceptual level. Yet, the performance on problems reflecting the highest – functional – level of mastery was relatively low, even by the end of elementary school. This result is not surprising as Vygotsky's theory predicts that the development of the highest level of understanding of academic content proceeds beyond the point when this content has been presented to children (i.e., the notion of learning leading development).

INSERT FIGURE 5 ABOUT HERE

Summary of Results.The empirical findings obtained with the new testing instrument, SAM-Math, indicate that theconstructstructure of the test itemscorresponded to our theoretical expectations. The dimensionality analysis shows that the test items captured a single dimension of children's performance (presumably, their mathematical achievement in

elementary school). The variable map indicates that the itemsdesigned to reflect each of the threelevels of mathematical mastery indeed clustered together and were ordered as expected.Finally, the benchmarks highlight a means of providing a practical interpretation of scores both in terms of current performance and development upwards along the mathematics continuum. In Study 2, we examine the developmental course of the acquisition of foundational mathematical concepts introduced in elementary school. To do so, we presented SAM-Math to students across four grade levels, spanning elementary to high school.

Study 2

Method

Participants. Study 2 participants included 396 studentsrecruited from 2 schools located in a large city inRussia (2 classrooms at each grade level per school).All students in participating classrooms took part in testing. This included 104 fourth-graders (53% girls), 103 sixth-graders (48% girls), 104 eighth-graders (40% girls), and 85 tenth-graders (51% girls).

Based on reports from local education officials and researchers, the participating schools served children from middle- and upper-middle class families and had a reputation as high-performing schools (statistical data on the socio-economic background of students from particular schools are not publicly available in Russia). Although this sample was not representative of the general student population in the county, our main goal was to obtain a sufficient number of students at different grade levels from comparable socio-economic and educational backgrounds so that we could conduct comparisons across grades. The choice of high-performing schools was purposeful. Since Vygotsky's theory predicted a gradual development of the functional level of knowledge (with a full mastery of basic concepts achieved well beyond elementary school), assessing high-performing students provided a strong test of this prediction. In contrast, observing a slow progress from procedural to functional levels in low-performing students could reflect a weakness of their educational input rather than the nature of the developmental process.

Instrument and Procedure. The testing instrumentand procedure were the same as those in Study 1. Assessment was conducted at the end of the school year. Data collection for all age groups was completed within one week.

Results

Before comparing students' performance across the three levels of mastery, we examined the overall characteristics of test items (average difficulty and discriminability) bygrade. The results are presented in Table 2. As shown in the table, the values of the discrimination index were in the acceptable range (i.e., above 0.2) and very similar across grade levels.The overall level of performance increased with age, as can be expected.

INSERT TABLE 2 ABOUT HERE

While the difficulty of the test as a whole changed across grades, an important question is whether the relative difficulty of individual items remained consistent fromone grade to another. To address this question, we computed correlations between items difficulties for different grades. Results presented in Table 3 show very high positive correlations for all pairwise comparisons among grades, indicating a high degree of invariance in the difficulty hierarchy of test items. Even the correlation between item difficulties in grades 4 and 10 was above .90, indicating that the items that were more difficult in grade 4 also tended to be more difficult, compared to other items, in grade 10.

INSERT TABLE 3 ABOUT HERE

Next, we tuned to the analysis of performance on the three types of items (procedural, conceptual and functional) across the five grades examined. (See Figure 6.) Students demonstrated highaccuracy on procedural items(Level 1) even in fourth grade, which is consistent with the characterization of our sample as high-performing.Conceptual and functional items (Levels 2 and 3) proved more challenging. In fact, performance on Level 3

items was far from ceiling even at grade 10. As shown in Figure 6, the difficulty hierarchy among the three levels remained invariant across grades.

INSERT FIGURES 6 ABOUT HERE

While Figure 6 depicts average results for the three levels of mastery, it is also useful to take a closer look within a given level. Figure 7illustrates the accuracy of performance on itemstapping children's functional competence (Level 3).This figurereveals widegrade-related differences in performance on individual items,but at the same time it shows very similar patterns of relative difficulty for Level 3items across grades.

INSERT FIGURES 7 ABOUT HERE

To compare statistically the accuracy of performance on different types of items across grades,we conducted a repeated-measures ANOVA with the proportion of correctly solved items as the criterion variable. Predictor variables included: item type (within-subject), grade and sex (both between-subject variables). The 3(Item Type) x 5 (Grade) x 2 (Sex) ANOVA showed main effects of Item Type, $F(2, 776) = 1451.55$, $p < 0.001$, $\eta_p^2 = 0.79$, and Grade, $F(3, 388) = 51.49$, $p < 0.001$, $\eta_p^2 = 0.29$, and no effect for Sex, $F(1, 388) = 2.76$, $p = 0.10$, $\eta_p^2 = 0.007$. In a follow-up analysis, we comparedthe means for the three levels of the Item Type variable and the four levels of the Grade variable, using the LSD method. This analysis showed significant differences for all pair-wise combinations between procedural, conceptual and functional items; as well as significant pair-wise differences between 4^{th}, 6^{th}, 8^{th}, and 10^{th} grades, all p's <.05. In addition to main effects, the ANOVA showed a significant interaction between Item Type and Grade, $F(6, 776) = 31.01$, $p < 0.001$, $\eta_p^2 = 0.19$.To examine the nature of the interaction, we conducted simple effect tests, which showed no differences among grades on Level 1: Procedural items, $F(3, 388)$ 1.80, $p = 0.15$, $\eta_p^2 = 0.007$, but significant grade differences on both Level 2: Conceptual items, $F(3, 388) = 20.32$, $p < 0.001$, $\eta_p^2 = 0.23$, and Level 3: Functional items, $F(3, 388) = 43.66$, $p < 0.01$, $\eta_p^2 = 0.32$.

We also examined our results by looking at the distribution of participants across the levels of proficiency based on the benchmarks described in Study 1.Figure 8presents this distribution for the four grade levels examined.As shown in the figure, towards the end of fourth grade, the majority of students in this sample demonstrated the second level of proficiency, which is indicative of their ability to solve more than a half of problems tapping their conceptual understanding of mathematical content covered in elementary school. Comparing the performance of fourth graders in Studies 1 and 2 as depicted in Figures 3 and 5, we see that the more selective sample (Study 2) had a higher percentage of students at the2nd level and a lower percentage of students atthe 1stlevel than the less selective sample (Study 1). Interestingly, there were no noticeable differences between the two samples at the 3rd level. The percentage of students at this – highest – level of proficiencyincreased steadily across grades, reaching the majority of students in middle and high school.

INSERT FIGURE 8 ABOUT HERE

Discussion

The present investigation tested the possibility of operationalizing the levels of knowledgeacquisition that were initially proposed by Vygotsky and further developed by his colleagues. The test used in the present study (SAM-Math) was designed to capture the distinction between procedural, conceptual and functional levels of knowledge. Specifically, the test included blocks of items that targeted the same content area but at three different levels of mastery. One of our key questions was whether the theoretically-based hierarchy of knowledge levels would be reflected in the empirically-based hierarchy of students' performance. Our analysis showed that the items designed to reflect each of the three levels of mastery indeed clustered together and were ordered as would be expected on theoretical grounds. Thus, we demonstrated a possibility to empirically distinguish between the different

types of knowledge within the same content domain. We believe that this demonstration has implications for psychological theory, test design, and educational practice.

Operationalizing a Theoretical Construct

From a theoretical perspective, an empirical testing of a hypothesized construct provides a powerful means of determining its plausibility and understanding its advantages as well as limitations. Vygotsky's notion of a gradual deepening(or internalization) of knowledge that is initiatedthrough teaching has attracted a lot of attention (e.g., Albert et al., 2012; Cole, 1996; Poehner, 2009), yet the levels of knowledge acquisition remained to be operationalized. The present investigation tested a way of operationalizing the notions of procedural, conceptual and functional levels of knowledge in the domain of elementary school mathematics. Note that the distinction between procedural and conceptual knowledge has been widely used in current educational research and practice (e.g., Silver, 1986; Byrnes &Wasik, 1991). Yet there is no systematic understanding of this distinction and different areas of knowledge often use very different criteria. Here, we used the theoretical foundation developed by Vygotsky and his colleagues to create a general framework that was applied to different content areas within mathematics. Furthermore, this framework can be used in other domains, such as science or language knowledge. In fact, our colleagues have created a parallel test of linguistic skills that is currently being piloted. Similar to SAM-Math, SAM-Language is comprised of blocks of items which tap different types of knowledge – procedural mastery of language (e.g., grammatical rules), conceptual understanding and functional competence.

In addition to its strong theoretical basis, the current distinction between levels of knowledge differs from the traditional conceptual/procedural distinction in that it introduces another level of knowledge– the level that we identified as functional competence. Vygotsky has broadly referred to this level as the ultimate stage of internalizing the concept. We must

acknowledge that the psychological nature of this level is not completely understood. In the present investigation, it was operationalized as an ability to mentally represent a range of operations that can be carried out within a given problem space. The kinds of items that were created to represent this level typically required that children carry out a mental experiment, generating and comparing several different approaches tosolving a problem and choosing the one that best satisfies given conditions.

This type of thinking is hypothesized to have a prolonged course of development. In fact, Vygotsky' theory leads to a somewhat unusual prediction that the basic concepts presented in elementary school will not be fully acquired until much later. Typically, we expect that assessment instruments designed to capture learning in elementary school students will be too easy for middle-school and especially high-school students. Yet, SAM-Math was designed to assess the same basic concepts (such as place value or unit of measurement) at different levels of depth, or different levels of mastery. Indeed our data showed that even high-school students did not perform at ceiling on functional-level items.Study 2 revealed a very gradual improvement in functional competence from elementary to middle school and then again from middle to high school. Perhaps this gradual improvement reflects the process of creating connections between students' understanding of these concepts andbroader cognitive skills (e.g., development of cognitive flexibility, logical ability).

Designing a Theory-Based Assessment

A key feature of the instrument used in the present study is that not only were the test items developed based on a systematic theoretical approach, but the same approach was used in interpreting students' performance. It should be noted that many other math tests, including the ones that are used in large-scale international assessments, include groups of items that reflect different types of knowledge, such as knowing number facts versus problem

solving. Yet, when the test data are collected, students' performance is often categorized based on a scale that does not map directly onto these substantive categories. In the present study, the initial theoretically-based distinction between knowledge levels was used not only to create test items but also to provide a framework for analyzing and interpreting students' performance. To characterize students' performance, we established benchmarks that reflected the three levels of our theoretical model. These benchmarks allow researchers and educators to capture the depth of a student's current knowledge of basic mathematical concepts and track the development in the knowledge structure across grade levels.

Implications for Further Research and Educational Practice

At the time of accelerating progress in science and technology, optimizing the process of mathematical learning and assessment has become one of the key educational goals. In this context, it is critical to provide educators with measurement tools that will allow them to better understand the level of their students' mathematical learning. In addition to obtaining an overall score as a basis for comparing achievement among students, teachers and psychologist will benefit from a more substantive characterization of students' knowledge. Vygotsky's theory offers a systematic view of knowledge development that allows for a design of testing instruments that specifically aim at identifying the depth of mastery of academic material. This type of assessment can be used to track the progress of individual students (or groups) and to compare different groups (such as classrooms, schools or even countries). For example, it can be used to compare the outcomes of particular curricular approaches. Some contemporary elementary math programs emphasize conceptual aspects of numeric development, whereas other, more traditional, programs primarily focus on counting skills in early grades. The approach introduced in the present study would allow researchers to determine to what extent the first type of program facilitates the development of conceptual understanding and functional competence.

It should be noted that the generalizabilityof the present study is somewhat limited. Study 2, in particular, only involved children from high-performing schools. On the one hand, this allowed us to show that even for the studentsexposed to favorable educational environmentthe functional mastery of mathematical concepts occurs over a long time period.On the other hand, it restricted the generalizability of the findings. Thus, in future research, it is important to use SAM-Math with a broader range of students –in order to both increasegeneralizability and address substantive questions about the relation between educational environment and the growth of mathematical skills along the three levels of mastery.

In sum, the present study offers initial empirical support for the hierarchical knowledge structure that is based on Vygotskian theoretical constructs. The new assessment tool, SAM-Math, which was designed to examine these constructs, has demonstrated sound psychometric propertiesand the Rasch analysis of scores obtained with this tool has indicated the clustering of items that directly maps onto the theoretically-hypothesized hierarchy of knowledge levels. We show that SAM-Math can be used as a tool to study developmental changes in the structure of mathematical skills and suggest that the ability of this type of test to provide a substantive characteristic of student performance is critical to facilitating a better understanding of student’s knowledge and designing effective instructional programs.

References

Albert, L., Corea, D., &Macadino, V. (2012). Rhetorical ways of thinking: Vygotskian theory and mathematical learning. New York, NY: Springer.

Byrnes, J.P., &Wasik, B.A. (1991).Role of conceptual knowledge in mathematical procedural learning.*Developmental Psychology*, 27, 5, 777-786.

Cole, M. (1996).*Cultural Psychology: a once and future discipline*. Cambridge, MA: Harvard University Press.

Galperin, P. (1998). *Psychology as an objective science*. Moscow, Russia: Institute of Applied Psychology.

Gredler, M. E., & Shields, C. C. (2008).*Vygotsky's legacy: A foundation for research and practice.* New York, NY: The Guilford Press.

Davydov, V. V. (1972). *Types of generalization in learning*. Moscow, Russia: Pedagogy.

Davydov, V. V. (1996). *The theory of developing learning*. Moscow, Russia: Intor.

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, N.J.: Prentice-Hall.

Elkonin, D. B. (1989).*Selected psychological work*. Moscow, Russia: Pedagogy.

International Journal of Testing, 10:4, 295 — 317.

Henson, R.K., & Roberts, J.K. (2006). Use of exploratory factor analysis in published research: Common errors and some comments on improved practice. *Educational and Psychological Measurement*, 66, 393-416.

Karlström, P., &Lundin, E. (2013). CALL in the zone of proximal development: Novelty effects and teacher guidance.*Computer Assisted Language Learning*,412-429.

Kardanova, E., Nezhnov, P. (2011): School achievements monitoring toolkit: Assessment framework. Paper presented at the *37-th Annual Conference IAEA*, Manila.

Kozulin, A. (1995). The Learning process: Vygotsky's theory in the mirror of its interpretation. *School Psychology International,16,* 117-131.

Linacre, J.M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2,* 266-283.

Linacre J. M. (2011). A User's Guide to WINSTEPS Program Manual 3.71.0. (http://www.winsteps.com/a/winsteps.pdf).

Ludlow, L.H. (1985). A strategy for the graphical representation of Rasch model residuals.*Educational and Psychological Measurement*, 45(4), 851-859.

Ludlow, L.H., Matz-Costa, C., Johnson, C., Brown, M., Bessan, E., & James, J.B. (2014). Measuring engagement in laterlife activities: Rasch-based scenario scales for work, caregiving, informal helping, and volunteering. *Measurement and Evaluation in Counseling and Development*.

Ludlow, L.H.,& Haley, S.M. (1995).Raschmodellogits: Interpretation, use, and transformation. *Educational and Psychological Measurement*, 55(6), 967-975.

Nezhnov, P. (2007). Mediation and spontaneity in the cultural development model.*Moscow State University Bulletin, 14,* 133-146.

Nezhnov, P., &Kardanova, E. (2011).SAM Framework.Center for International Cooperation in Education Development.

Nezhnov, P.,Kardanova, E., Ryabinina (2014).Investigating the process of internalizing learned concepts. *Educational Issues (VoprosyObrazovaniya), 1.*

Poehner, M. (2009). Dynamic assessment as a dialectical framework for classroom activity: Evidence from second language learners.*Journal of Cognitive Education and Psychology.*

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

Scrimsher, S., &Tudge, J. (2003). The teaching/learning relationship in the fist years of school: Some evolutionary implications of Vygotsky's theory. *Early Education and Development.*

Silver, E.A. (1986).Using conceptual and procedural knowledge: A focus on relationships. In J. Hiebert (Ed.), *Conceptual and Procedural Knowledge: The Case of Mathematics.* (pp.181-199). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Smith, E.V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, 2, 281-311.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals.*Journal of Applied Measurement*, 3:2, 205-231.

Smith, R.M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, 51, 541-565.

Social and Demographic Portrait of Russia (2010).http://www.gks.ru/free_doc/new_site/perepis2010/croc/Documents/portret-russia.pdf

Stone, M.H. (2004). Substantive scale construction.In E.V.Smith, R.M.Smith (Eds.), *Introduction to Rasch measurement* (pp.201-225). Maple Grove, MN: JAM Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher mental processes.* Cambridge, MA: Harvard University Press. (Original works published in 1930, 1933).

Vygotsky, L. S. (1986). *Thought and language*. Cambridge, MA: MIT Press.

Vygotsky, L. S. (1994a). The problem of the cultural development of the child.In Van der Veer, R., &Valsiner, J. (Eds), *The Vygotsky reader*. Oxford, UK: Blackwell. (Original work published in 1929).

Vygotsky, L. S. (1994b). The development of thinking and concept formation in adolescence.In R. Van der Veer&J. Valsiner (Eds), *The Vygotsky reader*. Oxford, UK: Blackwell. (Original work published in 1931).

Wertch, J. V., &Tulviste, P. (1992).Vygotsky and contemporary developmental psychology.*Developmental Psychology*, 28, 548-557.

Wright B.D., &Stone M.N. (1979).*Best Test Design.*Chicago: Mesa Press.

Wright, B.D., &Masters, G.N. (1982).*Rating Scale Analysis*. Chicago: MESA Press.

Table 1.

Summary of results for SAM-Math, Study 1

Test form 1	
Number of examinees	2216
Raw score out of 45 points: average (range)	26 (4-44)
Standard deviation	8.2
Item difficulty level: average (range)	0.61 (0.16-0.98)
Item discrimination level: average (range)	0.42 (0.15-0.59)

Table 2.

Comparison of test item characteristics at different grades, Study 2.

Grade	Difficulty level (proportion correct)	Discrimination index (point bi-serial correlation)
Grade 4	0.64	0.35
Grade 6	0.72	0.32
Grade 8	0.81	0.34
Grade 10	0.86	0.32

Table 3.

Correlations between item difficulties across grades, Study 2

	Grade 4	Grade 6	Grade 8	Grade 10
Grade 4	1	.93**	.93**	.91**
Grade 6	.93**	1	.94**	.94**
Grade 8	.93**	.94**	1	.96**
Grade 10	.91**	.94**	.96**	1

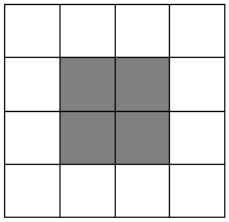
** indicates significance at 0.01 level

Problem 1A A school district received 10472 new textbooks. These textbooks have to be divided equally among 34 schools. How many textbooks should be sent to each school?
Problem 2A Peter copied a multiplication problem involving two numbers from the textbook. He wrote down the first number correctly: 7. In the second number, he accidently flipped two digits. The result he got was 147.What answer should Peter have gotten if he had copied the problem correctly?
Problem 3A What is the largest result that can be obtained if letters in the expression AB5 + BC2 are substituted with digits (different letters should be replaced with different digits)?

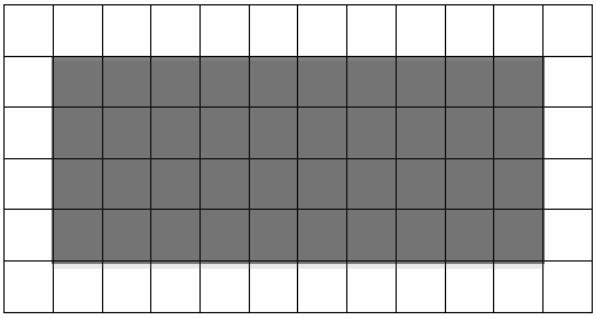
Figure 1A. Examples of numerical problems at the three levels of knowledge mastery

Problem 1B

This is a square unit.



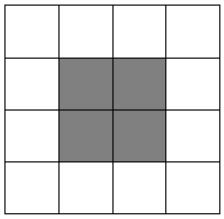
How many square units are in the figure below?



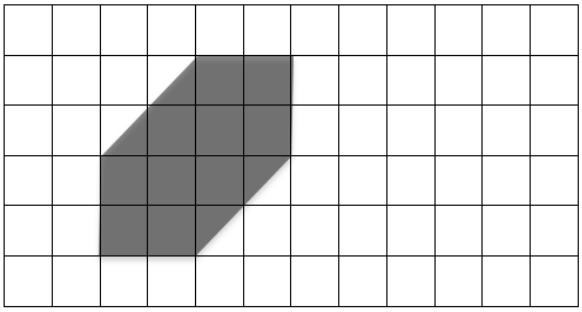
Answer: _____ square units

Problem 2B

This is a square unit.



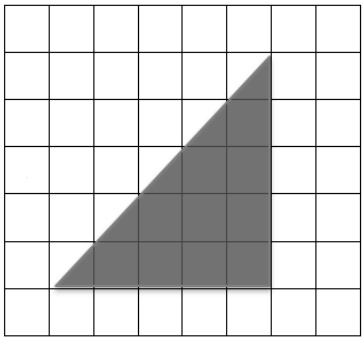
How many square units are in the figure below?



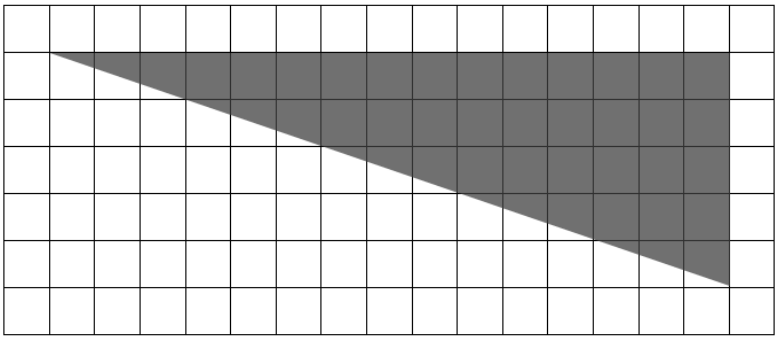
Answer: _____ square units

Problem 3B

This is a square unit.



How many square units are in the figure below?



Answer: _____ square units

Figure 1B. Examples of measurement problems at the three levels of knowledge mastery

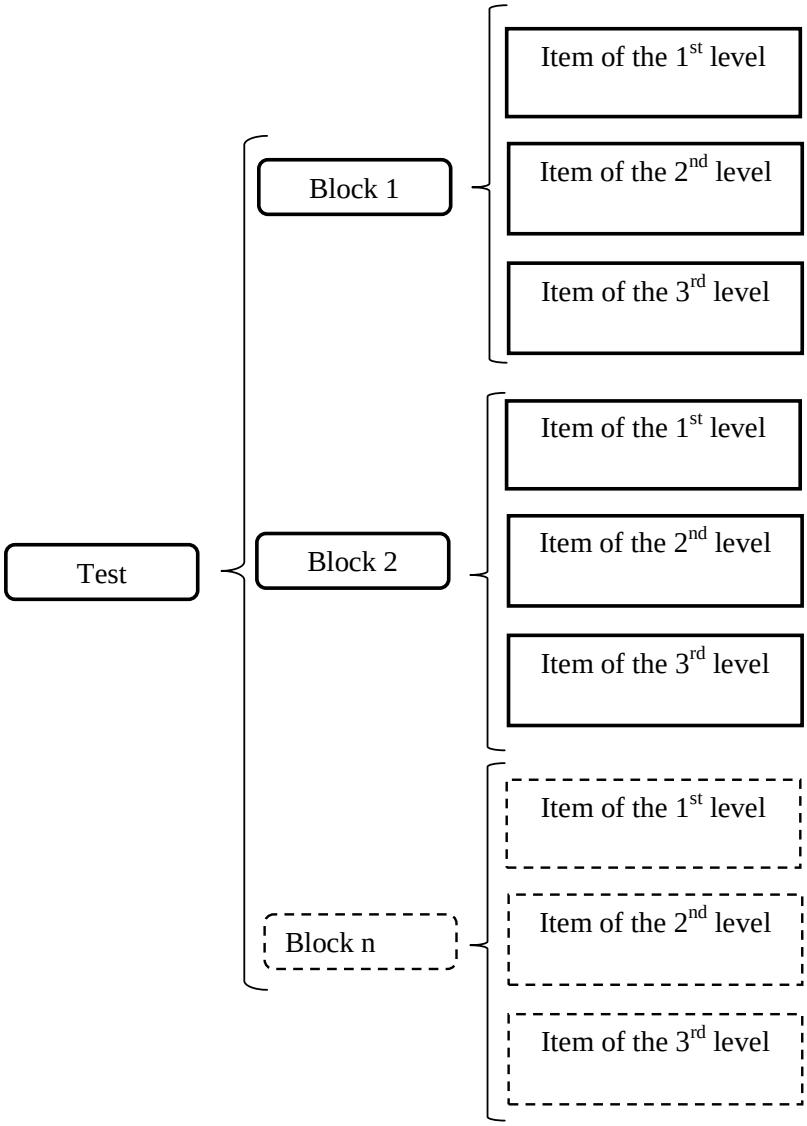


Figure 2. Structure of the test SAM-Math

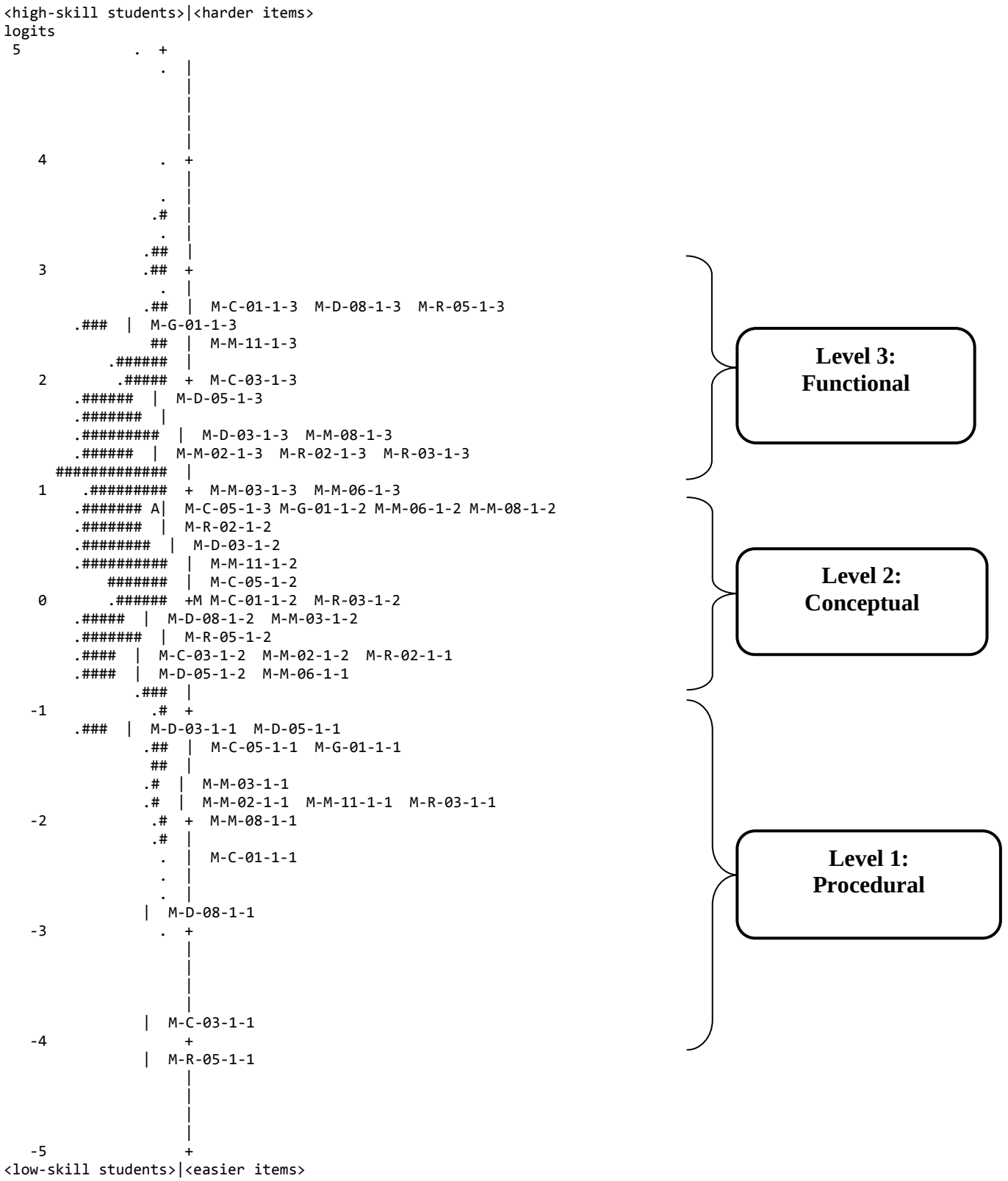


Figure 3.The SAM-Math variable map

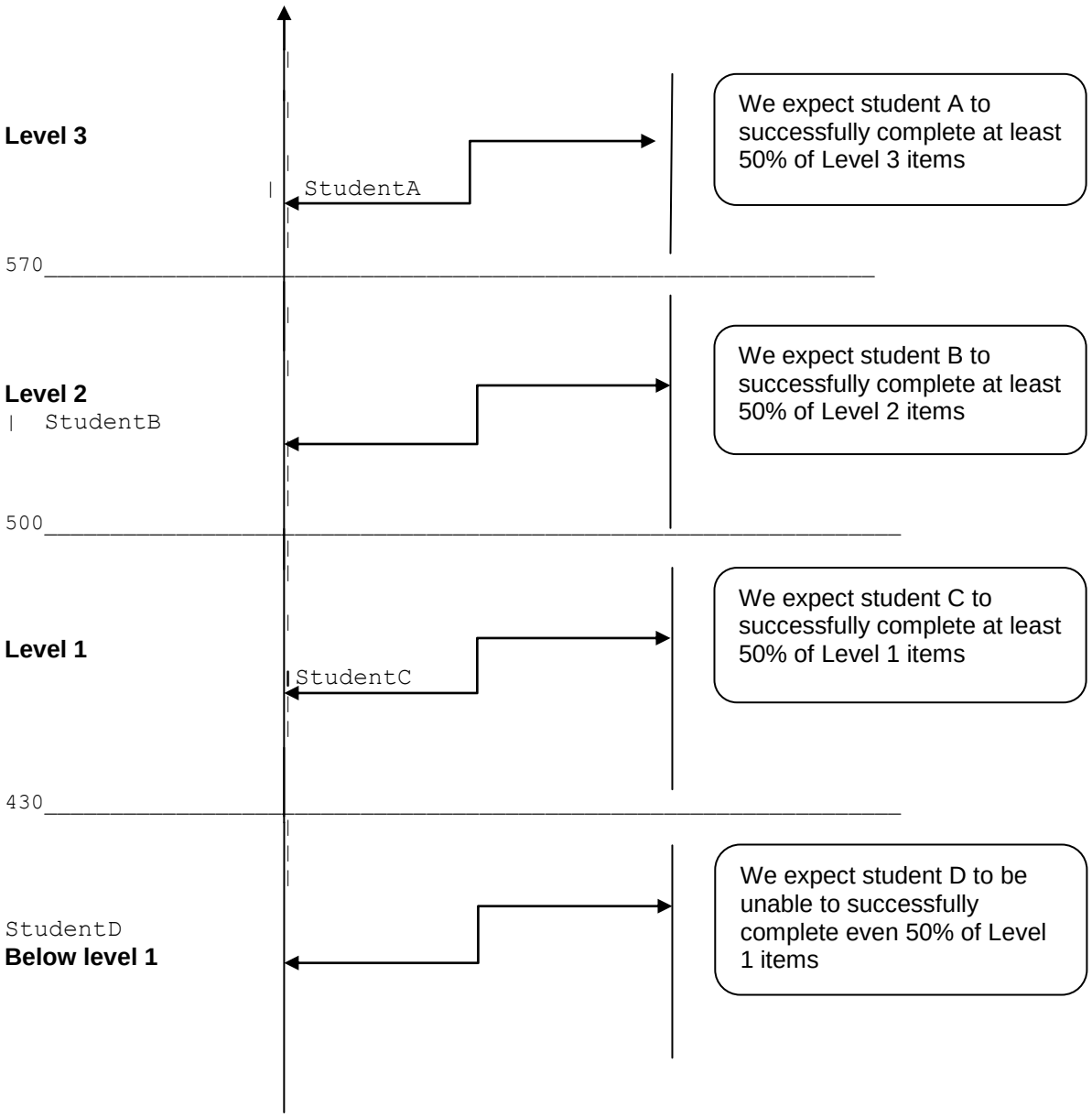


*Figure 4***.**Mathematical competence scale

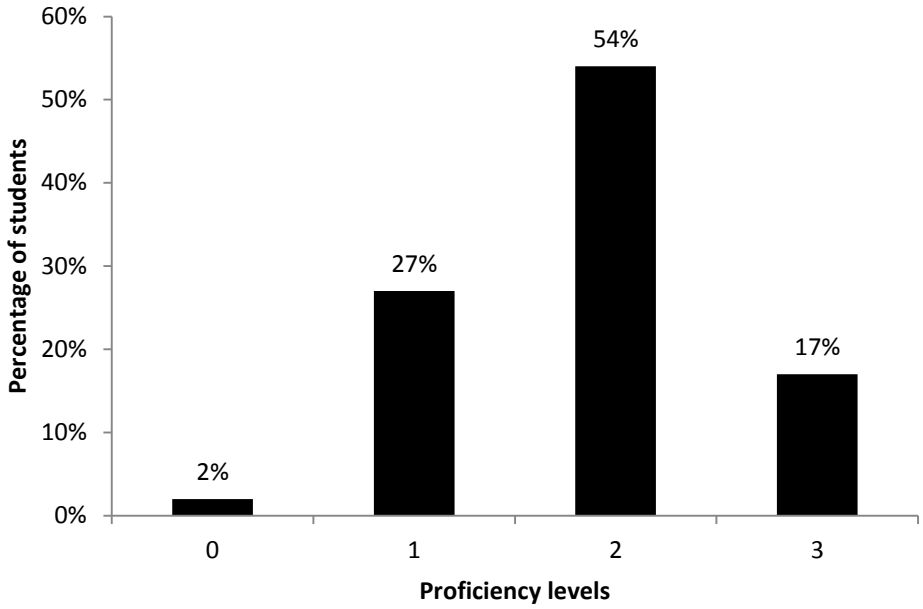


Figure 5. Distribution of fourth graders across proficiency levels: Study 1

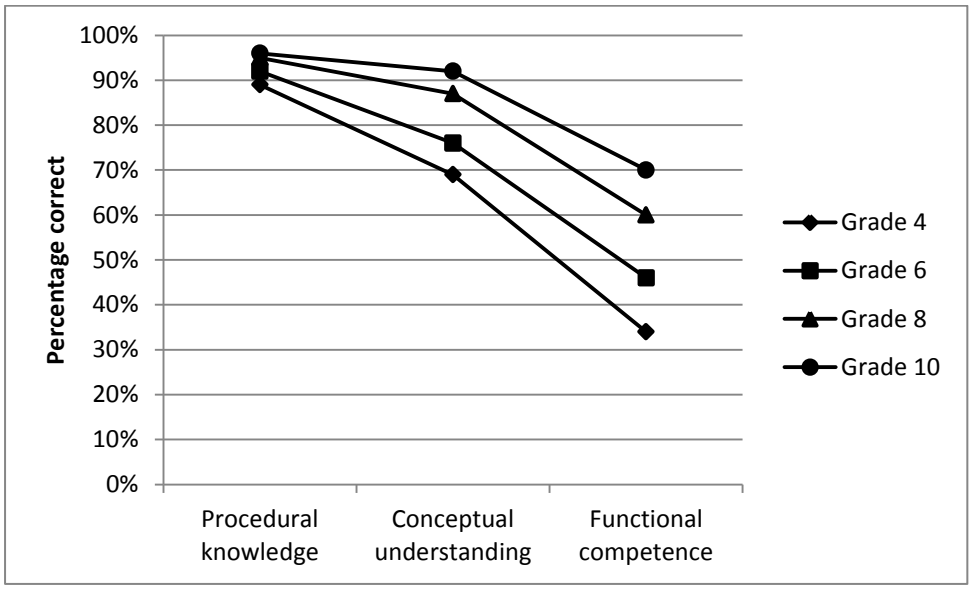


*Figure 6.*Percent of *Procedural*, *Conceptual* and *Functional*-level items solved correctly at each grade: Study 2

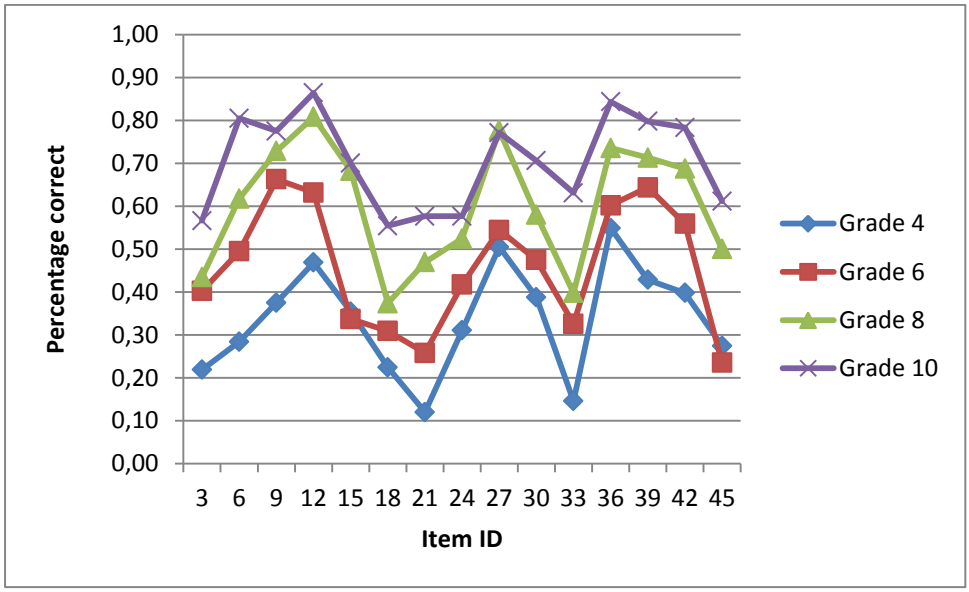


Figure 7. Pattern of item difficulty for Level 3 items across grades: Study 2

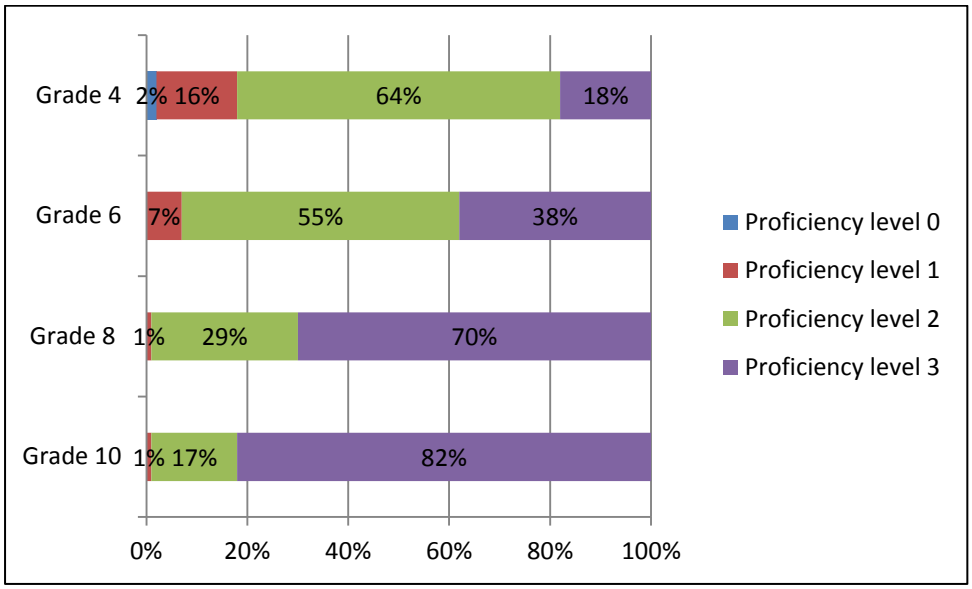


Figure 8. Distribution of students across proficiency levels: Study 2