Psychology Science Quarterly, Volume 50, 2008 (3), pp. 363-378

Rule-based item design of statistical word problems: A review and first implementation

HEINZ HOLLING[1], HELEN BLANK, KAROLINE KUCHENBÄCKER & JÖRG-TOBIAS KUHN

Abstract

Although a large body of research has been published with respect to arithmetic word problems, few studies have investigated statistical word problems in detail. This article therefore pursues two goals: Firstly, a review of current design and analysis of statistical word problems is provided. Secondly, results from a pilot study with systematically-designed statistical word problems are reported. Using the linear-logistic test model (LLTM) as well as the latent regression LLTM, we have found that the postulated cognitive model fits the data well overall. The study provides evidence that statistical word problems can be designed and analysed in a systematic way, and that the proposed cognitive model of solving statistical word problems can be successfully used in future assessments.

Key words: Statistical word problems, rule-based item design, Rasch model, linear-logistic test model, latent regression LLTM

[1] Prof. Dr. H. Holling, Chair of Statistics and Methods, Westfälische Wilhelms-Universität Münster, Fliednerstr. 21, 48149 Münster, Germany; email: holling@uni-muenster.de, phone: +49 251 – 83 38485, fax: +49 251 – 83 39469

Introduction

Statistical reasoning plays a key role in the social and natural sciences. For example, by translating a scientific question into a statistical hypothesis, a researcher can compare his theoretical expectations with empirical findings. The ability to understand and correctly handle statistical methods therefore lies at the heart of scientific work. Furthermore, statistical reasoning is of major importance in many real-life situations. For example, there are numerous everyday problems that can be solved by applying Bayes' theorem (Sedlmeier & Gigerenzer, 2001). Hence, a systematic assessment of statistical reasoning ability bears great relevance in explaining and predicting problem-solving behaviour in scientific as well as in applied settings.

Statistical reasoning is usually assessed by administering word problems with statistical content (e.g., Dimitrov, 1996). Using statistical word problems is straightforward as they imply the understanding and handling of statistical expressions and thus are good operationalizations of statistical reasoning in science and real life. Although statistical reasoning entails computation, the core of the ability pertains to building a mental model of the problem situation which needs to be translated into an equation that can be solved. Therefore, verbal information has to be transformed into a mathematical expression. While formalizing the problem structure it is necessary to focus on task-relevant data and parameters and to suppress irrelevant information. By solving the derived mathematical expression, a solution to the word problem under examination can be obtained. A typical application of statistical word problems pertains to the assessment of achievement in statistics courses at universities and schools (e.g., Jonassen, 2003).

Although some research has been conducted within this area, a framework for systematically designing statistical reasoning items still does not exist. Especially an explicit connection between cognitive processes necessary for correctly solving an item with systematic test design has not yet been established, although this can be considered critical for establishing valid measurement (Borsboom, Mellenbergh, & van Heerden, 2004).

The remnant of the current paper is organized as follows: First, a review of current research and models pertaining to solving algebra word problems is given, which bear many similarities to statistical word problems. Because little research was conducted considering statistical word problems, information from this related area has been used and, where possible, transferred to the field of interest. Furthermore, an overview concerning factors affecting the difficulty of word problems is provided and the rule-based design of a statistical word problems test is described. Finally, a pilot study is presented in which rule-based item design was utilized in order to assess the difficulty of cognitive components in statistical word problems.

Cognitive models for solving algebra word problems

In an important paper, Mayer (1981) established a framework for classifying algebra word problems from ten standard algebra textbooks used in California secondary schools. Based on the underlying source formulae (e.g., rate × time = distance), he classified these problems into eight families which could then be subdivided into problem categories consisting of different templates. Mayer showed that learning to solve algebra word problems requires the acquisition of large amounts of domain-specific knowledge (Mayer, 1987). During the solution process, both mathematical and real-world knowledge retrieved from long-term memory have to be combined with the problem context in working memory in order to develop a solution plan (Koedinger & Nathan, 2004).

Different cognitive models describing this solution process have been developed. Because algebra word problems consist of both a mathematical and a semantic part, both mathematical ability and verbal comprehension are required for a correct solution. The solution process has been subdivided in a different number of substeps which were often named differently by diverse authors (e.g., Briars & Larkin, 1984; Jonassen, 2003; Riley & Greeno, 1998).

Generally, solving algebra word problems requires the construction of a conceptual model of the problem. This model integrates the situational story content with the understanding of the semantic structure based on the mathematical principles in the problem (Jonassen, 2003). Moreau and Coquinviennot (2003) showed that the understanding of word problems leads to the construction of two complementary levels of representation. On the one hand, the elements which are indispensable for solving the problem are specified (problem model = PM). On the other hand, agents, actions and events in everyday concepts are represented (situation model = SM). Moreover, Hall, Kipler, and Wenger (1989) analyzed the quantitative and situational structure of algebra word problems based on written problem-solving protocols. Their results showed that comprehension and solution of algebra word problems are complementary activities.

Briars and Larkin (1984) proposed a computer-based model of word problem solution called CHIPS (Concrete Humanlike Inferential Problem Solver) emphasizing problem-solving procedures. The tasks in their paper were limited to addition and subtraction problems involving visualized sets of discrete objects. The CHIPS model was able to solve many common word problems by utilizing representations of physical counters. More difficult problems required augmenting this procedure with the information that one object is a member of both a set and its superset as well as with the knowledge that processes can be "undone" and that subsets are interchangeable. The authors characterized problems by the kind of knowledge the model used to solve them.

Nathan, Kintsch, and Young (1992) used a tutoring approach derived from a model of problem comprehension. According to these authors, in order to solve an algebra word problem, a subject must derive propositional and situational information and compose critical inferences. Next, this information needs to be coordinated with known problem models such that formal mathematical operations can be applied, and the exact solution can be found. According to Nathan et al. (1992), this task is highly reading-oriented. Consequently, poor text comprehension and an inability to access relevant long-term knowledge lead to serious errors. In line with this result, Cummins (1991) found that word problem solution errors are caused by misinterpretations of certain verbal expressions commonly used in problem texts. The major theoretical claim made by Nathan et al. (1992) is that based on the understanding of a word problem, a subject must establish a link between formal mathematical equations and her informal understanding of the situation described in the problem. A necessary assumption of their theory is that subjects are capable of understanding the stories given and that they form an appropriate situation model that can be mapped onto mathematical equations.

Furthermore, Nathan et al. (1992) claimed that a systematic approach to solve word problems is teachable. In order to investigate this issue, they designed ANIMATE, a learning environment that consists of two parts: Firstly, it requires the student to construct an explicit, graphical representation of the conceptual problem model (the algebraic problem schema) before deriving a corresponding equation necessary to solve the problem. Secondly, it links the formal domain of algebra with the given situation model in the real world using an animation to illustrate the conclusions implied by the student's problem representation. Nathan et al. (1992) compared groups of students who participated in different tutorials. Their results showed that by working in an environment that encourages situation-based reasoning as a normal part of the solution process, students would demonstrate an improved ability to make correct inferences. Additionally, situation-based reasoning, which was supported by exposure to the ANIMATE learning environment, helped students generate equations from texts and vice versa. The correspondence between the algebraic representation and the simulation was the main reason for successful problem-solving.

Riley and Greeno (1998) developed information-processing models of different levels of knowledge for understanding the language used in texts of arithmetic word problems, for forming semantic models of the situations described in the texts, and for making the inferences needed to answer the questions posed in the problems (Riley & Greeno, 1998). In the simplest cognitive models, inferences were limited to properties of sets that exist in a semantic model. In more complex cognitive models, relations between sets were represented internally, thereby supporting more complex reasoning.

The word problem solution model proposed by Mayer (1981) was modified in an important paper by Sebrechts, Enright, Bennett, and Martin (1996). The modified model consists of four steps, the first one being problem translation. When reading a word problem, subjects must use linguistic knowledge to translate the givens and goals into their own terms. In addition, they must often use a wide range of factual and commonsense knowledge. The second step pertains to problem integration. Some constraints and the relations between the problem elements are often implicit in the problem situation. A challenge in solving algebra word problems lies in uncovering these implicit relations and constraints and in organizing them into a larger structure. Word problems can be organized firstly by the category into which they fall (based on schemata or familiarity), secondly by the level at which they are stored in memory (categorization on the basis of structural or surface features), and thirdly by how the underlying structure is represented (formulae and equations or situational structure). The third step includes solution planning and monitoring. For any given problem there are multiple approaches to a solution. The chosen solution plan will depend on how the problem has been translated, available schemata, and the kind of strategic knowledge a subject has stored in memory. Both the effectiveness of the solution plan as well as the accuracy with which actions are executed need to be monitored. Since planning and monitoring are superordinate to other aspects of problem-solving, it is unlikely that many attributes of a word problem will have a unique impact on these specific problem-solving activities. One exception may be the nature of the problem goal, which can be classified as either a quantity or as an expression containing a variable. The presence of a variable in a goal would seem to make it more difficult to evaluate a result. The final step is solution execution. Once a sequence of steps has been planned, the solution must be implemented by executing those steps. In general, this consists of a series of computations as well as symbolic manipulations. It is generally considered desirable to minimize the effects of such low level procedural errors because they are thought to be poor indicators of quantitative reasoning.

The model proposed by Sebrechts et al. (1996) offers a number of implications relevant for the design of word problems. The impact of purely linguistic factors and computation errors should be reduced. Important factors that can be used in order to manipulate item difficulty are the category of the problem, its storage level, the representation of the underlying structure, and the nature of the problem goal.

Factors affecting the difficulty of word problems

Rule-based item design, which integrates findings from cognitive psychology with psychometric theory, is becoming increasingly common in psychological assessment (Irvine & Kyllonen, 2002). One of the core advantages of rule-based items design pertains to the fact that by making explicit cognitive processes necessary for item solution, the construct validity of the items can be tested, and items testing specific sub-abilities can be designed. Further, large pools of items can easily be designed and need not necessarily be calibrated in order to assess a subject's ability (Embretson, 1999). An IRT model that is suitable for analyzing the difficulty of cognitive processes in rule-based item is the LLTM. Closely related to the cognitive models that give explanations of how an item is solved are factors that influence item difficulty. Item properties that significantly affect item difficulty must affect the cognitive processes that occur during the solution process. Systematic variation of item properties in a test yields items differing in complexity, thereby establishing the necessary information for the estimation of a subjects' ability.

One well-designed study using algebra word problems was conducted by Enright and Sheehan (2002). They demonstrated three useful dimensions for understanding performance differences in solving algebra word problems: mathematical complexity, context, and "algebraicness". In this study, complexity referred to characteristics like number of operations, number of constraints and number of levels of parentheses. Context was varied by features like DRT (rate × time = distance), cost per unit or probability. "Algebraicness" referred to the possible manipulation of variables (i.e., ”T-Shirts that usually cost $8.00 per case are on sale for $6.00 per case. How many cases can John buy on sale for the price he usually pays for x cases?”). IRT and regression analyses showed that DRT items which required using variables were more difficult than those which did not. Among the items without operations on variables, the items with a cost context were significantly easier than the items with a DRT context (Enright & Sheehan, 2002).

In a related study, Lane (1991) used restricted item response models for examining item difficulty. She developed items in three contexts (DRT, interest, area) and varied the difficulty by the systematic modification of the item content. The items were manipulated by the number of assignments and relational propositions, the number of values that had to be derived, the amount of value derivation, whether the unknown needed to be manipulated, and whether the context was familiar. The results showed that items with a familiar context were easier and complex items were more difficult than simple ones.

Arendasy (2004) examined different restricted kinds of simple word problems (change word problems and compare word problems, respectively). A further study dealt with the automatic generation of quantitative reasoning items, in which the algebra word problems were technically constructed by the item generator Agen (Arendasy, Sommer, Gittler, & Hergovich, 2006). Arendasy based the item generation process on a set of pre-existing templates containing information on how to design each item. Arendasy's procedure can be considered a restricted sub-approach of the more general production of item variants based on a predefined set of radicals and incidentals. "Radicals" refer to item properties that affect item parameters such as item difficulty systematically. "Incidentals" do not affect item parameters and can be compared to surface characteristics (Irvine & Kyllonen, 2002). Items having the same structure of radicals can be considered "isomorphs". The radicals utilized by Arendasy et al. (2006) were the criterion typicality of the cover story, the number of partial equations, and the total number of unknown elements in the equations. However, no information could be obtained on how this item generator is supposed to work. Therefore, an evaluation of it is not possible.

From the studies mentioned above it can be concluded that the effect of the content of algebra word problems may be strong even for experienced problem solvers. The explanation for this effect relies upon familiarity. In many domains, the content of a problem (i.e., its surface cover story) provides useful clues as to the type of problem and to its solution. Blessing and Ross (1996) showed that solution probability, as well as problem categorization and determination of information relevance, are related to how typical the problem content is for its deep structure. A problem's deep structure refers to Mayer's (1981) classification of word problem families. Whether the content matched the deep structure or not depended on whether or not the typical cover story identified by Mayer was used for it. However, not in every case a significant influence of familiarity was found (e.g., Vlahovic-Stetic, 1999). The selection of non-familiar items for the comparison may at least partly account for that.

As can be seen, in general the main factor determining item difficulty was task complexity. However, the implementation of complexity differed vastly between studies. In addition to variations in the denomination of factors, until now there seems to be no consistent theoretical approach on how to define and implement complexity. Furthermore, there is a strong dependence on the type of item. Sometimes language aspects, the type of item, and its structure were regarded as complexity components and complexity was completely restricted to mathematical complexity. But even then, and even when the same type of item was used, the implementation usually differed.

As discussed above, an adequate statistical model for analysing the difficulty of item properties is the LLTM, which already has been applied in some studies concerning mathematical tasks (Fischer, 1973; Cisse, 1995; Dimitrov, 1996). The foundation for the application of the LLTM to such complex tasks as word problems was established by Fischer. He showed that the LLTM is generally appropriate for investigations in the area of instructional research if the students' tasks are solved by a certain number of cognitive operations (Fischer, 1973). Fischer analyzed problems in elementary differential calculus and showed that the difficulty of a task was not determined by repetition of the same operation within one problem but by the combination of different operations.

Cisse (1995) used a broader definition of complexity that was not restricted to mathematical complexity. He utilized the LLTM to explore the influence of six problem complexity features (part-whole, double role counters, comparative items, action cues, and language complexity) on the difficulty of addition and subtraction arithmetic word problems. The complexity factors were divided into two different sets: a logico-mathematical and a linguistic one. The full cognitive model, consisting of all six complexity factors, had significantly more predictive power than each of the two submodels consisting of the two different sets of complexity factors. But only three complexity factors contributed to problem difficulty:

knowledge of part-whole relationships, consistency of language, and double-role vs. single-role counters.

Regarding statistical word problems, little research has been conducted. A study by Dimitrov (1996) analyzed test data from university examinations on basic statistics as well as intermediate algebra courses. He utilized the linear-logistic test model (LLTM; Fischer, 1973) in order to assess the relative difficulty of cognitive components operating in these two task types. Dimitrov (1996) found nine relevant cognitive components for the statistical tasks and thirteen components for the algebraic ones. However, the items analysed in the study by Dimitrov (1996) were not designed according to a predefined set of complexity factors, because they were based on a pre-existing pool of items. Hence, the cognitive task components as well as their respective difficulty were analyzed in a post-hoc fashion.

Other studies in the field of statistical word problems dealt with the influence of frequency format versus probability format on task difficulty (e.g., Cosmides & Tooby, 1996; Evans, Handley, Perham, Over, & Thompson, 2000; Gigerenzer & Hoffrage, 1995). Evans et al. (2000) disagreed with the often-cited claim that frequency formats facilitate the correct solution of statistical word problems. In their experiments, frequency formats did not generally lead to better performance. Instead, they demonstrated that problem solving in statistical word problems is influenced by subtle variations in presentation of task information and format of the question. These results emphasize the need of systematic item construction. Different solution probabilities should not depend on superficial item formulations but on core difficulty factors which can be controlled during item construction.

In addition to systematic item properties as described above, other factors may affect item difficulty as well. The environment, including task instruction, may be such a factor. Xin (2007) examined the effect of learning opportunities in standard textbooks by analyzing item difficulty in relation to word problem distribution in adapted textbooks. He concluded that although task variables may determine the difficulty level of word problems, an instructional environment in which problem solving skills are developed may change the students' ability to tackle difficult tasks and make a difficult task an easier one. Person properties may also play an important role. Evidentially, the age of the subjects is crucial, especially as samples consisted of students in most studies. The impact of cognitive development phases may outweigh the impact of item features or interact with it (Wilkening, 1981).

Test design

Based on the literature review above, item properties affecting difficulty of statistical word problems were identified. Table 1 lists the item properties varied in this study. In addition, the most popular mathematics textbooks for schools were inspected. Word problems were searched and classified (e.g. combinatorics, dependent probabilities, independent probabilities). This information was taken into consideration during item design in order to develop a test of curricular relevance. We chose independent probability tasks for our pilot study because of two practical reasons. Firstly, this kind of task is taught at an early stage in school and therefore we were able to test more students. Secondly, the independent probability items can be used as the groundwork for designing dependent probability tasks in further steps.

The factors had to be implemented as rules for the construction of the items. Therefore, hierarchical templates were formed. The basic framework depended only on context.

Through the combination of complexity factors, a general formula was constituted. The terms of the formula in combination with the number format determined which sentences were added. Within each sentence, there were replacement characters whose values could be filled with irrelevant information if necessary. After the general design principles were established, an iterative process of item design, revision, testing, and again revision began.

The test consisted of independent probability items designed according to the five general factors shown in Table 1. Complexity comprised four construction parameters: the logical operators and, or, complementary, or rearranging. As can be gleaned from Table 2, there were eight factors with two levels each. Figure 1 gives an example of a typical statistical word problem from the test.

Table 1:
Overview of item difficulty factors manipulated in the current study

Factor	Description	Example
Context	Cover story given	Drawing cards, drawing balls, catching the murderer
Numbers	Number format of numerical elements	The probability is .5, 5 out of 7 balls
Complexity	Number of operations, number of constraints	
Irrelevant information	Information in the item that does not pertain to the solution	
Unknown variables	Unknown variables instead of numbers are presented in the items	The probability is X

Table 2:
Possible values of item difficulty factors

			Complexity					
Parameter	**Context**	**Numbers**	**Or**	**And**	**Complementary**	**Rearranging**	**Irrelevant**	**Unknown**
Possible values	Balls/ Murderer	Probability/ Absolute	Yes/No	Yes/No	Yes/No	Yes/No	Yes/No	Yes/No
Example (Figure 1)	Murderer	Probability	Yes	Yes	Yes	No	No	Yes

Figure 1:
Example item

> A murder was committed. The police want to carry out a DNA investigation because traces of DNA were found. Therefore, all subjects of a circle of potential offenders shall give a saliva sample. As this is very expensive, the police want to narrow down the circle so that only certain people are included in the DNA investigation. Relatives, friends and strangers are considered. All of them may either come from the same town as the victim or from a different town. The probability that a murder is committed by a relative is W. The probability that a murder is committed by a friend is X. The probability that a murderer comes from the same town as the victim is Y. How can the probability be specified that the police does not catch the murderer if exclusively all relatives and friends of the victim who live in the same town as the victim are tested?

Method

Participants

192 students from five different German grammar schools participated in the study (grades 11 to 13 and prospect students). We tested 119 male (62%) and 73 female (38%) students. The mean age was 17.48 (range 15 to 27). Subjects had to be excluded from further analysis if they answered all items or none correctly (20 subjects altogether), because the estimation method used was not suitable for dealing with those subjects. Furthermore, a residual analysis was conducted, which was based on a model with fixed effects for the person estimates (equal to logistic regression). Four subjects showed very large residuals accompanied with high values for the GLM equivalent of Cook's distance. These subjects were excluded as well. Hence, the final analysis was based on 168 subjects.

Materials and statistical analysis

11 independent probability problems were administered to the students. Additionally, the students answered some demographical questions. The test took about 45 minutes. Students were tested in their classrooms in groups of 15 to 30 pupils. At the beginning of the test, they received an instruction about independent probabilities. They read this instruction for seven minutes, but they could continue using it during the test. In the main part of the test, students were allowed to work as long as three minutes per task (33 minutes overall). They were not allowed to use a calculator. A month later, every student received a written feedback regarding the individual test result obtained.

Item 1 was used as an introductory item. It was rather simple and intended to help the subjects understand the general approach that had to be used. Therefore, it was not included in the analysis. Hence, 10 items per person remained. The design matrix of those items is shown in Table 3. The basic parameters for the items were "context", "number", "or", "and", "complementary", "rearrangement", "unknown", and "irrelevant".

Table 3:
Design matrix of the items

Item Nr.	Context	Number	Or	And	Complementary	Rearrange	Irrelevant	Unknown
1	1	0	1	0	1	1	0	0
2	0	0	0	0	1	1	1	1
3	0	1	1	0	0	0	0	1
4	0	1	1	1	1	1	1	0
5	1	1	0	0	1	0	1	0
6	0	0	1	0	0	1	1	0
7	0	0	0	1	1	0	0	0
8	1	0	0	1	0	0	1	0
9	1	0	1	1	1	0	1	1
10	1	1	0	1	0	1	0	1

In addition, person predictors were included. We therefore investigated whether the grade, which is related to age, had an effect on item difficulty, our assumption being that students in higher grades would obtain better test scores. Furthermore, we wanted to check whether the order of items had an effect on item difficulty, i.e. whether an item was easier if it was processed at the end of the test rather than at the beginning. Hence, the same set of items was presented to different groups of students in two different orders. In order to analyse the effects of all predictors mentioned, we first checked Rasch scalability of the items, and then computed both an LLTM (item predictors only) and a latent regression LLTM (item and person predictors; see De Boeck & Wilson, 2004) and compared model fit using likelihood-ratio tests.

Results

The average solution frequency of the items was .32. Merely item 4 was extremely difficult with a relative solution frequency of .04. The frequencies of the other items ranged from .22 to .59. Cronbach's alpha was .69, which is below the usual .80 boundary due to the shortness of the scale. In order to assess the Rasch scalability of the items, we computed the Q-Index for each item (Rost & von Davier, 1994). No significant deviations from the Rasch model were found. The Rasch model, as could be expected, fit significantly better than the LLTM, $\Delta\chi^2(2) = 67.87$, $p < .00$. In a next step, we computed both the LLTM and the latent regression LLTM (see Table 4).

Table 4:
Parameter estimates in the LLTM and latent regression LLTM

	LLTM		**Latent regression LLTM**	
Parameter	**Estimate**	**SE**	**Estimate**	**SE**
Context	**0.28**	0.13	**0.28**	0.13
Number format	**-0.66**	0.13	**-0.66**	0.13
Or	-0.02	0.12	-0.02	0.12
And	**-1.42**	0.14	**-1.43**	0.14
Complementary	**-0.62**	0.12	**-0.62**	0.12
Rearrange	**-0.70**	0.13	**-0.70**	0.13
Irrelevant	**0.64**	0.13	**0.64**	0.13
Unknown	0.03	0.12	0.03	0.12
Grade 11			**0.38**	0.19
Grade 12			**1.44**	0.22
Order			-0.22	0.18
Intercept	0.26	0.22	-0.14	0.27

Note. Significant coefficients are bold-faced ($p < .05$).

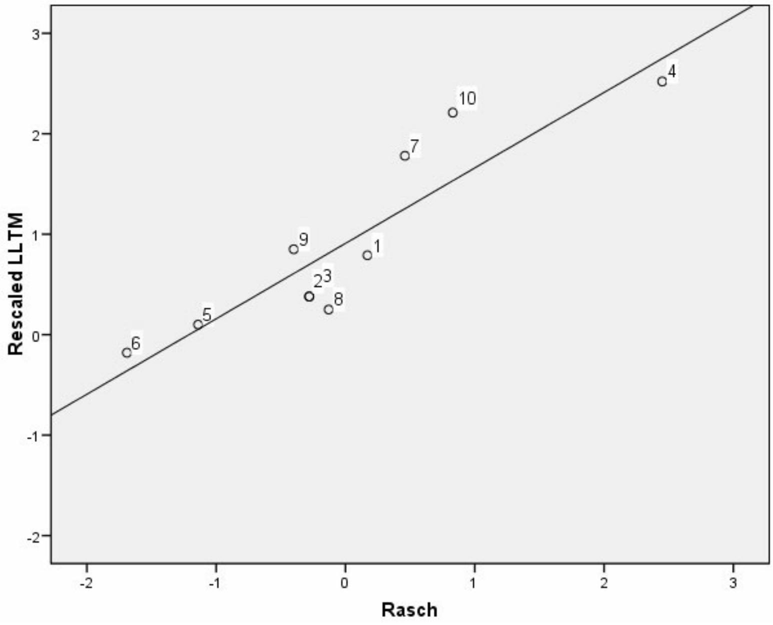
The estimations for the basic parameters in both LLTM and latent regression LLTM are shown in Table 4. The estimations for "or" and "unknown" are very low. The parameters do not seem to be very influential whereas all other parameters are significant. Especially "and" is a promising factor with an estimate of -1.42 and a standard error of .14. In addition, the person predictor grade seems to be important. Comparisons of grades 11 and 12 with grade 13, respectively, which are given by the parameters in Table 4 both reach significance. As expected, the latent regression LLTM fits the data significantly better than the LLTM, $\Delta\chi^2(3) = 42.65$, $p < .00$. Finally, as hypothesized, there is no evidence that item order plays any role. The LLTM-predicted (rescaled) item difficulties are very close to those estimated by the Rasch model, with a correlation of $r = .90$ (see Figure 2).

Discussion

Approaching complex tasks with systematic analysis and design procedures is a relatively new and important area in developing tests. The difficulties lie in the nature of the tasks. A high number of possible factors and difficulties in systematizing the items' components are the main obstacles. However, the widespread usage in many real life domains and the high ecological validity of these tasks document the relevance of research in this area. Word problems are a particularly important member of complex tasks due to their applicability in schools and universities as well as their capability of measuring applied, creative, logical, and mathematical abilities at the same time.

The literature review presented above showed that a variety of approaches were used to gather information about tests dealing with word problems. The sequential model by Sebrechts et al. (1996) was very influential and helpful in the course of this project. The identi-

Figure 2:
Rasch item difficulties and rescaled LLTM item difficulties



fied cognitive steps apply to word problems in general. The factors hypothesized to affect item difficulty in this study can be related to the cognitive steps outlimited by Sebrechts et al. (1996). When it is known which phase is affected by which factor, a better theoretical understanding of the test can be achieved. Aspects of the model were taken into account while developing the items used here. For example, items were created with similar wording in order to minimize language effects on understanding. Furthermore, an evaluation of individual difficulties with particular cognitive components (see the end of this section) is possible because of this model.

The pilot study reported here was supposed to explore possible factors and ways to construct such complex tasks systematically. By choosing statistical word problems, a practically important task type in an area less well researched was examined. Up to now, there has been no rule-based design of statistical word problems based on predetermined factors accounting for task difficulty. We showed that the LLTM is a suitable model for rule based generation of probability word problems. Our pilot study demonstrated that statistical word problems can be generated based on predetermined difficulty factors.

With regard to the parameters, “or” and “unknown” did not have an effect. The “or” operation itself was relatively easy. However, it is a basic construction component of the items and from a practical point of view in the context of probability, theory knowledge about this operation is crucial. A set of items would be incomplete without it. Also, some of the other operations such as rearrangement are necessarily based on an additional operation like “or”, “and”, and “complementary”. Thus, “or” can be used for items that include rearrangement without additionally increasing the difficulty.

The low estimate for “unknown” was unexpected. In other studies, it increased item difficulty considerably (e.g., Lane, 1991). A possible explanation might be that the entire test was very systematic. All items had the same general structure. In order to be successful, relevant parts of the items had to be found and the rules learned at the beginning had to be applied. A less consciously systematic solution strategy would hardly be successful regarding the high complexity of the items and the similarities between them. It appears that if a systematic solution approach is utilized, dealing with unknown elements is not very difficult.

If the complexity factors “and”, “Complementary”, and “rearrangement” are present, difficulty increases. “Complementary” and “rearrangement” change difficulty to a medium extent whereas items with “and” are considerably more difficult than the ones without. Moreover, because of this variation in parameters, they are very promising for usage in later test design. Items with a wide range of difficulties can then be easily designed.

Furthermore, “context” and “numbers” make a difference. The murder story is slightly easier than the drawing balls story. We assume that the latter might be easier because of higher familiarity. This cover story was found several times in all school textbooks. However, on the other hand, the murder story might be more attractive because of its uncommonness. Also it might seem more realistic to students. When the numbers are given as absolute instead of relative values, the difficulty increases. This is reasonable because, firstly, an additional calculation is necessary to derive relative values and, secondly, in order to make this calculation the concept of relative frequencies in probability, theory has to be understood. This was not self-evident to all students as some solution sheets showed. As this is a central concept in probability theory, “Numbers” might be a very gainful factor in word problem construction.

Finally, irrelevant information made items easier. This was an unexpected result. Irrelevant information must be suppressed while solving word problems, which usually increases work load. But, this was not the case in our study. An explanation might be that the additional information helped understanding the story. The information given was in no case wrong. It simply could not be used directly to solve an item. This would imply that more information leads to better understanding, even if only contextually. In conclusion, the information pertaining to the estimated basic parameters is valuable for future item design. Furthermore, the differences between significant and non-significant basic parameters appeared to be very large in most cases, thereby allowing the design of items with a wide range of difficulty.

The results showed that in addition to item predictors, the person predictor grade also affects solution frequencies. It is to be assumed that other person parameters (e.g., age and type of school) affect solution frequencies as well. This is relevant insofar as it provides insights for theories pertaining to group-specific differences in cognitive processes. There seemed to be differences between grades. In grade 13, the probability to solve an item was higher. There can at least be two reasons for this. Higher grades might have had more previous knowledge concerning probability theory or more general knowledge about mathematics or word problems that also helped solve the items. On the other hand, the higher age of the pupils in a more developmental sense might be responsible. We cannot distinguish between these two potential explanations based on the given data.

Finally, there did not seem to be an effect of the order of the items, an effect not necessarily anticipated. For example, it is known that the motivation of subjects decreases with increasing test length. With an altered sequence of difficult and easy items, different scores might be obtained. However, the test under investigation in this study was rather short, thereby possibly diluting this effect.

There are some possible limitations to this study that shall briefly be mentioned. The results obtained differ between samples, schools, and classes. We noticed that different teachers, schools, schoolbooks, and school curricula have a huge impact on students' performance (e. g. Xin, 2007). The necessary condition for applying the LLTM is that the subjects know the problem type and have learned an appropriate solution strategy in advance. Our short introduction given before the test could only refresh the students' knowledge about probability problems and did not aim to explain a completely unknown topic. Unfortunately, probability problems are often neglected during math lessons even though they are essential in a vast field of studies. Another pre-condition for our testing was that students solving our tasks were motivated during the test. Without a minimum of motivation, even simple items cannot be solved.

However, this study is just the first step in a project. The final goal of the project is to create a computer program producing automatically constructed items that are presented as an adaptive test. By corroborating the cognitive theory pertaining to statistical word problems advanced in this paper, specific cognitive components impeding item solution in individual students could be identified. Thus, the final assessment of a subject's ability will not only include an ability estimation, but also information about which cognitive components are understood by the subject and which are not. Such a test can be used not only for final achievement assessments but also as a learning tool. The components that are still difficult for a subject can be repeated and the test can be distributed again to evaluate learning improvements.

Acknowledgments

This research was supported by the research grant Ho 1286/5-1 of the German Research Foundation.

References

Arendasy, M. (2004). Dimensionalität und differenzielle Validität von Textaufgaben: Zum Einfluss von Bearbeitungsstrategien [Dimensionality and differential validity of word problems: The effect of strategies]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 18*, 231-243.

Arendasy, M., Sommer, M., Gittler, G., & Hergovich, A. (2006). Automatic generation of quantitative reasoning items. *Journal of Individual Differences, 27*, 2-14.

Blessing, S. B., & Ross, B. H. (1996). Content effects in problem categorization and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 792-810.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061-1071.

Briars, D. J., & Larkin, J. H. (1984). An integrated model of skill in solving elementary word problems. *Cognition and Instruction 1*, 245-296.

Cisse, D. (1995). *Modeling children's performance on arithmetic word problems with the linear logistic test model*. Unpublished doctoral dissertation, University of Alberta, Canada.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*, 1-73.

Cummins, D. D. (1991). Children's interpretations of arithmetic word problems. *Cognition and Instruction, 8*, 261-289.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Dimitrov, D. M. (1996). *Cognitive item subordinations in linear logistic test modelling*. Unpublished doctoral dissertation, Southern Illinois University at Carbondale, USA.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*, 407-433.

Enright, M. K., & Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items: Implications for item generation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Erlbaum.

Evans, J. S. B. T., Handley, S., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition,* 77, 197-213.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684-704.

Hall, R., Kibler, D., & Wenger, E. (1989). Exploring the episodic structure of algebra story problem solving. *Cognition and Instruction, 6*, 223-283.

Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development*. Mahwah, NJ: Erlbaum.

Jonassen, D. H. (2003). Designing research-based instruction for story problems. *Educational Psychology Review, 15*, 267-296.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences, 13*, 129-164.

Lane, S. (1991). Use of restricted item response models for examining item difficulty ordering and slope uniformity. *Journal of Educational Measurement, 28*, 295-309.

Mayer, R. E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. *Instructional Science, 10*, 135-175.

Mayer, R. E. (1987). Learnable aspects of problem solving: Some examples. In D. E. Berger, K. Pezdek & W. P. Banks (Eds.), *Applications of cognitive psychology: Problem solving, education, and computing* (pp. 109-122). Hillsdale, NJ: Erlbaum.

Moreau, S., & Coquin-Viennot, D. (2003). Comprehension of arithmetic word problems by fithgrade pupils: Representations and selection of information. *British Journal of Educational Psychology, 73*, 109-121.

Nathan, M. J., Kintsch, W., & Young, E. (1992). A theory of algebra-word-problem comprehension and its implications for the design of learning environments. *Cognition and Instruction, 9*, 329-389.

Riley, M. S., & Greeno, J. G. (1998). Developmental analysis of understanding language about quantities and of solving problems. *Cognition and Instruction, 5*, 49-101.

Rost, J., & von Davier, M. (1994). A conditional item fit index for Rasch models. *Applied Psychological Measurement, 18*, 171-182.

Sebrechts, M., Enright, M., Bennett, R. E., & Martin, K. (1996). Using algebra word problems to assess quantitative ability: Attributes, strategies, and errors. *Cognition and Instruction, 14*, 285-343.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130*, 380-400.

Vlahovic-Stetic, V. (1999). Word problem solving as a function of problem type, situational context and drawing. *Studia Psychologica, 41*, 49-62.

Wilkening, F. (1981). Integrating velocity, time and distance information: A developmental study. *Cognitive Psychology, 13*, 231-247.

Xin, Y. P. (2007). Word problem solving tasks in textbooks and their relation to student performance. *The Journal of Educational Research, 100*, 347-359.