Studies in Educational Evaluation 35 (2009) 71–76

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Automatic item generation of probability word problems

Heinz Holling *, Jonas P. Bertling, Nina Zeuch

Department of Psychology and Sport Studies, University of Münster, Fliednerstr. 21, D-48149 Münster, Germany

ARTICLE INFO

Keywords:
Rule-based item design
Probability word problems
LLTM
Student evaluation
Evaluation methods

ABSTRACT

Mathematical word problems represent a common item format for assessing student competencies. Automatic item generation (AIG) is an effective way of constructing many items with predictable difficulties, based on a set of predefined task parameters. The current study presents a framework for the automatic generation of probability word problems based on templates that allow for the generation of word problems involving different topics from probability theory. It was tested in a pilot study with $N = 146$ German university students. The items show a good fit to the Rasch model. Item difficulties can be explained by the Linear Logistic Test Model (LLTM) and by the random-effects LLTM. The practical implications of these findings for future test development in the assessment of probability competencies are also discussed.

Automatic generation of probability word problems

Mathematical competence is regarded as crucial for success in our modern sciences. The Programme for International Student Assessment (PISA) explicates the importance of mathematical competence and defines mathematical literacy as follows:

> Mathematical literacy is an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen (OECD, 2003, p. 24).

The mathematical content of the PISA programme is captured by four overarching ideas, namely, quantity, space and shape, change and relationships, and uncertainty. Uncertainty strongly relates to stochastic content. Stochastics are very important for the social sciences as well as for our everyday life because we are often confronted with uncertainty and probabilities. Acquaintance with this uncertainty can be regarded as an important academic pre-stage of stochastic competence. However, stochastics are studied less often than most of the other mathematical content areas and are often omitted in school lessons. Nevertheless, stochastics should be considered more seriously and should be assigned a more important role in education (e.g., OECD, 2003). Therefore, the measurement of stochastic competence becomes a major subject in educational research. Stochastic competence can be measured by several means, but one very good possibility is word problems.

Mathematical word problems are often used in school and university settings to assess mathematical competencies and achieved knowledge and skills in mathematical contexts (Jonassen, 2003). This ranges from the simple translation of verbal tasks into mathematical equations to complex mathematical reasoning and transfer. Word problems in this area show a high ecological validity as they measure creative, logical and mathematical competencies at the same time.

A large body of research concerning algebra and other mathematical word problems has been conducted by several authors (e.g., Jonassen, 2003; Koedinger & Nathan, 2004; Xin, 2007). Sebrecht, Enright, Bennett, and Martin (1996) presented a widely approved cognitive model for solving algebra word problems, including the four steps problem translation, problem integration, solution planning and, as a final step, monitoring and solution execution. This model has been proved to be essential in word problem solving and provides a helpful framework for the conceptualisation and evaluation of word problems.

The rule-based design and construction of word problems is very extensive. Automatic generation of word problems provides an economic way of creating a large item pool. For reasonable rule-based and automatic item generation (AIG) it is necessary to define elements that influence item difficulty. The current study is aimed at the identification of the impact of several basic constructive components in probability theory on item difficulty and the usage of these results for AIG. First, we will give a theoretical introduction on mathematical word problems, rule-based item design and AIG. After that, the Linear Logistic Test Model (LLTM), which is used to analyse our data, is introduced and the method of the current study is explicated. We present results from a first application of the newly generated word problems to a university context and discuss the current findings as well as their implications for further research.

* Corresponding author. Tel.: +49 251 8339419; fax: +49 251 8339469.
E-mail address: holling@psy.uni-muenster.de (H. Holling).

0191-491X/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.stueduc.2009.10.004

Probability and statistical word problems

While algebra and arithmetic word problems have been the focus of many studies, only few researchers have investigated the probability theory and statistical contents of word problems (e.g., Arendasy, Sommer, Gittler, & Hergovich, 2006). Probability and statistical word problems provide information about the competence to deal with statistics and probability theory beyond equations and formulas, which means the transfer of this competence and a deeper understanding of the relations and the sense behind the numerical expressions. The model proposed by Sebrecht et al. (1996) is easily applicable to this subtype of mathematical word problems. In fact, statistical and probability theory contents are conceptually similar to algebra and other categories. But as Arendasy et al. (2006) have shown, different subtypes of mathematical word problems are qualitatively different and, therefore, cannot be described on one common conceptual dimension. This underscores the necessity to investigate probability and statistical word problems as a separate problem type. An attempt to assign the findings for algebra word problems to probability and statistical word problems should not be made without a modification to the domain of statistical and probability word problems.

Rule-based item design and AIG

Generation of word problems in statistics and probability theory is of great importance for class exercises or university course tests, especially for the bachelor courses during which students have to collect credit points through test taking. Items have to be valid and reliable in order to provide test results that give sufficient information about the students' competencies. Test items should be robust enough to withstand guessing and practice effects, especially when similar tests are taken several times or by many students. The type of word problems investigated here is also well-suited to applications in training contexts (e.g., as part of tutorials to prepare students for exams). They can increase the motivation to learn and to understand statistics and probability theory by showing how formal concepts relate to applications in science and everyday life. The challenge lies in generating word problems which are as unambiguous as possible and which are based on a definite set of definite underlying rules.

Rule-based item design in the educational and psychological testing of competencies has a relatively short history, often dealing with figural or numerical content (e.g., Enright, Morley, & Sheehan, 2002; Freund, Hofer, & Holling, 2008; Irvine & Kyllonen, 2002). The main principle is to first analyse the components of items that influence item complexity and difficulty, and to then use these components to combine them and generate items of an arbitrary complexity level. Components which are crucial for the solution process presumably influence item difficulty and should therefore be well-defined for item design, item generation and item application.

There are several great advantages of rule-based item design: once the essential underlying cognitive components are known, items can be generated and applied on this basis without calibrating every single item. Also, item validity can be tested by comparing the cognitive components supposed to influence item difficulty to the response patterns and basic parameter estimates. Items can be designed to test specific competencies by the purposeful selection of construction components. Typical mistakes which occur during an item design which has not carried out a previous analysis of crucial components (e.g., invalid items that do not assess the desired underlying competencies) can be avoided if the design procedure is clearly defined along the identified components and design steps. Rule-based item design is also helpful for adaptive testing because items can be generated automatically as soon as the essential design principles are clear. Moreover, when the underlying construction principles are made explicit, item generation can be carried out automatically by an automatic item generator. AIG facilitates the assembly of large item banks without constructing items manually, thereby preventing typical mistakes which might occur during non-automated item writing.

However, verbal content poses several difficulties. For example, verbalisation per se provides room for interpretation; slight differences in wording can indeed have a strong impact on certain item properties such as validity, difficulty and complexity (e.g., Cummins, 1991). For this reason, verbal items should be built from unambiguous phrases to avoid misinterpretation and misunderstanding. These phrases should leave space for the explication of "free variables", such as numbers, variable characteristics and surface features while keeping the core verbal formulation constant. In such a way, an unambiguous matching between underlying mathematical expressions and wording (questions whose answers require the same cognitive steps should be formulated consistently) should be maintained.

The LLTM

In order to use rules for AIG, a test model is needed that incorporates the structure of the items and allows for the assignment of parameters (i.e. difficulties of the cognitive operations involved) to each of the rules. The LLTM (Fischer, 1973) was one of the first models to take these considerations into account. Starting from the idea that item difficulty can be conceived as a function of certain cognitive operations involved in the solution process (Scheiblechner, 1972), Fischer developed the model as an elaboration of the more general Rasch Model (RM; Rasch, 1960). The RM states the probability that the person j answers item i correctly as follows

$$P(X_{ij}=1|\theta_j,\sigma_i)=\frac{e^{\theta_j-\sigma_i}}{1+e^{\theta_j-\sigma_i}} \quad (1)$$

with θ_j the ability parameter for person j, and σ_i the difficulty parameter for item i.

The LLTM focuses on the basic parameters underlying the items of a test instead of the global difficulty of each item. In the context of educational and psychological testing, basic parameters are defined with regard to the cognitive operations associated with item difficulty. From this perspective, the core assumption of the LLTM is that differences between item parameters are attributable to cognitive operations involved in one, but not in another item. What determines an item's difficulty is the number and the nature of the cognitive operations involved. In the LLTM, the items are scored on stimulus features and q_{ik} is the score of item i on stimulus feature k in the cognitive complexity model of items. Estimates from the LLTM include η_k, the weight of stimulus feature k in item difficulty and θ_j, the ability of person j. The item difficulty is described as an additive function of basic parameters:

$$\sigma_i=\sum_{k=1}^{K}q_{ik}\eta_k. \quad (2)$$

Replacing σ_i in (1) with (2) yields the probability that person j passes item i in the LLTM:

$$P(X_{ij}=1|\theta_j,q,\eta)=\frac{e^{\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k}}{1+e^{\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k}}. \quad (3)$$

In the LLTM, the person effects are usually regarded as random and the item effects as fixed. If one wants to consider the abilities of