A SOLID FOUNDATION: WHY E-ASSESSMENT SHOULD BE BASED ON A SYSTEMATIC TYPOLOGY OF TEST ITEMS

Cerstin Mahlow
School of Social Work
University of Applied Sciences Northwestern Switzerland
Olten, Switzerland

Michael Piotrowski
University of Zurich
Zurich, Switzerland

Wolfram Fenske
Eudemonia Solutions AG
Magdeburg, Germany

ABSTRACT

The Bologna Process requires, besides other changes, more frequent assessment of students, both during and after modules. As e-learning and e-teaching scenarios already play important roles in many curricula, instructors are now starting to consider e-assessment as well. To enable automated evaluation, selected-response items are typically used in e-assessment. However, current e-assessment platforms offer only a limited and rather arbitrary selection of item types. This means that the decision on which item types to use in a test is often based on purely technical issues instead of pedagogical considerations. In this paper, we argue that both implementers and users of e-assessment platforms should abstract from current implementations and base the selection of item types for e-assessment on a sound typology of test items. This would allow instructors to choose the item types best suited for a test and it would allow implementers to generalize the test facilities of their systems, reducing maintenance and development costs. As an example, we outline Rütter's 1973 typology and discuss selected issues from the point of users and implementers of e-assessment.

KEYWORDS

Assessment, Typology of Test Items, Electronic Testing

1. INTRODUCTION

For several years now, European universities have been working on the implementation of the requirements defined in the *Bologna Declaration*. One aspect of the reforms initiated by this process is the restructuring of assessment in higher education. Instead of having one single exam at the end of the studies to complete one's degrees, as it used to be common in European universities, smaller units of assessment on the level of individual modules and courses are now required.

Universities are thus faced with the challenge of finding ways to assess the achievement of learning goals. The traditional two-hour (or longer) essay-type exams are no longer practicable: Lecturers give more than one lecture each term, the assessment results have to be reported no later than a few weeks after the assessment took place, and, most importantly, there are usually more than 50 and up to 1000 students attending one

lecture. Thus, there is a clear need for assessment types that are easy to deliver to a large number of students and that can be evaluated quickly. These requirements can best be fulfilled by *e-assessment*.[1]

It takes three minutes for the water pipe to fill a bucket that can hold 15 liters. You want to fill a bucket of 20 liters.	**Context information**
How long does it take to fill the 20 liter bucket?	**Question**
❑ 5 minutes ❑ 2 minutes ❑ 4 minutes	**Response**

Figure 1. Macro structure of a test item

Assessment in higher education must fulfill general requirements with respect to quality—i.e., objectivity, reliability, and validity (Lienert and Raatz, 1998)—as well as various technical and organizational requirements: The creation of test items should be efficient, the delivery and presentation of test items to students should be easy, and, if feasible, evaluation should be automated. Other factors, such as course type, number of students, and university rules and regulations will also influence the design of the assessment process. Most importantly, however, assessment should test skills on all levels of learning.

If automatic evaluation is required, multiple-choice (MC) questions[2] are the typical solution. While it is possible to create very sophisticated MC tests, as shown by Morrison et al. (2004) and Collins (2006), among others, it is often hard for instructors to create MC tests that go beyond testing of knowledge. Tests may thus fail to assess the skills they are supposed to assess—and thus fail the quality criteria. Thus, instead of simply using MC tests "by default", alternatives should be considered. Since educational testing is a well-researched issue in the field of education (see, e.g., Bennett et al. (1990) or Osterlind (1998)), serious e-assessment should be based on insights from educational research.

When evaluating test types, one should thus take a systematic approach for finding types that can be evaluated automatically *and* that are suitable for the specific assessment goals. We have chosen (Rütter, 1973) as our starting point, since Rütter defines an extensive typology of item types, presents examples of test items, and discusses most of the proposed item types with respect to their complexity for authors and candidates. A systematic approach is a much more suitable base for a typology of e-assessment item types than the ad-hoc created categories commonly found in the literature (e.g., Osterlind, 1998; Conole and Warburton, 2005) and in e-assessment and CAA systems.

In the rest of this paper, we will first give a brief overview of Rütter's typology. We will then review the item types with respect to their suitability for electronic processing to find potential alternatives or additions to MC tests. Finally, we describe the positive effects a sound typology may have on the implementation of e-assessment and CAA systems.

2. A TYPOLOGY OF TEST ITEMS

Rütter (1973) motivates his typology by the need for a systematic compilation of item types suitable for pedagogical purposes. Despite its age, we consider Rütter's approach still an outstanding work, which is unfortunately not well known outside the German-speaking countries.

Rütter's typology is a *structural typology*, i.e., it is based only on the objectively observable structure of test items (in contrast to, e.g., the taxonomy presented in Haladyna (2004)). The macro structure of test items consists of fields for *context information*, *question*, and *response* (Rütter, 1973, p. 51) as shown in figure 1. The typology is based on variations in the response field and the degree of freedom candidates have in formulating their answers.

This section outlines Rütter's typology and summarizes the findings of Fenske (2007) regarding the suitability of different item forms in an e-assessment or CAA scenario.

[1] We use the term *e-assessment* to describe assessment processes in which *all* phases are computer-supported; if this is not the case, we speak of *computer-assisted assessment* (CAA). Our definition of *e-assessment* is thus stricter than most others (see the survey by Rüdel et al. (2007)).

[2] Both in the form of single-answer ("check the correct answer") and multiple-answer ("check all that apply") items.

Table 1. Macro and meso typology of item types

Classes:	Open Items	Semi-Open Items	Closed Items
Categories:	Free Creation	Short Answer	Identification
	Free Interpretation	Association	Alternative
	Free Association	Completion	Selected Response
		Substitution	Selected Association
		Construction	Selected Completion
		Transformation	Selected Substitution
			Selected Extension
			Matching
			Reordering
			Proxy

2.1 Macro, Meso, and Micro Level

On the *macro level*, Rütter divides items into three *classes*, based on the type of response expected from the candidate. For an *Open item*, such as an essay, neither the candidate nor the scorer have a predefined correct answer. For a *Semi-Open* item, e.g., a cloze test, the answer is not shown to the candidate but predefined for the scorer. Finally, for a *Closed item*, such as a multiple-choice test, the correct answer is predefined and shown to both the candidate and the scorer.

By analyzing the *mesostructure* of items, each class can be further divided into *categories*, based on the kind of task that has to be performed. Table 1 gives an overview of categories and classes.[3] Several categories (e.g., association) can be realized in more than one class. Similarly named item categories share the same idea, but differ in the way the response has to be given.

On the *micro level*, item types can be further differentiated into *types*, e.g., depending on the number of subquestions in the item or the number of expected answers and their relationship to each other.

2.2 Examples of Item Types

For illustration, we will classify some well-known item types according to Rütter's typology. The classic multiple-choice item belongs to the class of *closed items* and the category of *Selected Response*. Considering the micro structure (*Sequence Type*), MC items can thus be classified as *Closed Sequence Selected Response* items (Reihenantwortwahl).

A graphical identification task, such as "color the trochanter of the schematically depicted insect leg," would be a *Closed Simple Identification* item (Einfachidentifikation).

Another well-known type of items are cloze tests, often used in language learning, where candidates are asked to fill in blanks in a text with the correct words. Such items would be classified as *Semi-Open Multiple Completion* (Mehrfachergänzung); if there is a list of words to choose from, we have a *Closed Multiple Selected Completion* item (Mehrfachergänzungswahl).

2.3 Electronic Realization

Fenske (2007) evaluates Rütter's item classes, categories, and types with respect to authoring, presentation, recording, and evaluation in an e-assessment scenario. Even though Rütter (1973) predates widespread electronic testing, Fenske finds that there are no items types that completely preclude the use of computers. Computer-based authoring and presentation is possible for items of all classes. Since test items often require textual answers, recording usually poses no problem, either. Non-textual answers, such as function graphs and mathematical or chemical formulae, on the other hand, may still be easier to record on paper due to the lack of suitable editors. The evaluation phase is where the biggest differences are found. The free nature of

[3] The terms used are our own translation of the German originals used by Rütter (1973). We give the original German terms in parentheses when naming specific item types.

items of the Open class forbids automatic evaluation in most—if not all—cases. Items of the Semi-Open class can be evaluated automatically if certain conditions are met (Fenske, 2007, p. 29).

Items of the Closed class can be evaluated fully automatically. The class of Closed items offers a wide variety of types suitable for e-assessment. On the one hand, there are well-known item types, such as those presented as examples in section 2.2 (Closed Sequence Selected Response (Reihenantwortwahl), Closed Homogeneous Alternative (homogene Mehrfachalternative), Closed Simple Identification (Einfachidentifikation) and Semi-Open Multiple Completion (Mehrfachergänzung)). On the other hand, the systematic nature of the typology helps to discover new types—for example *Closed Sequence Selected Extension* items (Reihenerweiterungswahl). A concrete implementation of this type in the form of so-called *sentence completion tests* is described by Mahlow and Hess (2004). These tests consist of a piece of information the candidate has to extend repeatedly by choosing suitable extensions from a list of alternatives in order to create a coherent piece of new information. Sentence completion tests are also an example of an item type that clearly benefits from an electronic realization.

3. USING THE TYPOLOGY IN E-ASSESSMENT

As Conole and Warburton (2005, p. 23) note, "CAA systems vary widely in the number of question types supported." In fact, most systems offer a random assortment of item types, not based on any systematic criteria but rather influenced by marketing considerations, ease of implementation, and user requests. This forces test authors to adapt their testing to the item types provided by the system, which may not necessarily be the most adequate.

Furthermore, item types used in e-assessment or CAA scenarios are often referred to using conventional, but effectively arbitrary names. For example, Conole and Warburton (2005, p. 19) list "multiple choice, multiple response, hotspot, matching, ranking, drag and drop, multiple steps and open ended." These names are based on completely different criteria: Some types are named for the task candidates have to perform (matching, ranking), some are named for implementation details (drag and drop, hotspot), some are named for aspects of the process (multiple steps), and others are named for yet other characteristics.

The haphazard naming reflects the lack of systematic categorization in the field of e-assessment. However, it is not by accident that nomenclature and taxonomy play such an important role in biology or chemistry, for example: A common systematic categorization and naming system enables not only effective communication, but also reasoning about the system and its constituents. The lack of a systematic categorization of item types in e-assessment hinders reasoning about item types and discourages their methodical exploration and evaluation. All of the item types mentioned above can be categorized according to Rütter's typology.

A systematic typology of test items could also improve implementation of e-assessment systems: From the point of view of computer science, an implementation should abstract from the pedagogical *differences* between item types and find *commonalities* (in appearance or functionality), since the goal must be to support as many item types as possible with a small, generalized implementation. Specific item types should thus be derived from general types by parameterizing, in order to avoid the duplication of code and functionality; this helps to reduce the potential for implementation errors. As a proof of concept, we have successfully created a framework for the implementation of test items, based on Rütter's typology, which allows creating a large number of the item types of the Closed and Semi-Open classes (Fenske, 2007).

Interoperability and standardized exchange formats are required to make the transfer of electronic items and tests between different e-assessment and CAA systems possible and to enable the reuse of items in *item banks*. However, the only public specification in this area, the IMS Question and Test Interoperability (QTI) format (IMS GLC), is effectively unusable for this purpose due to serious technical and conceptual problems (Piotrowski, 2010). One source of problems is again that QTI is not based on a systematic account of item types, but rather on an ad-hoc collection of item types taken from pre-existing implementations. This failure to abstract from appearances and minor differences between item types makes the QTI specification extremely complex. A usable interchange standard for test items would therefore also have to build on a systematic typology such as Rütter's.

4. CONCLUSIONS

We have outlined Rütter's typology of item types as an example of a systematic approach to item design. Based on this typology, we have briefly discussed the suitability of item types for e-assessment. We have found a significant number of item types that are viable alternatives to the well-known multiple-choice question. Thus, Test authors and instructors should not restrict themselves to multiple-choice questions, as there are no pedagogical, test-theoretical, or technical reasons for doing so, but they should also consider other item types. A systematic typology of item types is an essential tool for selecting the right item type to fulfill the pedagogical and technical requirements at hand.

We have also discussed how a systematic typology of item types can serve as a theoretical foundation for the implementation of e-assessment and CAA systems, and for the design of interchange formats for items, offering significant advantages over the haphazard choice of item types common in current e-assessment and CAA systems and the IMS QTI specification.

In our view, a systematic typology of test items is a crucial factor for the further development of e-assessment if it is to be more than just electronic multiple-choice tests.

REFERENCES

Bennett, R. E., et al., 1990. *Toward a Framework for Constructed-Response Items*. Tech. rep., Educational Testing Service, Princeton, NJ, USA.

Collins, J., 2006. Education Techniques for Lifelong Learning: Writing Multiple-Choice Questions for Continuing Medical Education Activities and Self-Assessment Modules. *In Radiographics*, vol. 26, no. 2, pp. 543–551.

Conole, G. and Warburton, B., 2005. A Review of Computer-Assisted Assessment. *In ALT-J: Research in Learning Technology*, vol. 13, no. 1, pp. 17–31.

Fenske, W., 2007. *Formen der elektronischen Testaufgabe*. Diplomarbeit, Fakultät für Informatik, Otto-vonGuericke-Universität, Magdeburg.

Haladyna, T. M., 2004. *Developing and Validating Multiple-Choice Test Items*. Routledge, Milton Park, UK, 3rd edn.

IMS GLC, 2005. *IMS Question & Test Interoperability Specification*. IMS Global Learning Consortium. URL http://imsglobal.org/question/. Version 2.0 Final Specification.

Lienert, G. A. and Raatz, U., 1998. *Testaufbau und Testanalyse*. Beltz, Weinheim, 6th edn.

Mahlow, C. and Hess, M., 2004. Sentence Completion Tests for Training and Assessment in a Computational Linguistics Curriculum. *In COLING-2004 workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning*. pp. 61–70.

Morrison, G. R., et al., 2004. *Designing effective instruction*, chap. 11. MacMillan, New York, NY, 4th edn.

Osterlind, S. J., 1998. *Constructing Test Items: Multiple-choice, Constructed-response, Performance, and Other Formats*. No. 47 in Evaluation in Education and Human Services. Kluwer, Boston, 2nd edn.

Piotrowski, M., 2010. QTI—A Failed E-Learning Standard? In Lazarinis, F., et al., editors, *Handbook of Research on E-Learning Standards and Interoperability: Frameworks and Issues*. IGI Global, Hershey, PA, USA. To appear.

Rüdel, C., et al., 2007. Risikomanagement für E-Assessment. *In* Merkt, M., et al., editors, *Studieren neu erfinden – Hochschule neu denken*. Waxmann Verlag, Münster.

Rütter, T., 1973. *Formen der Testaufgabe. Eine Einführung für didaktische Zwecke*. C. H. Beck, München.