Multiple choice tests in distance learning: Solving the problem of poorly evaluated question pools through nonmetric multidimensional scaling

Läge, D; Dobricki, M; Häne, M

Postprint available at:
http://www.zora.uzh.ch

Posted at the Zurich Open Repository and Archive, University of Zurich.
http://www.zora.uzh.ch

Originally published at:
Abstract

A principle tool for evaluating learner knowledge within the scope of distance learning is multiple choice questions. In view of the ever growing number of tests that are given to learners, evaluating the question pools in advance is becoming something of a practical problem, as the usual method of counting correct answers entails a great dependency on clarity and correctness of the questions and the response options.

The current article suggests a procedure for enabling knowledge, within distance learning, to be assessed using multiple choice questions, without the need to examine the correctness of individual responses: Using this procedure, the pairwise similarity of the response pattern between examinees is depicted in a low-dimensional space with the help of nonmetric multidimensional scaling. The quality of each candidate can then be discerned from the relations between the persons. An experimental evaluation using the external criterion of four groups with differing levels of expertise shows that with a well evaluated question pool, the procedure is as good as the traditional counting of correct responses, and in the case of a non-evaluated question pool, it is even markedly better. This solves a fundamental problem inherent in tests conducted through distance learning.

Introduction

Over the last few years, E-learning processes have become increasingly differentiated across the board, in terms of both technology and content. Indeed, nowadays, they cover practically the whole range of traditional core subjects of universities and higher education institutions. Within each of these core subjects, there are specific areas that lend themselves particularly well to being offered through electronic media, and those in which a direct contact of the learners with the trainers is central. However, thanks to high technological standards and the broad range of possibilities offered by educational media, an astonishing amount of facets within training and education are also possible without direct face-to-face contact with the trainers.

Next to this positive preliminary conclusion regarding the teaching opportunities, the evaluation of learning progress, which nearly always accompanies education and training, seems like the proverbial poor cousin. Indeed, the development of evaluation procedures remains largely in its infancy, particularly if one considers that in distance learning, the personal examination discussion is generally no longer an option.

The multiple choice procedure, which draws questions and response options from a pool, is certainly the most advanced of the automated tests. However, a critical issue regarding the multiple choice procedure – and the focus of the current article – is that the evaluation of individual knowledge might take place against the background of possibly ambiguous or even misleading question-response combinations.

In order to illustrate this problem, let us consider a scenario in which a test with multiple choice tasks might usually take place: The aim of a test is generally to assess a person’s knowledge regarding a particular teaching goal and also in comparison to the quality of knowledge of other persons (fellow students, experts etc.). Traditionally, to conduct a multiple choice test, an examiner thinks up a series of questions, determines one or several correct solutions to these questions, and additionally formulates a series of wrong answers which function as distractors. Whether or not the resulting product is helpful in terms of the test’s aim can be shown through a test run using a representative sample from the population to be examined: The degree of difficulty and the discriminatory power of the individual questions can be determined such that it is possible to differentiate within the population. And whether this differentiation is accompanied by a qualitative scale of expertise can be ascertained by surveying specialists from the field under examination: If the specialists deem the same response options to be correct as the examiners, then the test can be conducted as a valid one.
While the aforementioned process might be acknowledged as the norm, it does not necessarily describe the real-life situation, as the time resources required to construct this kind of validity test are simply not available for every multiple choice examination. In practice, the usual compromise is to forgo presenting the questions to specialist colleagues in advance, and to only conduct an item analysis on the basis of the already implemented test (if necessary, poorly posed questions can then be eliminated in the evaluation). Mostly, due to time constraints, the subsequent item analysis is also dispensed with and the fate of the examinees is therefore left to the intuitive quality of the questions provided. And considering the fact that the amount of tests is increasing due to the Bologna process, in reality this situation is hardly likely to improve.

In light of this situation, it is clear that the traditional counting of correct responses cannot be seen as a safe way of evaluating knowledge against a backdrop of poorly evaluated test questions. Consequently, in the following, an experiment is described that demonstrates the effect of poorly evaluated multiple choice questions and, most importantly, suggests a procedure for enabling a reasonable evaluation of the knowledge of individual test candidates in spite of poor-quality questions. This procedure relies not only on there being differences in persons with and without knowledge in terms of the number of correct answers they give, but also presupposes that the response patterns of the knowledgeable persons are systematically more similar to one another than the patterns of those without knowledge (in other words: If knowledgeable people make mistakes, then these will frequently be “smart” mistakes that emerge based on specialist knowledge.)

For this purpose, the test data of all participants are compared not only with the correct solution, but above all pairwise among each other. The method best suited to evaluating the resulting matrix of proximity values between all pairs of persons is nonmetric multidimensional scaling [1]. Through the ordinal consideration of the proximity matrix, this stepwise approximation algorithm leads to a (generally Euclidean) map, which contains all scaled objects as points. The calculated configuration converts the similarity relations of the matrix into distances such that large similarities correspond to a small distance and small similarities are represented by a large distance. In robust NMDS, a low dimensionality ensures that dominant covariances are brought to bear in the map and any randomness in the proximity values is suppressed.

The aim is to show that such a procedure, which is based solely on the variances between the test candidates, is able to assess the examinees' expertise just as well as counting the correct answers. Which response option was rated as correct by the examiner and which constitutes a distractor can therefore be completely disregarded.

If the number of correct responses in a test is compared with the process of depicting covariances in an NMDS map, then an external criterion is required. For this purpose, we suggest a priori groups for which the general degree of expertise in a subject area is known. These groups should systematically differ from one another in terms of their knowledge on the group mean value level, but do not necessarily have to be completely disparate.

**Methods**

The experiment was conducted using 30 multiple choice questions from the area of psychopathology, each of which had four response options. A differing number of these options could be correct for each question. Fifteen of the questions originated from a pool of questions which, including all response options, had been evaluated and deemed as good in terms of factual correctness and unambiguousness by at least two experts (textbook authors in this area) in addition to the examiner’s assessment. The other 15 questions were constructed by a person experienced in drawing up actual examinations in the field under study, but were not additionally evaluated by peers. In the following, we will refer to “evaluated questions” and “non-evaluated questions”.

The following four expertise groups were defined as an independent external criterion for the experiment: Experts (university lecturers and researchers in the area of psychopathology), advanced students (advanced students in the subject of psychopathology), novices (psychology students in their first year of study) and laypersons (persons who did not study either psychology or psychopathology and did not work in any profession related to psychopathology).

All participants took part in the experiment individually and answered the questions via Surveymonke on the Internet. A total of 101 persons took part in the experiment, and were divided into the following four groups: 12 experts, 44 advanced students, 33 novices and 12 laypersons. The experts ticked an average of 24 wrong
answers (standard deviation 7.3), the advanced group 29 wrong answers (SD 5.6), the novices 38 (SD 5.5) and the laypersons 42 (SD 7.3). The mean performance of the groups therefore corresponds to the expected hierarchy of expertise. The differences between all groups were found to be significant in a t-test (alpha < 0.05), albeit overlapping. The effect sizes were as follows: experts/advanced students d=2.54 (huge effect), advanced students/novices 2.14 (huge effect) and novices/laypersons 0.63 (small effect).

Thus, the benchmark was set: If the four groups are reconstructed solely based on the rank order of the number of errors, then for the 15 evaluated questions, the number of errors correctly assigns 59 of the 101 persons. For the 15 non-evaluated questions, the number of correctly assigned persons decreases substantially to 45 (table 1).

Parallel to this evaluation based on the number of errors, the concordances were calculated between each pair of participants in terms of the number of identical responses ticked by both, independently of which questions these responses were assigned to and regardless of whether they were correct. The concordance value can therefore lie between 120 (identical response behaviour) and 0 (answered differently for all of the 120 alternatives). The matrix of all pairwise concordances was subject to a two-dimensional NMDS using the robust variant of the RobuScal algorithm [2]. In addition, as the 102nd point, another perfect solution with 120 correctly ticked responses was included in the NMDS (figure 1). This NMDS was also calculated separately for the pool of evaluated questions and the pool of non-evaluated questions (no diagrams; results reported numerically).

Table 1  number of correctly assigned subjects by following the two evaluation approaches

<table>
<thead>
<tr>
<th></th>
<th>by rank order of errors</th>
<th>by NMDS map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluated questions</td>
<td>59 subjects</td>
<td>56 subjects</td>
</tr>
<tr>
<td>Non-evaluated questions</td>
<td>45 subjects</td>
<td>54 subjects</td>
</tr>
</tbody>
</table>

Results

Figure 1 shows the two-dimensional NMDS map for the total pool of 30 questions. It is indicated to which group participants belong (type of symbol) as well as their performance (number of incorrectly placed ticks). At a glance, it is possible to see the high association between the distance to the perfect solution and the number of individually made errors. At the same time, however, the differing quality inherent in the wrong answers is apparent, insofar as the four expertise groups are focused mainly in different areas of the map: From the experts, only one single person can be found below the perfect solution, while almost all of the others lie above it. In this area, there are also several advanced students, but not a single novice or layperson. Most of the advanced students can be found just on the left below the perfect solution, where they are mixed with several novices. The group of novices is the most broadly scattered across the field, with the exception of in the area of the experts (indeed, if a convex hull is placed around all of the experts, not a single novice or layperson is found). The left-hand area of the map is reserved for the laypersons, with the exception of two novices (of poor performance), who also lie in this area of the space.

In terms of the visual description of the map, it can be seen from the distribution that there is a series of advanced students who converge extremely highly with the experts in terms of their response patterns and therefore also lie in that area. However, the numbers of errors made by these persons does not differ significantly from those of the advanced students who lie in the “typical” area for their group, namely on the left below the perfect solution. Nevertheless, the concordance in the errors makes a part of the advanced students more similar to the experts than the other part. Thus, there is potential here for evaluating expertise. Even without knowing of the coordinates of the perfect solution and of the numbers of errors made, the sole distribution of points in the NMDS space indicates would adequately performed and who failed.

Let us look now at the differences found in the assessment of evaluated and non-evaluated questions (these NMDS maps are not shown due to lack of space): Here, the measure for estimation lies solely in the distance of each point to the coordinates of the perfect solution. If a rank order is formed from these distances and the participants are assigned in terms of sectors to the four a priori expertise groups (just like the procedure performed with the numbers of errors above), then 56 hits emerge for the pool of 15 evaluated questions (cf.
Table 1). This value is therefore slightly lower than it would be for the group assignment based on the number of errors. However, the effect of the procedure is demonstrated in the 15 non-evaluated questions: Here, with 54 hits, the measure of distance is practically just as good and thus does not substantially decrease like it did for the assignment based on number of errors.

![Figure 1: NMDS map of all 101 subjects (similarity counted over the total pool of 30 questions), figures indicating the total number of errors made by the respective subject]

**Discussion**

With the formation of four expertise groups, an external criterion was created that enables the correctness of the responses and the concordance in the response behaviour to be tested. As the performance of these four groups overlaps, a perfect reconstruction is in any case not possible. However, both procedures – concerning both the evaluated and the non-evaluated questions – are in an area that lies clearly above the random hit rate of 33 (calculated on the basis of the given unequal sizes of the four groups).

In terms of the evaluated question pool, no substantial difference is apparent in the performance in terms of group assignment between the number of errors (59 hits) and the distance from the perfect solution (56 hits). Therefore, it can be stated that the measurement of distance measures the expertise for these types of question just as well as the deviation from a sample solution that has been deemed as correct by peer expertise.

However, if the sample solution becomes uncertain with regard to correctness and above all unambiguousness in determining the response options (no peer review of the response alternatives), then the deviation from a sample
solution deemed as correct loses more than half of its predictive value (from 59 to 45 hits for an expected value of 33). The assignment to the degree of expertise based on the distances in the NMDS map, however, only decreases to 54 and therefore proves to be clearly superior.

As described in the introduction, the suggested procedure makes use here of the concordances between individual participants: Persons with knowledge are indeed evidently able to exclude a series of responses as definitively meaningless, while persons without knowledge are unable to do so or at least do so to a lesser extent. Persons with knowledge tend to make the same mistakes as one another, above all in the non-evaluated question pool, in which unfortunate formulations can potentially give rise to certain response options being wrongly provoked as “correct” even though they are seen as “false” by the examiner (and vice versa). In the low-dimensional NMDS map, this manifests itself not only in a relative clustering of the degree of expertise (concordances of persons among one another), but also in the fact that at the same time, the perfect solution comes close to the cluster with the highest expertise (concordance of the perfect solution with the individual persons). As the novices and laypersons, whose subject-based knowledge is poor, show little concordance among one another, they are correspondingly scattered so broadly that a more fan-shaped structure emerges and they do not form any real cluster.

So far, this study has not systematically evaluated the diagnostic potential of belonging to a particular cluster: A more differentiated external criterion than the general expertise level would be necessary in order to show that, for example, those advanced students who lie within the convex hull of the experts can be categorised as better than the advanced students who lie outside of the expert region even if they show the same number of correct responses. If one observes how well two areas on the Y-axis of the NMDS solution are separated here for the advanced students, such an interpretation looks sound: In the top half, the persons with “smarter” errors can be found, and in the bottom half are those with “beginner-type” errors. However, this presupposes that the errors of the experts really are the “smarter” errors. Although this is plausible, it cannot be proven using the current data.

In order to gain a clear reference point for the evaluation, the perfect (in the examiner’s opinion) solution was taken into account as an additional point in the NMDS map. Interestingly, however, this model also works without using the criterion of correct/false at all. Under the assumption that knowledge is convergent, this type of fan-shaped structure is always found (in the extreme also a radix structure, and in that case spreading into a fan shape in all directions). Whenever knowledge is convergent, those persons with knowledge converge more strongly among one another than those without knowledge. This is sufficient in order to define the area in the NMDS map of good expertise. In the extreme case, this would mean that it would only be necessary to assign half-way plausible statements to the test questions and the better participants would themselves define which response options are consistent with the experts and which are not. Admittedly, this extreme perspective goes far beyond the concern of the research presented here. At this point, it should merely be demonstrated that even in the case of increasing pressures from tests, it is possible to maintain a high quality of assessment in spite of poorly evaluated test questions if one takes into account the different variances in response behaviour between persons with different degrees of expertise. In terms of the use of multiple choice questions in E-learning curricula and for tests conducted through electronic means, this comes as a great relief, as it is not always possible to evaluate all questions in a normatively perfect manner in advance.

References