EXAM QUESTION EVALUATION WITH ITEM RESPONSE THEORY

Evert Jan Bakker and Elly J. H. Korendijk
Chair of Applied Statistics, Biometris
Wageningen University, Wageningen, the Netherlands
evert-jan.bakker@wur.nl

We describe our first results with the analysis of the answers to 21 exams of the same Statistics course using Item Response Theory. We used a model with one-parameter (difficulty) and a model with two parameters (difficulty and discriminatory power). The second one appeared to be significantly better in 20 out of 21 exams. Using the outcomes of the analysis with the two-parameter model, we discuss some of the elements that can be analyzed with this model: the discriminatory power of individual questions, the information value of an exam, and we draw some conclusions about the type of exam questions used in this course of which the students' correct-wrong answering pattern aroused suspicion about how good the question was.

INTRODUCTION

At Wageningen University Statistics is taught to students of nearly all programs of Plant Sciences, Animal Sciences, Environmental Sciences, Food and Nutrition and Social Sciences. The course Statistics 2 (S2) is the first course in which Inferential Statistics is taught. The course covers the following topics: the Normal distribution, the distribution of the mean, the z-test for one mean, the one-sample, paired and independent samples t-tests, confidence interval estimation and simple linear regression and correlation.

The continuous increase in student numbers of the last ten years has led the teachers group to opt for exams that consist of multiple choice questions only. In the calendar year 2013 some 1200, mostly first-year, students followed the course in one of five periods, and seven exams were held. The 25 (or 20) questions all have four possible answers, of which exactly one is correct, the others are wrong. Some of the questions come in groups, as they relate to the same situation and/or data set. The exam is given to students in two versions, A and B. The versions are different with respect to the order in which (groups of) questions are asked. For a particular question, the order of the four answers in the two versions may also be different. We combine the results of the two versions into one data set.

For an exam of 25 questions, the mark is determined as follows. The maximum score is 10, the minimum score is 1. From the assumption that random guessing would, on average, lead to 6.25 correct answers, this minimum score is given to all with 6 correct answers or less. The mark of others is calculated as: $1+9\cdot(NC-6)/(25-6))$, where NC is the Number of Correct answers. A pass is given when the mark is at least 5.5, which requires 16 correct answers or more. (In fact, the mark can be increased by a 0.5 bonus for computer practical work, but we will ignore this here.)

Now one could wonder, how good this procedure is: how well the abilities (basis for the student marks) are estimated, how much each of the questions contributes, ... Stimulated by a request from within our university about 'how we (should) deal with the correct-guess probability' we decided to use Item Response Theory (IRT) as a framework to analyze (some of) the issues involved in judging exams, exam questions and exam results. The initial choice for IRT was based on suggestions from some colleagues, and after going through Hambleton & Swaminathan (1985) and some initial experience, we decided to stick with it. No higher math involved, except that we heard that IRT is also used by the group that makes the annual grade-eight tests, the results of which receive considerable weight in deciding to which kind of secondary education Dutch eighth graders will go. We used the program R to carry out the analyses.

ITEM RESPONSE THEORY

Applied to our exam situation: in an IRT model for dichotomous data (correct/wrong answer per question per student), it is assumed that each student has a latent one-dimensional quality, termed *ability*. The probability that a randomly drawn student from a group with the same ability, correctly answers a question (generally, an *item*) is a function of this ability. The simplest model that we used is the Rasch model, also called *the one-parameter logistic model*. See Figure 1

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2014 ISI/IASE]

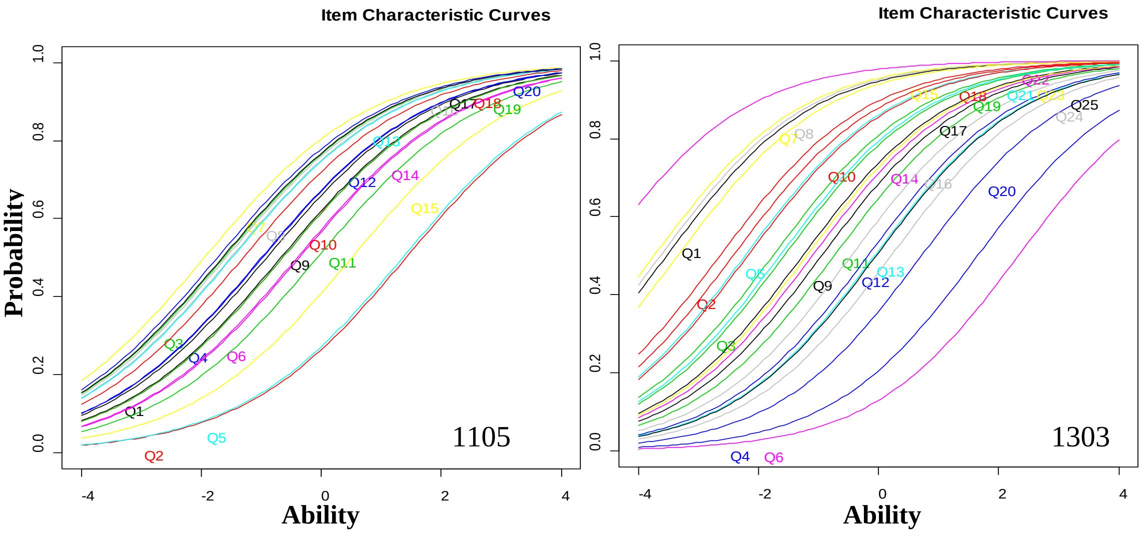


Figure 1 For two exams (coded 1105 and 1303): estimated relationships between correct-answer probability and (z-scores for) ability for each of 25 questions, using the Rasch model.

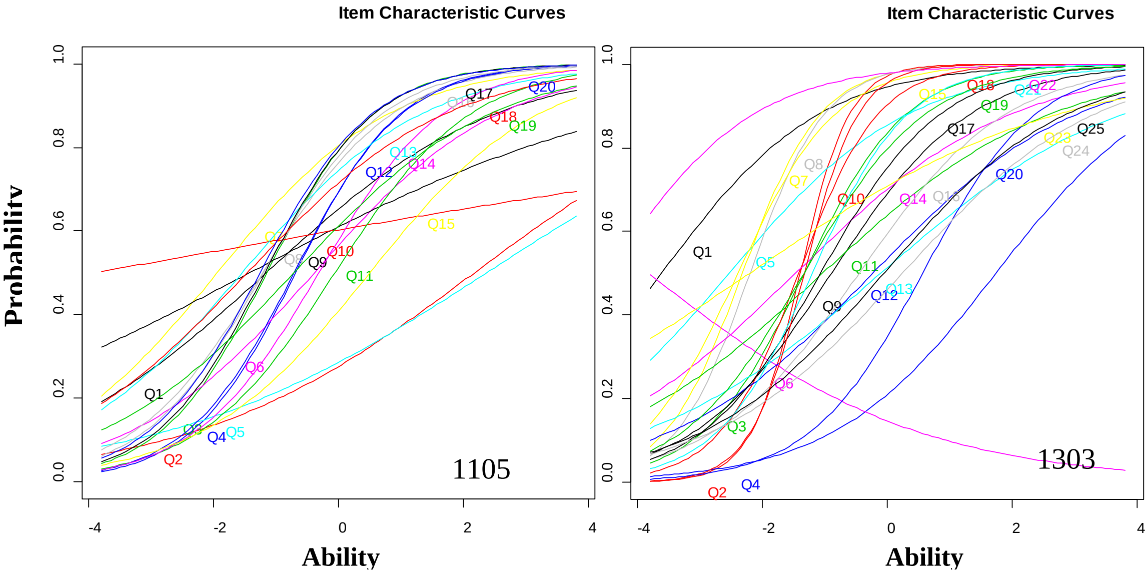


Figure 2 For two exams (1105 and 1303): estimated relationships between correct-answer probability and (z-scores for) ability for each of 25 questions, using the two-parameter logistic model.

for an example of the estimated relationships (called item characteristic curves) between the correct-answer probability and ability, as measured on a z-scale. The one parameter for each question is regarded as its (relative) difficulty. It is equal to the ability for which the correct-answer probability is 0.5. (Note that there is a problem of estimability: a more difficult exam could show the same results as a student group with lower average ability.) The Rasch model forces the lines to run parallel. Questions with curves on the right side of the pack are 'difficult' (relatively few students gave the correct answer), those with curves on the left side, are 'easy'. Estimation of a student's ability is based on the estimated graphs and the students' correct/wrong configuration. It appears that the number of correct answers is a sufficient statistic for the estimation of a student's (z-score for) ability. The two graphs depicted in Figure 1 are extreme in our set of 21 with respect to their width. The narrower band in the first graph suggests that the questions in the 1105 exam have a smaller difficulty range than the question in the 1303 exam.

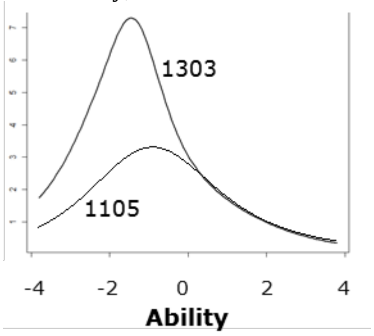
The *two parameter logistic model* allows the slopes of the S-curves in the logistic model to be different. These slopes have the interpretation of the discriminatory power of the questions, steeper curves having higher discriminatory power. The exam results as used for Figure 1 are also used to produce Figure 2. In the first graph (1105), three or four curves stand out, in the sense that they are the least steep, and have the lowest levels when ability is high (3-4). In the second graph one curve runs very low, and is sloped downwardly, indicating that low-scoring students were more likely to give the correct answer than high-scoring students. This may indicate that something is wrong with the question: the situation or the question is not stated clearly, no answer is really correct, some answers are very similar, ...With the two-parameter model, the ability z-score is estimated by a weighted number of correct answers, the weights being the discrimination coefficients. The remark about the width of the band covered by the curves made for the Rasch model, also applies here.

RESULTS

In our R-program we tested the fit of the two models. It appeared that the two-parameter model fits significantly better in 20 of the 21 exams. We will therefore discuss our findings for the results from the analysis with the two-parameter model.

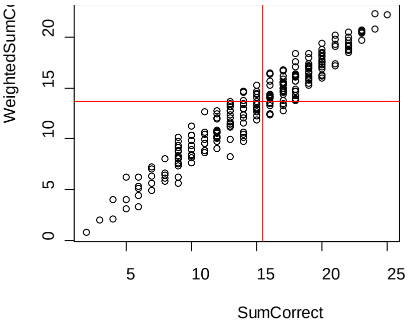
Merging the results for the A and B version per exam, has the advantage of a larger data set, which improves the precision of the parameter estimates. The item characteristic curves (as in Figure 2) for the two versions were usually very similar. This was especially true if the numbers of students in the A and B groups were large, say, both above 120. Having less than 100 students per group, often led to A and B showing different questions being conspicuous, i.e. curves for which the slope was small or even negative.

Figure 3
Information value as a function of ability, for two exams



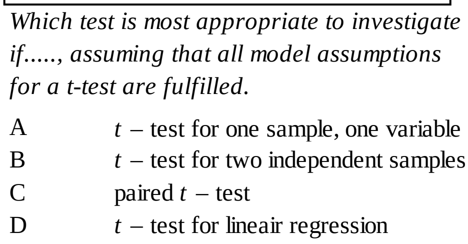
The suggestion from the difference in band width of the curves between the two graphs in Figure 2, is that the1303 exam has a wider range of abilities for which it has discriminatory power between ability levels. This idea is confirmed in Figure 3, where we display the so-called *test information value* curves for the two exams. For above-average abilities (ability z-value > 0) the information values are similar, but for below-average abilities, the 1303 exam is a much better tool to judge the ability of students than the 1105 exam.

Figure 4
Weighted vs. Unweighted Number of Correct answers: an example.



Given the fact that the two-parameter model generally gives a better fit than the Rasch model, it seems to make sense that the exam grading should be based not on the number of correct answers *per se*, but on this number weighted by the discrimination parameters. Figure 4 shows a scatterplot with the weighted number of correct answers vs the non-weighted number of correct answers, for an exam with 239 students. The vertical line gives the norm for a 'pass'. The horizontal lines is the norm for a 'pass' based on the weighted number of correct answers, if it is assumed that the same number of students would pass. From (a larger version of) the graph, it can be seen, that choosing the second criterion would result in a 'pass' for nine students, that did not pass in reality, and nine students who did pass, would not have passed if the weighted criterion had been used. In general, over all exams that we considered, this would be about 4%.

Figure 5
Question with answers, often resulting in an a-typical curve.


Which test is most appropriate to investigate if....., assuming that all model assumptions for a t-test are fulfilled.

A t – test for one sample, one variable
B t – test for two independent samples
C paired t – test
D t – test for lineair regression

We set out to analyze the 21 exams expecting it to lead to general conclusions about the type of questions that show 'bad behavior', i.e. rather flat or even downwardly sloped curves. We made question categories and investigated if certain categories showed a-typical behavior (flat or downward sloped curves). Categories were made by topic, and by question formulation. In the latter case, there was one type of question, given in Figure 5 that in 9 of 18 cases resulted in flat curves. It is quite an essential question, in the sense that a student in his or her later work should not make a mistake here, because the analysis would then be wrong, and, possibly, the conclusion as well. We think that this result is indicative of the tendency in our teaching to focus on what the students have to do in a given situation. At the time of the teaching, it is often obvious which type of situation is encountered (one sample, two samples, paired observations), as these situations are discussed one by one. But in the end, the student has to decide not so much how to carry out the t-test (the computer will do the calculations), but rather which analysis to choose. So what does this result tell us? It does not indicate that the question is wrong, but that we may need to pay more attention to the problem of "when to choose which test", which, even to good students is apparently not easy. Secondly, we think that this type of question may appeal to a different type of ability than most of the questions.

CONCLUSION

This paper gives the results of an analysis that is a beginning of what may grow out to be a long-term study on the use of Item Response Theory to analyze the difficulty and discriminatory power of exam questions, and on how to use the results to improve the quality of exams. In this paper we have only scratched the surface of the possibilities of this approach. We are also aware that there is a large body of literature on the subject that we have not (yet) read.

In our presentation we will have some more room to show results, and discuss the merits of IRT and the benefits it has had for Statistics education at Wageningen University.

DISCUSSION

Several implicit assumptions are made when using an IRT model. A thorough discussion of the assumptions is given in Hambleton and Swaminathan (1985). It is important to be aware of them when analysis results are used for making inferences about an exam or exam questions. The next two paragraphs illustrate this point.

In the two-parameter model, a flat or even downward sloped line seemingly indicates a 'bad' question. However, it is also possible that correctly answering the question requires an ability that does not correlate well with the ability to which most of the questions appeal. To give the question a small or negative weight in the determination of the ability, and hence in determining the mark, would be a mistake.

The two models used in the analysis are associated with different ways of placing the students in order of ability. This begs the question which of the two criteria should be preferred: the weighted or the unweighted number of correct answers. The statistical evidence points at the weighted score, as the two-parameter model gives a better fit. That this is not the last word about the matter, already follows from the fact that sometimes the estimated weight may be negative. Another warning comes from the conclusion in the previous paragraph: not every flat curve (which corresponds to a low discrimination parameter, and hence a low weight) indicates that the question is a poor one. Only the latter case would warrant a low weight in the student score.

The 'guess-correct' probability can be included as the third parameter (per question) in the model. Its estimate is seen in an item characteristic curve as the y-value at ability = - ∝, which then is no longer zero, as in the one – and two-parameter models. We tried this model for a few exams, but it did not lead to a significantly better fit.

REFERENCES

Hambleton, R., & Swaminathan, H. (1985). *Item Response Theory*. Boston/Dordrecht/Lancaster: Kluwer-Nijhoff Publishing.