A dialogue about MCQs, reliability, and item response modelling

Daniel B. Wright & Elin M. Skagerberg,
University of Sussex

Multiple choice questions (MCQs) are becoming more common in UK psychology departments and the need to assess their reliability is apparent. Having examined the reliability of MCQs in our department we faced many questions from colleagues about why we were examining reliability, what it was that we were doing, and what should be reported when examining the reliability of MCQ exams. This paper addresses the most frequently asked questions.

I N ORDER TO cover the breadth of the psychology curriculum, many departments have included multiple choice questions (MCQs) as part of their assessment. These have the further advantage that they can be marked more rapidly than essay questions. In our department MCQs were recently introduced as one component for several of our second year courses. Fifteen MCQs were included with some essay questions for most exams. The courses in this year provide the main material that is required for the national recognition of our psychology degrees by the British Psychological Society. Therefore it is important that all areas of the curriculum are assessed, and this is more easily done with MCQs than by more traditional methods. We were asked to examine the *reliability* of the MCQ portions of these exams.

Reliability has several meanings both within methodological discourse and in general. The purist definition is how related the scores from a *test* would be if it were given in exactly the same conditions to the same people. As this cannot happen methodologists have created different ways to estimate reliability. The most popular of these are based on how associated the individual items are with each other, or the internal consistency of the test. The reliability is a function of the number of items and some average internal consistency.

The purpose of this paper is to describe how the reliability can be assessed and how this information can be given to the people who construct exams. By sharing our experiences we hope to help others improve the quality of their MCQ exams. This paper can also serve as a primer for item response modelling for students wishing to learn some of the basics of the procedure.

This paper is based on the types of questions that we received both from staff and students (and also some questions that we expect people had but did not ask). This paper is a dialogue between a hypothetical colleague and ourselves. Our colleague's queries are in *italics*. They are about why we should being interested in reliability, how the statistics are conducted and reported, and how these results should affect how exam questions are constructed. The questions are divided into three sections. The first set of questions deal with why we should be interested in reliability and what we mean by reliability. The next section covers details of item response modelling. The final section is about how we recommend providing this information to the course convenors. The questions are a mixture of statistical questions, like 'what is reliability?' and pedagogical questions, like 'why should we measure reliability?'. In fact, many of the questions require using some statistical procedures to help address pedagogical questions.

Conceptual Questions about Reliability

What is reliability?

The word 'reliability' is used in many different situations. A simple mathematical definition is that it measures how correlated scores would be on a test if the exact same test was given to the same people in exactly the same situation. This is not possible because, among other things, people would learn from taking the test once. Therefore methodologists have developed several ways to estimate reliability. Some of these require multiple testing sessions (so-called test re-test reliability), but the most popular involve seeing how well the scores on items within a test relate to each other. A conceptually simple approach is to say how well a score based on the odd numbered items correlates with a score based on the even numbered items. This is called split-test reliability. This is based both on number of questions and how related all the questions are. Cronbach's alpha is a kind of average of all possible split-test reliabilities.

Why should we measure Reliability?

Formal assessment can serve several different purposes. This includes the estimation of students' achievement for the particular course that is being assessed. The percentage of items correctly answered on the exam provides one possible measure of this achievement and often this measurement is used. Is it important to examine whether this measurement is reliable? Phrased another way, with respect to MCQs, would we expect similar questions to produce similar measurements and therefore similar estimates of achievement?

Everyone we have spoken with agreed that being able to measure achievement reliably, when described in this way, was important and to be able to show that an exam was reliable or not was also important. Discovering that an exam has poor reliability could be a catalyst for improving the exam in subsequent years. From a pragmatic perspective, having some measure of reliability is useful for exam boards (and for assessment of the exam process) and is now often being requested by external examiners.

I expect some of my students to do well on one part of the exam and poorly on another part, but I expect other students to have the opposite pattern. How would this affect reliability?

Some students do better on certain parts of an exam compared with other parts. Often this is can be accounted for by chance fluctuations, but some students do specialise. This will lower the estimated reliability because most measures of reliability assume that all the items are measuring a single latent variable, like knowledge of social psychology. This is called the unidimensionality assumption. If this is not a valid assumption then alternative methods are necessary. Many tests have blocks of questions about certain topics. These are sometimes described at testlets and there are appropriate statistical procedures designed to analyse these.

What should be done if the statistics show that there are two or more distinct 'things' that are being measured?

If we were constructing an attitude scale to measure a single attitude, and we found out that the scale actually measured two distinct attitudes, our published account would presumably describe this and each respondent would have two attitude scores. Using this (sound psychometric) logic we should report two marks for students' achievement on an exam if it turned out that two distinct types of achievement were necessary to account for the data.

But there is only one column in the Excel spreadsheet for the marks' array!

The requirements of most exam boards do mean that you have to come up with a single score. However, it is often possible to provide more detailed feedback to the students and often conveners will be interested in the different types of achievement. Further, the procedures described here help educationalists examine this assumption empirically, rather than just tacitly assuming it which is what people have traditionally done.

Daniel B. Wright & Elin M. Skagerberg

What happens if I have little variability among my students on their level of achievement?

This can happen for two reasons. The first is that the students are all approximately the same as each other. Because there is little variability in achievement the reliability statistics will not be very useful.[1] However, with respect to undergraduate psychology exams these situations are rare because there is usually great variability among psychology students. The second reason is more common, and that is where the exam is poor at differentiating the vast majority of students. This can be done on purpose, particularly with threshold exams, where the exam is used to assess whether students have achieved some bare minimum threshold. In these cases the exam's purpose is not to differentiate among most of the students, so the standard methods for checking reliability are not appropriate. In other cases the exam writer may be at fault. For example, an exam composed of some very easy questions and others so difficult that most answers are guesses will not differentiate students.

Should we also be checking the reliability of the essay questions and laboratory reports, and should we measure reliability across courses since at the end of a degree we assign students to a single classification?

Yes. But with a caveat. We should not strive for every correlation to be extremely large. The variable that we are trying to estimate, academic achievement, is complex. To capture the complexities it is necessary to measure many different things. It is possible to achieve high reliability, but of a narrow aspect of what you are really trying to measure.

How to see if MCQs are reliable

I was taught Cronbach's alpha. Is this all that people use?

While Cronbach's alpha was a significant advance in its time, further advancements have been made. There is a huge field of edu-cational measurement. Many postgraduate programmes in the USA, Canada, the Netherlands, etc. offer psychometric courses and there are journals that specialise on the topic (for example, *Applied Measurement in Education, Educational and Psychological Measurement,* and *Journal of Educational Measurement*). Cronbach's alpha is a greatly misunderstood and mis-applied statistic (Thompson, 2003b). Using alpha as the sole criterion for the quality of a test is problematic. For example, simply repeating questions will generally increase alpha. The two main advances since alpha are generalizability theory (Cronbach *et al.*, 1972; see Crocker & Algina, 1986; Thompson, 2003a; for introductions) and item response modelling (see Embretson & Reise, 2000; Henard, 2000). Item response modelling (IRM, sometimes denoted IRT for item response theory and sometimes called LTM for latent trait modelling, for reasons explained below) is more often used in education and is better suited for our purposes. Embretson and Reise (2000, chapter 2) describe some of the differences between using alpha and IRM, and Crocker and Algina (1986, particularly chapter 15) compare generalisability theory with IRM.

All this said, Cronbach's alpha is still reported more often than any other measure (Hogan, Benjamin & Brezinski, 2003), is understood by many, and has some value and therefore is often still worth reporting. Below we suggest reporting this and suggest an additional, easily calculated, measure which decreases as the number of questions increases, so is affected by the number of questions in the opposite direction to alpha.

Conceptually, and without equations, what is IRM?

It is probably easiest to understand IRM conceptually with reference to factor analysis. The two procedures are very similar (Bartholomew *et al.*, 2002). The difference is that in the typical exploratory factor analysis

---

[1] This is one of the reasons why it is inappropriate to say that a test has certain reliability. The reliability of a test is dependent on the sample.

the researcher has recorded several supposedly interval level variables, but in a typical exam the manifest variables are binary, the student either gets the question right or wrong. Thus, diagrams like Figure 1 depict the IRM situation in the same way as a researcher would conceptualise exploratory factor analysis for measuring an attitude using a questionnaire. Responses for each individual item are based on the latent variable achievement and the item specific error term. The latent variable is often called a latent trait and therefore the procedure is often called *latent trait modelling*.

IRM provides a useful graph for analysing the individual questions. These graphs, called item characteristic curves (ICCs), plot the probability of correctly answering the different question as a function of the latent variable achievement. There are different types of IRMs. One of the ways in which they differ is how complex the curves are in the ICCs. The simplest models, sometimes called Rasch models, assume items only vary in how easy they are to answer (Figure 2a). More complex models allow the items to vary in

how well they discriminate according to the achievement latent variable.[2] The steeper the curve the better the item is at discriminating among people. Where the curve is steep is where the item is best able to discriminate. Thus, in Figure 2b Q1 is a poor question because students who have low achievement (the scale for the achievement variable is usually shown in units of standard deviations) have a similar probability for correctly answering the question as those with high achievement. Q2 and Q3 both have high discriminability being able to differentiate very poor students from above average students. Q4 is has less discriminability but does differentiate among better students.

Sometimes it is helpful to allow a 'guessing' parameter. This is the probability of correctly answering the item if you have no achievement. Figure 2c shows this. For Q1 and Q2, even those with very low achievement have about a 50 per cent chance of getting the question correct. The difference between these questions is that the high achievement students do not perform much better than this for Q1 but they are very
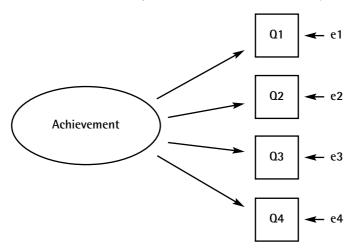

Figure 1: A conceptual item response model (IRM) in which one latent trait (achievement) accounts for the shared variation among four questions. Each question's variation is based on the latent trait plus a unique error term.

---

[2] The word 'discriminate' has several meanings in English. Here, 'discriminate' means how well the item is at differentiating people who have high latent variable values for the attribute being measured and those who have low values.
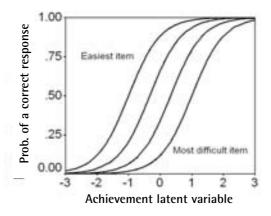

Figure 2a: Item characteristic curves (ICCs) for the one parameter or Rasch model which allows items to vary only by how easy they are to answer.
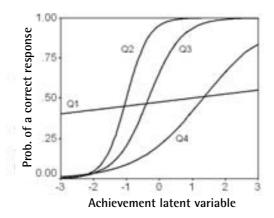

Figure 2b: The two parameter model which allows items to vary by ease and discriminabilty.
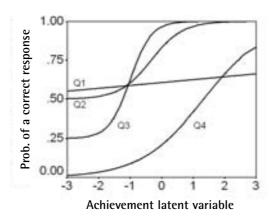

Figure 2c: The three parameter model which also allows different guessing parameters.

likely to correctly answer Q2. Q3 has a 25 per cent chance of being correctly answered by people with no achievement. People with low achievement are very unlikely to answer Q4 correctly. This is unusual for MCQs, but can occur for short answer questions.

Isn't 'guessing' just 25 per cent if there are four options?

Guessing completely at random would be 1 over the number of options. However, in many MCQs there are responses which are clearly wrong. Consider the following question for potential drivers (from http://www.dsa.gov.uk/mockpaper/ down-loaded 26-5-05):

You are going through a congested tunnel and have to stop. What should you do?
a. *Pull up very close to the vehicle in front to save space.*
b. *Ignore any message signs as they are never up to date.*
c. *Keep a safe distance from the vehicle in front.*
d. *Make a U-turn and find another route.*

Somebody with no driving knowledge (but with a little common sense, which is not what these tests are supposed to measure) should be able to eliminate one or two of the alternatives thereby raising the 'guessing' percentage above 25 per cent. 'Guessing' estimates are usually higher than the amount predicted by guessing at random. It is important to consider other ways of estimating how people with no knowledge about a course should perform because these estimates should be important for determining how to map students' responses onto grade classifications. In the next section we address how the guessing factor can be used to adjust marks.

The ICCs in Figure 2 all go up smoothly? What happens if an item does not have a smooth increasing function?

These curves, as with all models in science, are simplifications. If it were possible to find the exact relationship between the probability of correctly answering an item and the

achievement then it would look more complex. However, if the item is a good question the probability should increase as achievement goes up. However, bad questions exist where this is not the case. Consider the following (not from any Sussex exam):

What is the appropriate statistical test to run in SPSS if you are interested in testing the null hypothesis that there is no association between gender and performance on a psychometric test? Assume the test produces normally distributed interval data.
a. *correlation.*
b. *t-test.*
c. *no preference.*

This is an example where the relationship between giving the correct response and knowledge is non-monotonic (i.e. it goes up and down). Someone with no knowledge could have 'no preference' because they do not know what 'correlation' and 't-test' are. A student with a little knowledge might look for words like 'association' in the question and think 'correlation'. A little more knowledge might guide the student to those 'which test' diagrams in the back of some statistics textbooks. This would lead them to 't-test'. Even more knowledge and the student would realise that 'correlation' and 't-test' are the same model and they might opt for the 'no preference' option. A further increment of knowledge might lead the person to choose 'correlation' because it is arguably a more useful measure of the effect size. Finally, somebody might opt for one of the two tests because of the output that SPSS gives (and probably opt for 't-test' because its SPSS procedure provides more useful output for understanding the effect compared with the correlation output).

This is a bad question. The assumptions of the basic IRM limit the complexity of the ICCs. Several methods are necessary to weed out bad questions, preferably before they are included in an exam. While this is an extreme example, many 'trick' questions do exist, where the answers that seem right to people with some knowledge are not correct.

Daniel B. Wright & Elin M. Skagerberg

But some of these 'trick' questions allow the really good students to show their knowledge?
There is a difference between difficult and 'trick' questions. It is good to have some difficult questions. The problem with some 'trick' questions is that very poor students sometimes outperform mediocre students. It is worth trying to re-frame the questions so that you would expect them to monotonically increase with achievement. It is worth also realising that it is very difficult to measure all aspects of a student's achievement with MCQs.

Looking at the ICCs in Figure 2 helps me to think about what I am trying to measure and the role of individual questions. The ICCs suggest that if I want to be able to differentiate students at different levels of achievement I need to make sure that I have some questions that discriminate students with low levels of achievement, some questions that discriminate people with medium levels of achievement, and some questions that only the best students correctly answer.
Yes! This is true with any form of assessment where you are trying to estimate achievement across a broad spectrum. Examiners are often reluctant to include 'easy' questions and overestimate how much the weaker students know. Given that discriminating between students who just pass and those who fail is probably the most important border, it is important to have questions that accurately discriminate among students at this level of achievement. ICCs are a useful way to show where different items discriminate along the achievement latent variable.

What are the equations for Item Response Modelling?
If you have used logistic regression, then the equations look fairly similar to those that you use to plot the probability of a 'success' by a

continuous covariate. The main difference is that rather than the continuous covariate being an observed variable, it is the latent variable. The equation for ICCs of the one parameter model, where items are assumed to differ only in how easy they are to answer, is as follows. Let p_{ij} be the probability of the ith person correctly answering the jth question:

$$p_{ij} = \frac{\exp(b(X_i - a_j))}{1 + \exp(b(X_i - a_j))}$$

where $\exp(x)$ means e^x and a_j is a measure of the ease of item j.[3] X_i is the value on the achievement latent variable for the ith student. The parameter, b, is for discriminability and is assumed to the same for all items for this model. In Figure 2a the only differences between the lines are due to the different values of a_j (1, .33, -.33, -1 from left to right; b is 2 for all items). The smaller values of a_j are for the more difficult questions. With the two parameter model shown in Figure 2b, the b_j values, now including the subscript j for items, are allowed to be different for each item. The value of $b_1 = .1$ for Q1 means that the ICC is essentially flat and therefore unable to discriminate according to achievement. The other values are 3, 2, and 1 for Q2, Q3, and Q4, respectively.
The three parameter model is:

$$p_{ij} = c_j - (1-c_j)\frac{\exp(b(X_i - a_j))}{1 + \exp(b(X_i - a_j))}$$

The c_j parameters are for guessing. If someone has no ability (which would be $X_i = -\infty$), then this person would have a probability of c_j of answering the jth question correctly.[4] In Figure 2c the values are 0 for Q4, .25 for Q1 and Q3, and .5 for Q2. For Q1 it is necessary to imagine the X-axis extending to large negative numbers to imagine that this ICC has a lower asymptote at .25. For some exams it is worth assuming that all the items have the

---

[3] Sometimes the equations used are slightly different. In particular, often it is $(X_i + a_j)$ and the a_j are measures of difficulty.
[4] If $(1-c_j)$ is changed to $(d_j - c_j)$, this becomes the four parameter model. This allows there to be a ceiling probability for correctly answering the question (i.e. even the brightest student does not have 100 per cent probability of correctly answering the question). While this could be useful, it is seldom used in practice because of the increased complexity of the models.

same value for *c*. This can simplify the computation and is often necessary for a solution to be found.

Are solutions not always found?

For more complex IRMs, the estimation will often not converge to a solution. If different a, b, and c parameters are estimated for each item, if there are 100 items, then this is 300 parameters. This can create problems. When researchers evaluate these models they should often focus on a small subset of items. This is also true for factor analysis and structural equation modelling. Sometimes cowboy analysts will use many manifest and latent variables exploring complex models without due consideration of the statistical problems inherent in these approaches. We are reminded of the approach described in the preface to one of the most insightful books about latent variables:

> *The reader already familiar with factor analysis may be surprised that our emphasis, in theory and examples, is on models with only one or two latent variables. On the other hand, we pay more attention to questions of sampling variability and goodness of fit than is usual. This shift of emphasis is deliberate because we wish to stay within the bounds of what is statistically defensible. (Bartholomew & Knott, 1999, p. xi).*

So we should concentrate on simpler models, but before you mentioned that the unidimensionality assumption is often suspect. Does allowing extra dimensions (or latent variables) make the model more complex?

It does. This is why it is important to be careful in factor analysis or structural equation modelling if you have more than one or two latent variables. However, in IRM the tradition is just to have one. It is worth examining whether more variables appear to be needed. The equation for two latent variables does not appear much more complex. Depending on the text the equation is often written in different ways, but to make it most similar to equations presented earlier we can write it for the two latent trait model as:

$$p_{ij} = \frac{\exp(b1_j X1_i + b2_j X2_i + a_j)}{1 + \exp(b1_j X1_i + b2_j X2_i + a_j)}$$

This can be expanded to more latent variables (Xk_is), but with each additional Xk_i the complexity of the model increases greatly (here $k = 1, 2$). The a_j here are different from before. We use the multiple trait model simply as a way to assess the validity of the single trait models although a more detailed exploration of an exam could focus on these different latent variables.

What happens if we think the latent variable is not continuous, but classifies people into a small number of groups?

This is an area of enquiry that is likely to become more important throughout the next decade (Wright, 2006). Recall the phrase 'latent trait models', which is what many statisticians use to refer to IRM. There are also 'latent class models' that estimate the probabilities for people belonging to different latent groups. This type of analysis is very popular in sociology and is becoming more popular in psychology under the banner of taxometric analyses (*taxon* meaning a group). Waller and Meehl (1998) provide a detailed introduction.

While latent class models are used in many areas, the applicability of these models for MCQs is uncertain. This is because when examiners hypothesize different groups these groups usually can be ordered along an achievement dimension. This makes it difficult statistically to differentiate between the model which assumes a continuous latent variable and one which assumes several latent groups that are ordered along some dimension.

Are there other techniques that I could also use?

IRM can either be thought of as an extension from exploratory factor analysis, but where the observed variables are binary, or as an extension from logistic regression, but where the covariate is a latent variable. Thus, similar techniques to these can also be used. Principal component analysis (PCA), which is sometimes thought of as a descriptive alternative to factor analysis can be used. Binary

data can be used with PCA. Binary data meet all the measurement level requirements of interval data and as a descriptive technique PCA works with binary data.

Other techniques which might appear promising, like signal detection theory and multilevel logistic regression, could be useful for some purposes, but are not designed to allow people to focus on particular items. Depending on the purpose of the analyses, different techniques could be used. Because a large number of statistical techniques under the general umbrella of IRM have been created for analysing exams, it is usually best to look at these before trying to reinvent the wheel.

What statistical packages do you recommend?
The most popular package among academic psychologists is SPSS. Unfortunately, as of version 13.0, SPSS does not have a procedure for IRM. It can do some useful analyses for checking the reliability of exams, but not IRM. This may be one reason why IRM is not as common as it should be. The popular statistics package SYSTAT does offer IRM facilities and there is a procedure (called LTM for latent trait modeling) that can be used in both S-Plus and R. As R is free this is a cost effective option. The LTM procedure is similar to GENLAT, which is also free and available at http://multilevel.ioe.ac.uk/team/aimdss.html. The above packages are all fairly limited in what they output. There are lots of specialist packages available, some of these are reviewed in Embretson and Reise (2000, Chapter 13). Some of these are free; some are not.

We felt the most complete suite of software was the IRT suite which is available at http://www.ssicentral.com/product.htm#la3. This suite contains four packages, each of which costs $250. However, they allow people to rent the software which is a very cost effective way of seeing whether the software meets your needs. We strongly recommend spending $40 on the manual (du Toit, 2003) if you are planning on using this software. Besides describing all 4 packages in detail

and with examples, it provides an in-depth introduction to the area.

What to report
Our role has been to create reports for course convenors. In this section we describe how we approach this.

Are these fancy statistics the only things worth reporting?
No. Fancy statistics should only be reported as a last resort. Often it is much better to communicate aspects of the data using simpler procedures (Wright, 2003; Wright & Williams, 2003). Often the simplest analyses are appropriate for the questions people have.

What is the first thing that you report?
Begin with univariate statistics for both items and students. Rank the items by percentage of accurate responses and also look at the errant responses. In describing the percentages it is important to highlight any items that were so difficult that they appeared near chance guessing levels. By looking at errant responses it is possible to identify any 'trick' questions. For the reports we have conducted there have been a few question where more people gave a particular errant response than gave the correct response. This shows that however the course has been taught it has led more students providing a particular wrong answer than the correct one.

It is also useful to provide more qualitative analyses of the questions. Ideally this should be done prior to having students take the exam, but some questions with complex sentence structures, double negatives, 'all the above'/'none of the above' responses, etc. are likely to make it into some exams. With respect to 'all the above'/'none of the above' responses, educationalists usually are against this type of question (see for example, Haladyna, Downing & Rodriguez, 2002). The view is that these tend to lower reliability, but the opinion is mixed. Dochy and colleagues (2001) examine this type of question. They found that people who are guessing tend to opt for this option.

Question number	% correct
10	18
12	23
8	3 3
5	4 1
9	4 5
14	45
1	5 5
3	5 6
2	5 8
7	6 2
6	6 8
11	69
13	70
15	71
4	7 7
total	53

```
Freq # correct
 6  0 ******      6 no shows
 2  1 **
 4  2 ****
 7  3 *******
 8  4 ********
14  5 *************
30  6 *****************************
39  7 **************************************
38  8 *************************************
30  9 ****************************
42 10 *******************************************
31 11 *****************************
10 12 **********
 3 13 ***
 2 14 **
 0 15
```

Figure 3: Univariate statistics for items (percentage correct) and for people (number of correct responses).

The univariate statistic for students can be printed in tables or histograms, depending on the numbers of questions and people. There are lots of different types of histograms. Figure 3 shows a table with the univariate statistics for items and people. The number of correct responses for each item is listed in a table and the students' number correct is shown in a histogram.[5]

From this histogram most people answered correctly between 40 per cent (six correct answers) and 67 per cent (10 correct answers) of the items. The histogram shows six 'no shows'; it will depend on departmental policy how these should be treated. If the percentage accurate is to be used to estimate achievement, it is important to decide if any adjustment is necessary to map these values onto the 0 to 100 per cent scale that is used to classify students. Most departments have verbal descriptions of the different bandwidths corresponding to the grade classifications.

How can this adjustment be done?
There are several ways to do this and several considerations when deciding which way to use. The fail boundary is arguably the most important. Suppose this is set on the 0 to 100 per cent scale at 40 per cent. One way to conceptualise this is to assume that somebody at the pass/fail boundary should be able to answer 40 per cent correctly. If this were the case then the raw observed percentages could be used.

However, an examiner might want to say that to pass a student should know 40 per cent of the items. Because students could guess correctly some reduction for guessing is needed. Because of the need for transparency to students, the level for 'guessing' should be calculated beforehand, rather than based on the three parameter IRM. It is often assumed that students either know the answer or do not, and if they do not know the answer they guess (sometimes avoiding clearly incorrect alternatives). This is sometimes called the all-or-none or threshold model from signal detection theory. If K is the actual knowledge, G is the percentage of time that a guess should be accurate, and O is the observed percentage, then O is predicted by:

$$O = K + (1-K)G$$

---

[5] In this section we report data from several different exams and do not name the exams.

Solving for K produces:

$$K = \frac{O-G}{1-G}$$

If there are four alternatives for each item, then if somebody does not know the answer and guesses completely at random they have a 25 per cent chance of correctly answering the question. If G = .25, this means a person needs O = .55, or 55 per cent correct, to have K = .40. If this is taken as the pass mark this will increase the number of fails and if K is used for the entire scale all the other grade classifications are also changed. With the observed percentages in Figure 3, there is a large group who just passed with the unadjusted scores (the 30 people with six correct), but once the adjusted is done they fail as do several other students. If using O = 40 per cent, only 35 fail (not counting 'no shows'), but if using K = 40 per cent then 142 fail. Thus, the choice of O versus K is important.

The G = .25 is probably too low because

there is often at least one response that is clearly incorrect. It is usually not as obvious as in the driving question given earlier, but often one incorrect answer stands out. As G increases the percentage that need to be correctly answered to have K = .40 also increases. If True/False questions are used then G is at least .50 and students would need to get at least 70 per cent correct to pass if K is used and 40 per cent is the fail boundary. G can be calculated in other ways, for example, having people who do not take the course answer the questions. However, these alternatives require more work on the part of the examiners.

Given the difficulties setting G, and that it looks bad for departments to fail a lot of students, would a good solution be just to use the observed percentage? That is a solution, but a bad solution. It ignores the problem that students are not correctly answering enough questions. A

Table 1: The correlations (lower triangle), odds ratios (upper triangle), and percentage correctly answered (the diagonal) for five questions. 95% Confidence intervals are in parentheses below the estimate.

	Q1	Q2	Q3	Q4	Q5
Q1	69%	1.71	1.35	1.79	2.21
	(63%, 74%)	(0.88, 3.33)	(0.77, 2.37)	(1.05, 3.06)	(1.26, 3.86)
Q2	.10	23%	0.94	1.09	2.88
	(-.02, .22)	(18%, 28%)	(0.51, 1.75)	(0.62, 1.93)	(1.34, 6.19)
Q3	.07	-.01	70%	1.13	1.46
	(-.06, .19)	(-.13, .11)	(64%, 75%)	(0.66, 1.92)	(0.83, 2.58)
Q4	.13	.02	.03	45%	1.74
	(.01, .25)	(-.10, .14)	(-.09, .15)	(39%, 51%)	(1.01, 3.01)
Q5	.17	.17	.08	.12	71%
	(.05, .29)	(.05, .29)	(-.04, .20)	(.00, .24)	(65%, 76%)

Note: The 95% confidence intervals for odds ratios were calculated in SPSS. Those for percentages and correlations were calculated using http://glass.ed.asu.edu/stats/analysis/. All sample sizes were n = 266.

better solution is to increase the number of easy questions. Marks should be adjusted, but it is important first to make sure that the exam does assess people across the spectrum of achievement.

Which bivariate statistics should be reported?
Bivariate statistics can be reported for people and items. Reporting statistics comparing pairs of people would tell you if two people were responding nearly identically, which along with information like seat locations could be strong evidence of collusion (McManus, Lissauer & Williams, 2005).

Bivariate analyses of items can be particularly informative. An association matrix can be calculated. Most statisticians prefer the odds ratio (or the log of it) as a measure of association rather than the correlation, but most psychologists are more comfortable with correlations. Reporting both, with their confidence intervals, in a square matrix with one measure on the upper triangle and one on the lower triangle, is an option. The odds ratio and correlation can produce different results, particularly when the proportion of right or wrong answers is very high (Goodman, 1991). Univariate information, like the proportion correct, can be reported on the diagonal. Table 1 is an example of this.

What do you report for multivariate analysis?
Because of its popularity it is worth reporting Cronbach's alpha (which for binary data is sometimes called KR-20, which is the name of a formula used to calculate it). Following what is often done in research methods courses for attitude scale construction and measurement, we report alpha, and which items if removed increase alpha. In most of the exams that we have looked at there were items like this, in some exams several items. The recommendation in attitude measurement is usually that these items are removed. This is difficult on exams because a student who got a psychometrically poor question right would be penalized and could complain.

Alternative statistics can also be reported. One is the proportion of variance accounted for by the first principal component. This tends to decrease as the number of questions increases, so moves in the opposite direction of alpha. Other alternatives include the median correlation and the median odds ratio. Their means can also be used, though you might wish to use Fisher's z and the natural log transforms if calculating their means. For the data in Table 1, Cronbach's alpha is .32, the first principal component accounts for 28 per cent of the total variation, the median correlation is .09, and the median odds ratio is 1.59.

What do you report from the unidimensional IRM?
Rather than reporting many statistics, we focus on the ICC graphs. This allows the convenor to see which items are easy and which are difficult, and which items discriminate well along the achievement dimension. Often the models will not converge, particularly with the more complex models when items are not associated with the others are included. Therefore, this needs to be explained to the convenor. If, for example, the three parameter model cannot be estimated, it is worth fixing the guessing parameter, c_j, so that it is the same for all items. If this still does not converge then the ICCs for the two parameter model should be reported. With all ICCs, it is worth stressing that these are idealisations.

We used the package BILOG-MG. It produces ICCs and many other useful graphs for our purposes. However, it is often necessary to create new graphs from the output that will best suit the needs of your readers. In Figure 4 the numeric output was used to create a single graph with all the ICCs using the scatterplot procedure in SPSS (the scatterplot procedure in SPSS is used to graph functions). Graphing the ICCs allows them to be compared more easily. In this case it shows that two questions (Q11 and Q12) have flatter slopes than the others and therefore discriminate less well along the latent dimension.
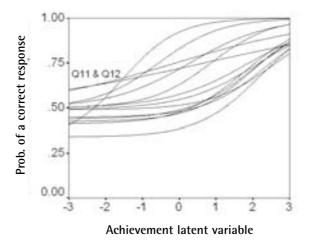
Figure 4: Multiple ICCs plotted separately for the different items as produced by BILOG-MG (4a) and plotted together using the SPSS scatterplot option (4b).

There is only a single dimension, achievement, plotted in Figure 4. How do you see if this is appropriate? To test this we used the program TESTFAC (the freeware GENLAT can also be used). TESTFAC provides several possible measures, including the proportion of variance accounted for. Because this is a measure with which people are familiar we used it to produce scree plots. The cumulative percentages of variation accounted for (what is in a scree plot; shown here for up to nine traits) are: 14, 21, 28, 37, 44, 53, 62, 72, and 82. This produces a fairly linear scree plot, suggesting that there is little underlying structure to the data. The Cronbach's alpha for these data was .51 and the first PCA accounted for only 14 per cent of the variation (and the subsequent percentages show the same pattern as the scree for TESTFAC). The fact that the PCA solution will produce similar findings means that this procedure can be used to check for the dimensionality assumption if you do not want to spend the money on TESTFAC (which costs additional money to BILOG-MG) or do not want to download GENLAT.

What else can item response modeling be used for? This paper is just a brief introduction to the most basic IRM procedures. There are many other things that IRM can do. For example, you can include:

1. Allow partially correct answers;
2. Track items over years so that cohorts can be compared;
3. Allow students to answer different questions, which is useful with computerised adaptive testing;
4. Compare groups of people;
5. Treat missing values, 'don't know' response, and errors differently, and more.

What were the main findings from your analyses? Showing people the ICCs made them aware of two important aspects of their exams. First, items were identified where the probability of answering them correctly was not related to answering other questions correctly. The ICCs for these items were flat. From a psychometric perspective these items should be removed, although this might be difficult as it would penalized those students who answered them correctly and who may complain. They should be changed for subsequent exams. Second, most of the ICCs showed that many of items discriminated best at the high end of the achievement latent variable, but few at the low end. This means that there were not many easy ques-

tions. We encouraged convenors to add some simpler questions to their subsequent exams in order to differentiate those students near the pass/fail boundary. It is common for lecturers to overestimate the proportion of students who will answer an item correctly.

Address for orrespondence
Daniel B. Wright, Psychology Department, University of Sussex, Brighton, BN1 9QH.
E-mail: danw@sussex.ac.uk

Note

References

Bartholomew, D.J. & Knott, M. (1999). *Latent Variable Models and Factor Analysis (Kendall's Library of Statistics 7)*. London: Arnold.

Bartholomew, D.J., Steele, F., Moustaki, I. & Galbraith, J. (2002). *The analysis and interpretation of multivariate data for social scientists*. London: Chapman & Hall/CRC.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Raharatnum, N. (1972). *The dependability of behavioural measures: Theory of generalizability for scores and profiles*. New York: John Wiley.

Dochy, F., Moerkerke, G., De Corte, E. & Segers, M. (2001). The assessment of quantitative problem-solving skills with 'none of the above' items (NOTA items). *European Journal of Psychology of Education, 16*, 163–177.

du Toit, M. (Ed.) (2003). *IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact*. Lincolnwood, IL: Scientific Software International. Inc.

Embretson, S.E. & Reise, S.P. (2000). *Item response theory for psychologists*. Mahmah, NJ: Lawrence Erlbaum Associates.

Goodman, L.A. (1991). Measures, models, and graphical displays in the analysis of cross-classified data. *Journal of the American Statistical Association, 86*, 1085–1111.

Haladyna, T.M., Downing, S.M. & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309–334.

Henard, D.H. (2000). Item response theory. In L.G. Grimm & P.R. Yarnold (Eds.) *Reading and Understanding More Multivariate Statistics* (pp. 67–97). Washington, DC: American Psychological Association.

Hogan, T.P., Benjamin, A. & Brezinski, K.L. (2003).

Reliability methods: A note on the frequency of use of various types. In B. Thompson (Ed.) *Score reliability: Contemporary thinking on reliability issues* (pp.59–68). Thousand Oaks, CA.: Sage Publications.

McManus, I.C., Lissauer, T. & Williams, S.E. (2005). Detecting cheating in written medical examinations by statistical analysis of similarity of answers: Pilot study. *British Medical Journal, 330*, 1064–1066.

Thompson, B. (2003a). A brief introduction to generalizability theory. In B. Thompson (Ed.) *Score reliability: Contemporary thinking on reliability issues* (pp.43–58). Thousand Oaks, CA.: Sage Publications.

Thompson, B. (2003b). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.) *Score reliability: Contemporary thinking on reliability issues* (pp.3–23). Thousand Oaks, CA.: Sage Publications.

Waller, N.G. & Meehl, P.E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua*. Thousand Oaks, CA.: Sage Publications.

Wright, D.B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology, 73*, 123–136.

Wright, D.B. (2006). The art of statistics: A survey of modern statistics. In P.A. Alexander & P.H. Winne (Eds.) *Handbook of educational psychology* (2nd edn), pp.879–901. Mahwah, NJ: Lawrence Erlbaum.

Wright, D.B. & Williams, S. (2003). Producing bad results sections. *The Psychologist, 16*, 644–648.