COMMENTARY ON STATISTICAL ISSUES ARISING FROM CHAPTERS

Harvey Goldstein

Several contributors have emphasised that statistical techniques can at best provide evidence to assist comparability judgements, but cannot themselves define comparability in the absence of experienced interpretation. To quote Murphy 'Statistical methods cannot resolve the most fundamental dilemmas… The best that statistical treatments of this type can do is to indicate possible interpretations'. Even so, there remain considerable problems with the statistical methods currently in use. In addition to specific comments on individual chapters in this note I want to draw from the different chapters to highlight the generic issues and suggest possible directions for future exploration.

A key issue addressed by several authors, especially Murphy, is that of 'interactions', for example in the case of a reference test, different 'slopes' for different sub-groups such as males and females. In one sense identification of these (and they do exist in at least some cases) implies that comparability cannot be achieved, yet on the other hand it is possible to argue that while there is no single common adjustment, differential adjustments, for example separately for males and females, are possible. The problem is that this will raise equity issues. To avoid this, what is effectively done in practice is to ignore any such differences and fit a single adjustment model. This has the effect of averaging over all the actual differences in the data, which may be justifiable but needs to be explicitly stated. If we do model differential relationships then this can lead to improved understandings about how candidates respond, and the exploration of such models should be encouraged.

A similar issue arises when basic item response models (IRMs) such as Rasch are used, as advocated by Bramley and Coe. A problem with simplistic models of this kind is that they ignore not only group interactions but, in the case of Rasch, also differential discriminations. Thus, for example, in equation (17) in Bramley's chapter, if the discriminations denoted by α_{AB} are not equal then the assumed properties of the Rasch model do not hold and interpretations based on the model are not justified. Thus, for example, the probability of one object exceeding another will depend on which two are being compared, and so one cannot claim that the method is 'sample free'. Moreover, the fact that there is 'consensus' on the use of the model is not the same as the existence of good evidence for its use. A further problem is that conformity to the model often results in 'discrepant' data being discarded – thus helping to confirm the model. Of course, if the data are manipulated so that judges and objects that do not conform to the assumptions of a simple model are removed, then Bramley is likely to be correct. That, however, is bad science; it prioritises making the data conform to a particular statistical model rather than trying to understand how the data are actually structured and fitting models that attempt to describe such structures. If there are good, large-scale, studies that have explored this issue it is a pity that they are not discussed. If there are no such reliable studies then we need to be very cautious about interpreting results.

This point is relevant to the Bramley's discussion of 'misfit'. Essentially what happens with the Rasch models is that a model is fitted to the data and various diagnostics used to determine whether it is indeed a good description of the data structure. This approach is the one typically used in this kind of item response modelling work, but at best it will only indicate where a model may be inadequate – it does not itself constitute a proper exploration of alternative models, which is a crucial issue. To do this, as I have suggested, requires the setting up of a statistical model that explicitly incorporates parameters for the effects being sought, for example examiner interactions, and then making inferences based on their estimated values. One of the problems, of course, is that to properly explore models that allow different discriminations, interactions etc., requires very large samples that typically may not be available. Nevertheless, I would suggest that it is the responsibility of those who promote and use any particular technique to justify its assumptions and compare its performance against reasonable alternatives.

A further issue arises with item response model-fitting, namely that of multi-dimensionality. Coe discusses this in the context of large common factors implying a single underlying dimension. Without getting into a general debate about dimensionality, it does seem important that whenever a model is fitted, the existence of more than one dimension is studied, in particular after adjusting for social and other background factors. There are other concerns when using very simplistic models such as Rasch. Thus Bramley's equations (10)–(12) are valid only if one assumes no interactions, for example between judges and boards. It is possible to elaborate these models to study such possibilities directly, yet this seldom seems to be done, and it is worth pointing out that studies of item 'misfit' are technically very poor substitutes.

Those who would use the results of statistical models need to be careful. Certainly, as Schagen and Hutchison point out, all the techniques used, whether based upon item response models or more traditional regression procedures, should have a multilevel component, and software for fitting such models is available. In addition, such models should routinely incorporate interactions as well as adjustments for background factors in order properly to inform any subsequent adjustments. One of the uses of multilevel modelling not covered by Schagen and Hutchison is their extension to handle general item response models, of which the Rasch model is just the simplest version. These models are in fact just factor analysis models with a binary (correct/incorrect) response and can readily be extended to explore more than one factor and to handle multilevel structures where, for example, there can be factors (dimensions) at the level of the school as well as at the level of the pupil. A discussion of such models with educational examples is given by Goldstein *et al*. (forthcoming). It should also be noted that the most efficient modelling procedures for fitting IRMs use a random effects formulation that not only allows the generalisation to multilevel structures but also avoids the need to discard instances where all the responses are the same (see Bramley's chapter, section 3). This would in fact seem to be an obvious development from the current use of 'fixed effect' models.

There are several further directions that multilevel modelling could follow. An important one, mentioned by Schagen and Hutchison, is the use of random coefficient models, where the relationship between an examination mark or grade and another variable is allowed to vary across centres such as schools. As already mentioned in the case of interactions, the existence of such differential effects can impose important questions of interpretation. For example, we may not only find that the relationship with a common test varies from boys to girls, but also from school to school. We may also discover, say in the context of a common examinee analysis, that the relationship between a pair of marks varies from school to school; differential entry policies or other unobserved factors may be responsible. The existence of such variation will again raise difficult questions about interpretation and use, but it should not be ignored. More advanced versions of the basic multilevel model (see, for example, Goldstein, 2003) allow us to take account of data cross-classifications such as when pupils move from the Key Stage 4 period to a post-16 school or college, and we can apply multiple membership models such as when, during the course of a Key Stage period, pupils move across schools.

In short, multilevel models can and should introduce more complexity than is typically assumed in current procedures. If there is to be any really useful development of statistical methods for monitoring comparability in the near future, the adoption of these techniques is both desirable and necessary.

References

Goldstein, H. (2003). *Multilevel statistical models*, (3rd ed.). London: Arnold.

Goldstein, H., Bonnet, G., & Rocher, T. (forthcoming). A study of procedures for the analysis of PISA reading data. *Journal of Educational and Behavioral Statistics*.

RESPONSE TO COMMENTARY ON STATISTICAL ISSUES

Tom Bramley

In those of his comments which related to my chapter Harvey Goldstein focused on the perceived inadequacies of the Rasch model for analysing the kind of data collected in a paired comparison study. Within the chapter I showed the connection between the Rasch model and Thurstone's Case 5 specialisation, and between the discrimination parameter in a 2-parameter IRT model and Thurstone's 'discriminal dispersion'. I indicated one alternative possibility for analysis (logistic regression), and Goldstein has suggested some more (multi-dimensional and/or multilevel models).

However, I feel that to concentrate on improving the statistical models would be a misdirection of effort. The current constraints on cross-moderation studies in terms of available scripts, numbers of judges and time, mean that the data sets are not large enough to support more complex modelling, as Goldstein acknowledges. The real issue is to produce an outcome which is interpretable in terms of the performance standards (grade boundaries) in the different boards.

To do this I have argued that it is necessary to relate the latent trait implicit in the judges' ordering of scripts to the latent trait implicit in the raw mark totals created by applying the mark schemes. Understanding the meaning of these latent traits is therefore very important. This is what is so striking about Thurstone's work – his attempts to get to grips with the underlying philosophy of 'subjective measurement' and his concern to relate it to psychological processes. This concern is not always apparent in the writings of some of those who spurn 'simplistic' measurement models.

If researchers in this field are not to be haunted by Gene Glass's well-known quote about the language of performance standards (but which could be applied much more widely) – that it is '… pseudoquantification, a meaningless application of numbers to a question not prepared for quantitative analysis' (Glass, 1978) – then we need to take the issue of measurement seriously, as Thurstone and Rasch did. The Rasch model can be seen as specifying requirements which the data must meet to yield additive, linear measures. Assessing whether the data do indeed meet such requirements is not 'bad science', nor is trying to understand the causes of any specific, localised, misfit to such a model.

It is interesting to note that psychometricians generally have been accused of 'bad science' in their refusal to consider the hypothesis that the latent variables in their statistical models might not possess the structure necessary to support quantification (Michell, 2000). Michell would see this issue as logically prior to any fitting of a simple *or* complex statistical model. In my view we need to draw on, and integrate, developments in cognitive psychology, measurement theory and statistical modelling. Progress in the latter has resulted in a 'psychometric embarrassment of riches' (Borsboom, 2006) in the variety and complexity of models available, but this has yet to be well integrated with the first two. In the context of cross-moderation exercises, I therefore feel that our efforts are better devoted to understanding the psychological processes underlying the experts' judgements, discovering the features of candidates' performances which influence them, and knowing exactly what we mean when we say one script is 'better' than another.

References

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440.

Glass, G.V. (1978). Standards and criteria. *Journal of Educational Measurement*, *15*, 237–261.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory and Psychology*, *10*, 639–667.

RESPONSE TO COMMENTARY ON STATISTICAL ISSUES

Robert Coe

Harvey Goldstein claims that the use of the Rasch model is 'simplistic', fails to take account of a number of commonly found characteristics of the data such as multi-dimensionality, differential discrimination and different relationships for different sub-groups, and so constitutes 'bad science'. He suggests that a solution to all these problems is available in the form of multilevel modelling.

These criticisms confuse modelling, which, as Goldstein says, aims 'to determine whether [a model] is indeed a good description of the data structure', with measurement, which is the aim of the Rasch model. The requirements of the Rasch model are precisely those of adequate measurement; if data do not fit the model, one cannot simply change the model.

The issue of multi-dimensionality is particularly interesting, and Goldstein is right that it is 'important that whenever a model is fitted the existence of more than one dimension is studied', though a statistically purist view of uni-dimensionality could prevent any measurement ever being allowed. We must remember that any valid measure of anything that is worth measuring will always contain more than a single pure dimension. The crucial question is not whether we can find more than one dimension in a measure – we almost always can – but whether its interpretation in terms of a single construct is defensible and useful. Andrich (2006) refers to the 'fractal' nature of measurement, in which the same construct can be seen as both uni-dimensional and multi-dimensional, depending on how it is viewed. For example, it may be appropriate to view a set of GCSE examinations as measuring a common construct, 'general academic attainment', while simultaneously acknowledging that a particular subject, such as mathematics, measures something interpretable uniquely as attainment in 'mathematics'. This too may be subdivided into components, and for some purposes the construct 'mathematics' may be too broad. We might, for example, wish to talk about performance specifically in 'algebra', though for other purposes we might want to subdivide further and talk about understanding of 'simultaneous equations'. The fact that our original construct can be subdivided so many times does not necessarily make it inappropriate to view it as a single 'uni-dimensional' construct.

This apparent paradox seems counter-intuitive and this may account for some of the resistance to the use of statistical methods to establish comparability across different examination subjects. It seems obvious that examinations in different subjects measure quite different things, so the basis for comparison must be problematic (e.g.

Goldstein and Cresswell, 1996). In fact, in the application of the Rasch model referred to in my chapter (Coe, forthcoming), the latent trait measured by the 34 GCSE subjects in the analysis proved to be impressively uni-dimensional, with a person-reliability (internal consistency) estimate of 0.94 and 83% of the item variance explained by the latent trait in Principal Components Analysis, and the eigenvalue of the biggest residual contrast being just 1.9. This suggests that despite the perspectives of the examiners, teachers and candidates in GCSE examinations in different subjects that they are measuring very different things, the empirical data seem to tell a rather different story; examinations in many different subjects are remarkably consistently measuring the same thing.

References

Andrich, D. (2006, April). *On characterising the roughness of an educational measurement*. Paper presented at the 13th International Objective Measurement Workshop, Graduate School of Education, University of California, Berkeley.

Coe, R. (forthcoming). Comparability of GCSE examinations in different subjects: an application of the Rasch model. *Oxford Review of Education*, *34*.

Goldstein, H., & Cresswell, M.J. (1996). The comparability of different subjects in public examinations: a theoretical and practical critique. *Oxford Review of Education*, *22*, 435 441