Correspondence

Applying the partial credit method of Rasch analysis: language testing and accountability

Two questions occupy the thinking of many language testing researchers at this time: what is language testing research *for*?; and how can nonclassical measurement theory further the aims of language testing research? Bachman and Clark's proposal of a new 'framework of communicative language ability', to be investigated through a programme of collaborative research and instrument development, has set the stage for a new phase in language testing research:

> . . . language proficiency testing has reached an important watershed, at which recent statistical advances, increased attention to the development of detailed theoretical models of communicative language proficiency, the existence of a number of useful prototype instruments, and growing interest in proficiency-based language teaching and assessment on the part of both language teachers and researchers all combine to produce a very opportune moment for the field to make rapid, synergistic advances in both the theory and the practice of language proficiency assessment. (Bachman and Clark, 1987: 33)

An important element is missing from their exciting proposal, an element that less careful researchers than Bachman and Clark may as a consequence omit from their research, with damaging results. That element is the explicit accounting that we, as testing researchers, must make.

I want to remind us of the distinction between *test development*,

i.e., the design, construction, piloting and implementation of language tests which will be brought into operational use, and *testing research*. There are really two types of testing research, although they may be done by the same people and they may use the same investigative methods. The first type, which I will call *test validation*, has as its purpose the investigation of a test instrument which will be, or which is being, used for a nonresearch purpose. The investigation typically encompasses reliability, criterion-related validity and content validity, and recently has become more likely to include face validity and construct validity. Test validation is generally carried out through a range of psychometric procedures carefully matched to the validation aims, although recently there have been moves toward less formalist, more philosophical epistemologies, in particular *a priori* construct validation (Weir, 1986; Hamp-Lyons, 1986). The second type of testing research, which I will call *metatesting*, has as its purpose the investigation of how, why and when language is acquired or learned, not acquired or not learned, the ways and contexts in which, and the purposes for which, it is used and stored, and other such psycholinguistic questions. When this type of testing research investigates existing test instruments, the concerns are not with the test's adequacy for operational purposes, but stem from the need to ensure that appropriate tools for investigating these psycholinguistic questions are chosen. When research test instruments are designed and administered in this type of testing, again the concerns are not with how the instruments would perform operationally, but with what useable information for furthering the theoretical investigation can be gained.

Clearly there is, in most contexts, a continual interaction between the two purposes of testing research. As understanding of language acquisition issues grows through careful theoretical research involving hypothesis formation and testing, i.e., through metatesting, that understanding can be applied to the design of potentially operational test instruments which can be shown to be reliable and valid, i.e., to test validation. At this point the other needs of operational tests enter the balance. Equally, as language testing researchers are asked to engage in test validation, not only practical but also theoretical insights will be gained from data configurations. These insights will lead to the forming of generalizable hypotheses which can be tested through a careful programme of testing research of the second, theoretical type, that is, through metatesting.

The problem is the need to keep it clear in our minds which area we are dealing with at any particular time and report our work to others with the same clarity. A further concern has been that the emphasis of language testing research should be on theory-building, that is, on

research which will contribute to a growing understanding of the psycholinguistic issues and thus to a knowledge foundation from which test development for operational purposes, and the procedures for test validation, can increasingly draw. This, they would argue, and I would concur, is pre-eminently what language testing research is for. Psychometric procedures and other research tools take their value from the explanatory power they have in fulfilling the purposes of language testing research.

Inseparable from the above, however, is the fact that language testing (and here I include all kinds of language testing research) is a political act. Language testing researchers must always be aware of the potential consequences of what they do, whether their focus at the time is test development, test validation, or metatesting. What does this mean? As Stevenson pointed out some years ago, it often means that 'language testers are more concerned, or should be, with stopping bad testing than with developing new tests' (1981: 18). Bad tests are not only those which have low reliability, or which claim to measure one construct while in fact measuring another. Bad tests are also those which discriminate against certain classes of individual; bad tests are those which can be subverted to the purposes of a government which wishes to disenfranchise a subset of the population. Too often, bad tests turn out to be those which have detrimental washback onto the curriculum, which is painfully sensitive to changes in testing practices and very apt to interpret such changes as statements about values and philosophies. It will be clear from the foregoing that I take what Stevenson (1981: 17) refers to as an 'expanded view' of language testing and language testing research: in this view we cannot validate a test in isolation from its adminstration, scoring, score analysis and reporting, nor from the interpretation and use of the information from the test, nor from its short-term or long-term social, cultural, economic, educational/curricular effects. For me, all of this *is* the test.

In this view, every decision made by a language tester, at either the research or operational stages, cannot avoid being a political and an ethical decision. We may not close our eyes to the eventual outcomes of our testing research activities because we are engaged in theory-building, or metatesting; it is too difficult to be certain that, in the inevitable iterative process involved in the sharing of our ideas with our colleagues around the world, something will not be lost in the retelling and inappropriate conclusions drawn from our work.

Because of the views I have outlined above, when I read accounts of language testing research, I look at the care with which a study has been constructed and carried out, and the clarity with which it has been reported; but I also look at the impact it is likely to have in terms

of our expanding knowledge base; the directions it may suggest for future testing research; the implications it may suggest to the knowledgeable and (perhaps more importantly) the less knowledgeable about how language is acquired and learned and therefore how it might be taught; the questions it may suggest as being worthy of attention. Inevitably, because the real-world considerations of language testing research are ultimately derivable as curriculum concerns, I look at how the study can be expected to affect curriculum in either or both the pre-test context and the post-test context.

Two papers in a recent issue of *Language Testing* (Volume 4, No. 1) by Adams, Griffin and Martin (1987) and by Pollitt and Hutchinson (1987) have prompted these comments. Both make use of, and make certain claims for, the value of item response theory, specifically Rasch analysis using a partial credit method, and both have a clear impact on curriculum.

In their study, Pollitt and Hutchinson start from the classroom: and they stay connected there, stating that 'As a formative assessment package, TELS Profile is designed to give teachers and pupils as much information about developing language skills às possible' (p. 75). The purpose of their study is to 'develop a framework for the (formative) assessment of competence in writing' (p. 75), and the purpose of this paper is to use 'the Rasch item response theory . . . in its partial credit form to show how graded assessment that does take account of the specific features of a particular task (as in the TELS) may be used to assess competence in writing' (p. 73). In the terms I established earlier, their concern is with test validation: they are investigating a test instrument which may be used for a nonresearch purpose.

Pollitt and Hutchinson provide a table to summarize the item difficulties and they interpret the fit statistics. They provide us with complete data for person ability estimates. Figure 2 and the discussion of it on p. 85 are likely to be particularly helpful to readers unfamiliar with Rasch partial credit in allowing them to see exactly how items and persons interact, since it represents visually on one diagram the distribution of subject responses and the distribution of item step difficulties. The discussion of the three misfitting subjects on p. 82 allows us to see what it is that is causing the psychometric model to flag misfit, and to consider for ourselves the plausibility or otherwise of the actual response patterns shown by the misfitting subjects. Wright has pointed out that 'the most important information could be about the misfitting person, as a diagnostic profile of what's going wrong, in the person or in the test, even though the response is not a valid test response' (Seminar at the University of Chicago, March 1988).

Once the item/subject interaction has been made clear in Figure 2, we can go on to Figures 3a and 3b, where the items have been reorganized by difficulty (note that line 1 on p. 86 should read 'difficulty of the *last* step'): we can see in Figure 3a the visual representation of the statement in the text (p. 86) that the components of competence are almost completely separated in terms of the relative difficulty of achieving a full score, and (p. 88) that there is a strong tendency for the tasks to separate in terms of the relative likelihood of getting at least a score of 1. They are able to extract a good deal from these data which can help them understand and refine the various facets of their model of writing assessment: the notion of both general skills or components based *a priori* on a theoretical model of language competence; of specific writing skills each related to a certain task type and measured through a range of tasks; of a theoretically defensible scale length; and of the combination of difficulty and judgement strategies for scoring through careful specification of all facets of the measurement instrument. While they did not claim to set out to conduct metatesting research, but rather test validation research, the work Pollitt and Hutchinson have done can be generalized and used by others in designing new studies based on similar questions in other contexts. Equally, the account illustrates, without the need for assertion or argument, how the partial credit model can be applied and how results can be interpreted taking into account all the information made available by the psychometric procedure.

Pollitt and Hutchinson take an area which is both of educational significance and which is highly problematic in testing terms, and make only very limited claims as to their intentions. The outcomes of their study belie their modesty, for they move from test development and validation into theory building, providing results of significance beyond the bounds of their own study. The curriculum implications which can be drawn from the conclusions of their study (which space prevents me from elaborating) are many. This study provides us with a model of what accountability in language testing research should be. One can perceive that the less gifted schoolchildren they were concerned with, and their teachers, will reap benefit from the outcomes of their test development and validation work. The study has clear curricular implications, for example, in the USA, where many, including White (1985), have argued that multiple-choice tests discriminate strongly against certain groups while performance-based tests display less bias. Pollitt and Hutchinson have made a contribution not only in presenting a responsible new test instrument, but also in revealing some important issues for the evaluation of existing instruments for the assessment of writing.

In contrast, the study by Adams *et al.*, at least as far as it is reported here, leads me to serious questions about the educational consequences of their work. They begin by saying that 'A lack of sound testing procedures can . . . lead to problems in research design and ultimately to inappropriate theory development' (p. 10). They criticize studies which have used other, classical psychometric methods to explore the UCH/DCH question, because the choice of psychometric tool used predisposes either the unifactorial or multifactorial solution. Like Pollitt and Hutchinson, their focus is on the psychometric tools used in theory-building at least as much as on the theories themselves. They go on, in section IV, briefly to introduce the Rasch partial credit model and provide an overview of the rest of the paper, in which 'the partial credit model is used to illustrate how one type of "authentic language test" can be constructed and validated, and how confirmatory approaches to test development can be used in research settings' (p. 12). Their study 'sought to define . . . [a] . . . dimension as an example, without any claim to importance or to dominance among other possible dimensions' (p. 13). In seeking to define a dimension of language proficiency they are engaged in construct validation, that is, in metatesting, in theory-building about the nature and acquisition of language. Adams *et al.*, posited the existence of a dimension of language proficiency which they call 'grammatical competence'. This 'dimension' was selected on *a priori* grounds, i.e., classroom observations, interviews with teachers, and a survey of the literature indicated that it was a 'general development area of general concern' (p 13). However, because 'The differences observed . . . meant that language acquisition or developmental models based in achievement of course-specific objectives would not be appropriate for large-scale testing', they then decided to construct a 'generalized proficiency measure'. As far as we can tell the classroom was then left behind as the researchers moved, through a process they do not share with us, to 'a test of spoken language focusing on the structural elements' (p. 13).

At this stage the original construct is also left behind, of course, and we have instead a hypothesized construct which is being validated, one which is not grounded in pedagogy. But I'll come back to that. I have two serious concerns about this paper: the first relates to the application of the partial credit model. It the authors' main purpose is to show how the partial credit Rasch measurement model is a better measurement model for investigating the hoary issue of whether language competence is unitary or divisible, they should rehearse not only the failings of factor analysis, but also the virtues and limitations of the partial credit model, especially the question of the unidimensionality assumption of the partial credit model, the

conditions under which that assumption can be said to be violated, and the significance of this for the psycholinguistic questions they are investigating. They do not do this. In particular they need to note that the model is very robust to violations of unidimensionality (Henning *et al.*, 1985), which is a psychometric property independent of any concept of 'dimensions' of language proficiency, which are psycholinguistic properties or concepts. Having selected the 'dimension' of grammatical competence in section V, they go on in section VI to characterize the Rasch partial credit model, not the 'Independent Grammatical Competence' (my naming) model, as one might have expected. We are faced, then, with a situation where we do not know anything about the model which the researchers are validating: we cannot properly examine their assertions without the test objectives and scoring criteria, which have not been appended to the paper. Item 1.2 appears to be a test of the lexicon, and I wonder how it found a place in a test of grammatical competence. Looking at item 1.3, I wonder whether it tests either verb mastery specifically or grammatical competence more generally: it may simply be a test of the ability to apply provided rules, since it is a transformation table with two completed examples. The ICC also suggests it is a dichotomous item rather than a scalar one. The ICC for item 4.4 shows excellent scale separability, but the item has negative misfit. According to Ben Wright (personal communication) this suggests the presence of another, positively correlated, variable. Reading Adams *et al.*'s description of the scoring criteria for the item (p. 18) makes me wonder if the other variable might not be 'fluency' or 'communicative competence' or some such less prescribed variable than that measured by the rest of the test. Item 4.4 is more like I would expect a test of grammatical competence in an authentic, interview context to be, and the researchers' comments lead me to assume that no other items were like it, from which I conclude that the other items were all more or less discrete point, like 1.3 and 1.2.

Adams *et al.* do not provide us with the person fit statistics, which we need in order to judge their claim that their test is unidimensional. Their assumption that if the data fit the psychometric model they *de facto* validate the model of separable grammatical competence is questionable. If you construct a test to test a single dimension and then find that it does indeed test a single dimension, how can you conclude that you have shown that this dimension exists independently of other language variables? The unidimensionality, if that is really what it is, is an artifact of the test development. They have not shown that this dimension could be separated from any others because they have not included any others. (In fact I suspect they get closest with item 4.4, which they would like to apologize away.) Lack

of misfitting data cannot be a sufficient condition for acceptance of the grammatical competence model, since: (1) the psychometric model *assumes* unidimensionality, and as the study by Henning *et al.* (1985) shows, the model interprets unidimensionality very liberally, thus it is not at all certain that the model could detect multidimensionality; (2) misfitting data can in any case be explained by a range of causes other than lack of unidimensionality, some of which are indeed offered by Adams *et al.* in explaining away the misfit in their data.

My other concern about this paper is more serious. Adams *et al.* say they chose this dimension to test whether the Rasch partial credit model could be usefully applied in the empirical investigation of the structure of language proficiency. The underlying rationale seems to be that if they could show through this method that one dimension exists, the psychometric procedure could be applied to other dimensions. The concern with grammatical competence, then, grew out of classroom and curriculum. A test was developed which is described (p. 12) as an 'authentic language test', but nowhere in this paper are we given the characteristics of such tests in general, nor of this test in particular, nor of how the test relates to the classroom and curriculum from which it grew. The establishing of the existence of a dimension, if this is indeed established, says nothing about curricular implications, but the impact of testing on curriculum *is* a real concern, and in my view we are always ethically required to take account of the effects and uses of our tests.

It seems from their description that, having found the real world of second language learning in classrooms rather messy and unpredictable, Adams *et al.* imposed their own order. If Adams *et al.* had gone the route of pure research I would have had no quarrel with their study, as long as it was clearly reported as pure research. Such a study would have had only remote classroom implications. However, their statement that 'many linguists and language instructors would gauge this (the kind of test they chose to develop) a controversial or even an incorrect decision, however, our purpose here is to construct one tool that will be useful in language testing, both in the classroom and in research' (p. 13) causes me considerable anxiety. This statement makes it clear that this is *not* metatesting, at least not solely metatesting, but that (apparently) the research aim is coupled with a test development aim. The conclusion makes it even clearer that the instrument being developed will be used not only as an example of the use of the Rasch partial credit model in research into the structure of language proficiency, nor even just to argue that a dimension of grammatical competence exists, both of which are valid research objectives, but also as an operational test of real learners in real

learning and teaching contexts. This returns me to my opening question of what language testing research is for. It may be for wholly theoretical purposes or wholly practical ones, or it may be part of the attempt to move forward our understanding of how language is acquired, used, etc., and then build on that understanding to design new test instruments for use in institutional test administration programmes - but these are not all the same thing and the same constraints do not apply in each case. They must be kept clearly separated in the minds of everyone concerned, and this cannot be done if they are not kept separate in the minds of the researchers themselves. It seems to me that Adams *et al.* have fallen into the trap I described above: they have failed to distinguish consistently between their roles as theoretical researchers and the disinterested questions it is legitimate to ask in that role, and their roles as test developers in which there are much greater constraints on what it is legitimate to do. Having made the decision to move away from the reality of learners, and what and how they learn, they cannot at the end of their research study simply shift gear and change track, as though they had never left the road of pedagogic practice.

All tests create washback, even pure research ones. Throwing the spotlight so powerfully, and only, on a measure which begins with 'isolated elements of vocabulary' and proceeds to 'basic formulaic language and basic structures' (p. 13) makes a statement about how English should be taught. Early on (p. 11) Adams *et al.* state that 'it should be possible to develop teaching programmes around each factor completely isolated and unrelated to programmes for other factors or dimensions'. Of course it is. But what most teachers do not accept is that because it can be done, it should be done. On the contrary, for most teachers and their learners, integration and communicative language teaching have been a liberating influence in the classroom. For me this study, in contrast to that by Pollitt and Hutchinson, is a backward step for both language testing and language teaching.

References

Adams, R.J., Griffin, P.E. and **Martin, L.** 1987: A latent trait method for measuring a dimension in second language proficiency. *Language Testing* 4, 8–27.

Bachman, L.F. and **Clark, J.L.D.** 1987: The measurement of foreign/second language proficiency. *Annals of the American Academy* 490, 20–33.

Hamp-Lyons, L. 1988: Proficiency, profiling and M2. *ELTS Research Reports* 1(ii). London: The British Council; and Cambridge: University of Cambridge Local Examinations Syndicate.

Henning, G., Hudson, T. and **Turner, J.** 1985: Item response theory and the assumption of unidimensionality for language tests. *Language Testing* 2, 141–54.

Pollitt, A. and **Hutchinson, C.** 1987: Calibrating graded assessments: Rasch partial credit method analysis of performance in writing. *Language Testing* 4, 73–92.

Stevenson, D.K. 1981: Language testing and academic accountability: on redefining the role of language testing in language teaching. *IRAL* 19, 15–30.

Weir, C. 1986: Construct validity. *ELTS Research Reports* 1(ii). London: The British Council; and Cambridge: University of Cambridge Local Examinations Syndicate.

White, E. 1985: *Teaching and assessing writing*. San Francisco: Jossey-Bass.

University of Michigan *Liz Hamp-Lyons*