Measuring pedagogic practice for widening participation in mathematics

By
Maria Pampaka*, Julian Williams, Graeme Hutcheson, Pauline Davis, and Geoff Wake

Under review in
Assessment in Education: Principles, Policy & Practice

Abstract

A new self-report instrument was constructed to measure the 'pedagogy' of mathematics teachers for pre-university programmes in the UK. Data from 110 cases of pedagogy are employed to validate this measure using the Rating Scale Rasch model. We report on the outcomes of this analysis that resulted in a one-dimensional measure of what we call 'teacher-centrism' (ranging from 'teacher- centred' to 'learner- centred'). The validation process is also enriched with case study data from teachers and results of a parallel student survey. The paper concludes with the methodological and educational significance of the study.

*** Corresponding author**
School of Education
Room B4.10 Ellen Wilkinson Building
University of Manchester
Oxford Road
Manchester M13 9PL
Email: maria.pampaka@manchester.ac.uk

1. Introduction: The need for a measure of pedagogic practice

We report the construction and validation of an instrument used to measure teacher-centered-ness or 'teacher-centrism' of pedagogy in a very particular, widening participation, project context. We seek to show how its method of development involved interactive processes of qualitative interpretation and statistical modelling, and the role this validation process plays in helping our research to make substantive claims about widening participation in mathematics. The qualitative data came from case studies of classrooms and institutional practices in five of the institutions involved in our project, while the statistics came from surveys of almost 1800 students taught in over 100 classrooms by 90+ teachers in 39 Colleges. Thus, the study is about measuring pedagogic practice and involves the interplay of these different data sources.

The need for this instrument arose in the context of a research project which aims (i) to understand how cultures of learning and teaching can help widen and extend participation in mathematically demanding courses in Further and Higher Education, and (ii) to measure the effect on learner outcomes of pedagogy and programmes designed to widen participation. The project contrasts a traditional mathematics programme with a new 'use of mathematics' programme in the U.K. at "AS" level, i.e. for students usually aged 16-19 who have completed compulsory schooling and have chosen to study some mathematics further. The 'traditional' (Trad) programme is designed to help students prepare for mathematically-demanding courses in university. The new 'Use of Mathematics' (UoM) programme is designed to widen participation to include those who may need to use mathematics in the future generally, but who may or may not progress into very mathematical courses and who may not have very strong previous mathematical backgrounds. This course is designed, then, to make mathematics more accessible to a wider range of students, and it tries to achieve this by providing opportunities for students to develop understanding through 'uses' of mathematics, mathematical modelling, uses of technology (e.g. graphic calculators and spreadsheets) and assessment via comprehension tests and coursework. Collectively we view these as 'modelling and communicative' mathematical practices. We have shown elsewhere that this programme 'works' for retention, and that it enables students with weaker math background to participate (Williams et al., 2008).

In our case study colleges we had the opportunity to observe teaching and learning in some considerable detail and use an analytic framework that includes dimensions of pedagogic practices and mathematical narrative to attempt to make sense of the ways in which teachers engage students in learning new mathematics. This allowed us to identify ways in which the different mathematics programmes might constrain or afford different pedagogies. In fact, we observed in pilot studies that while most traditional teaching is 'transmissionist' and 'teacher-centred', there were teachers of

traditional mathematics courses who described their teaching as 'connectionist' and whose practices were consistent with this description (Swan, 2006; Askew et al., 1997): this approach is student-centred and involves students in conceptually-focussed discussion that is certainly 'communicative', but not necessarily use- or modelling- orientated. We therefore set about measuring pedagogic practice that might gain purchase on this dimension. The research question is threefold:

- How can we measure the pedagogy?
- How does it relate to the two programmes (UoM, Trad)?
- How does pedagogy affect learning outcomes?

However, it was clearly impractical to measure teachers' practices directly across the wider sample of some thirty plus additional colleges: our solution was to develop an instrument to measure the balance of the pedagogies a teacher employs (according to the teacher's self-report) which can be used as an explanatory variable in modelling the student dispositions and achievement (as will be presented in more detail below).

In this paper, therefore, we report how we constructed and validated an instrument to measure the pedagogic practices of teachers of pre-university mathematics. Following a brief review of existing instruments of teachers' pedagogy, we move on to a description of our methodology. In the main part of the paper we deal with the validation procedures employing a Rasch measurement methodology and then triangulate these results with qualitative data from interviews with some of the teachers and case study findings, reported in detail elsewhere (Williams et al., in press). Finally, we illustrate the applications and implications of the use of this measure of teachers' pedagogy.

2. Existing Literature on Pedagogy Instruments

Our framework for the conceptualization of teaching practices began with the concepts of 'connectionist', 'discovery' and 'transmissionist' practices developed by Swan (2006) in the sixth form and Further Education College (6fFEC) context, but drawing on previous literature relevant across the age range even to Primary pedagogy (Askew, Brown, Rhodes, Johnson, & Wiliam, 1997). A very common categorisation of classroom practices in the relevant literature is that between "teacher-centered" and "learner-centered" instruction or practices. Both are broadly applied to include a variety of views and strategies for teaching and learning (Cuban, 1983; Kember & Gow, 1994). The first, according to Schuh (2004), is usually associated with 'transmission' models of teaching where teacher and instruction are the focus, whereas 'learner-centered' practices move the focus to students and learning. In a similar vein Roelofs, Visser, and Terwel (2003) report on six dimensions in which learning environments can differ and they describe them as opposites. However, the authors stress that they can also be considered as continua with two extremes, representing the

transmission model for the latter opposite each time (e.g. construction Vs transmission of knowledge, personal Vs teacher-led meaning, etc).

Other 'opposites' are quite often met in literature that contrasts some kind of reform teaching with the existing dominant 'traditional' practice: 'facilitating as against 'telling'; 'scaffolding' (from a Vygotskian perspective), 'individualised learning approaches' (Bell, 1993a, 1993b), discussion-based approaches (Swan, 2000), guided discovery teaching approach, and 'dialogic teaching' (Ryan & Williams, 2007).

A literature review revealed a variety of instruments designed to measure classroom practices, either based on the above distinctions or more generally oriented. A classification of these instruments and the relevant literature gives three broad categories briefly summarised here. First, there are instruments for teachers used as part of broader international studies like PISA (OECD, 2003) and TIMSS (Hiebert et al., 2003; Mullis et al., 2000; NCES, 2000; Webster & Fisher, 2003). Second we find commercially developed instruments are also available, like those emanating from Horizon Research, Inc. (Horizon Research, 1996) and their modifications (McCaffrey et al., 2001). Third, there is a variety of instruments referred to and employed by other researchers in a variety of contexts and levels mostly in small single-studies (Harwood, Hansen, & Lotter, 2006; Roelofs et al., 2003; Swanson & Stevenson, 2002).

None of these instruments (and others that could not all be mentioned here) fulfil our needs, either because they are not applicable to the British post – compulsory system or because they were not developed for mathematics lessons. They mostly focus on measuring teachers' beliefs rather than practices and they are usually targeted at younger students in compulsory schooling. We should also note Argyris and Schon's (Argyris & Schon, 1978; Schon, 1990) distinction between espoused theory and theories in action. In this context, we must think of the teachers' self-report as being their account of what they do to us, refracting their 'espoused theory' of teaching practice, in relation to the items in the instrument that refer them to their concrete, practical actions.

3. Methodology

3.1. The instrument development

For our study, in post-compulsory, "advanced" level mathematics college classrooms, the work of Swan (2006) was particularly appropriate. Conceptually, Swan built on the research findings of Askew et al. (1997) and Ernest (1991) as a basis for the development of his instruments and the interpretation of his results. He adapted three components that can be used to characterize the

teachers' belief system (i.e. the nature of mathematics as a subject, the nature of mathematics teaching and the nature of the processes of learning mathematics). From the work of Askew and colleagues he derived the 'ideal' categories of teachers' orientation towards each component (i.e. Transmission, Discovery and Connectionist). We decided to use the items of Swan's 'practice scale' (Swan, 2006) because this instrument was designed to evaluate the implementation of a national development project aiming to promote students' understanding of GCSE[1] algebra through collaboration and discussion using almost exactly the same population of Mathematics teachers as our study: therefore the items were highly appropriate for our target teacher population.

An analysis of Swan's (2006) original[2] data with 120 teachers showed acceptable fit to a one-dimensional scale using Rasch modelling techniques, but also suggested some modifications. Even though Swan finally used 25 out of his initial list with 28 we decided to pilot his original including the three he deleted. Preliminary findings highlighted some items that may be problematic; hence we slightly reworded 6 of them. This revised version of the instrument was then given to a teacher conference to test its face validity. The items were presented to the teachers in the form of statements asking them to report the frequency with which certain activities take place in their classroom (using a 5-point Likert scale). An example item is the following: "Students work through exercises" (Almost never, Occasionally, About half the time, Most of the time, Almost always). The functionality and usefulness of the instrument were discussed and teachers commented on the presentation and meaningfulness of the items. This led to the split of an item into two and the deletion of another item, and hence the final instrument, which was then used in the main survey study with the second, 'end of year' student survey point.

3.2. Data sources

A total of 28 items were selected for the instrument that was sent to the teachers of the (approximately 1800) students involved in our cross-college student questionnaire survey. Teachers were asked to complete one survey for each of the mathematics classes they teach, so that students could be matched to the corresponding teacher's practice. Data for this paper come from 110 'cases' of pedagogy from 31 of the 39 further education colleges in the UK. We use 'case' (for most of the analyses) instead of 'teachers' because teachers were asked to report a different survey for each of the surveyed classes that they taught, hence our unit of analysis is a classroom case of pedagogy. There was a total of 95 individual teachers in our sample, and 12 of them responded more than once (9 twice and 3 thrice). Some of the colleges of our survey sample teach both the traditional AS 'Mathematics' (hereafter ASTrad) course as well as AS "Use of Mathematics" (hereafter UoM). In

[1] GCSE is a major end qualification for the compulsory phase of education in the UK (approximately for age 16)

total we have 31 cases of UoM pedagogy and 78 of AStrad. We additionally draw on interviews and observations of pedagogy with nine of these teachers from our case study colleges and will report extracts from these interviews.

3.3. The Rasch Rating Scale Model

The data were analysed using the one-parameter Rasch rating scale model. The Rasch model was selected because it provides the means for constructing interval measures from raw data and because the total raw score is sufficient for estimation of measures. When data can be selected and organized to fit a Rasch Model, the axiom of additive conjoint measurement is satisfied, a Guttman order of response probabilities and hence of item and person parameters is established, and items are calibrated and persons measured on a common interval scale. Models of the Rasch family are hence governed by certain assumptions, the most important of which are unidimensionality, local independence, and common item discrimination. In its simplest form (i.e. for dichotomous responses) the model proposes a mathematical relationship between a person's ability, the difficulty of the task, and the probability of the person succeeding on that task (Acton, 2003; Wright, 1999; Wright & Mok, 2000).

For the analysis and results reported in this paper we employed the Rating Scale Model (RSM, with the FACETS software), from the family of Rasch models, which is an extension of the simple model to rating scale observations like ours (Likert type response format). When the response rating scale works, it yields ordinal data which need to be transformed to an interval scale to be useful. In Andrich's (1999) terms, the response categories serve to define a continuum, and the ratings can be seen as extensions and refinements to dichotomous responses such as disagree and agree. The transformation of the ratings to measures that gives the probability of passing or failing each threshold or step is given by the following model (Linacre, 2003):

$$\log (P_{nik}/P_{ni(k-1)}) = B_n - D_i - F_k$$

where:
D_i is the item's difficulty;
B_n is the person's ability;
P_{nik} is the probability of observing category k for person n encountering item I;
$P_{ni(k-1)}$ is the probability of observing category k-1; and
F_k is the difficulty of being observed in category k relative to category k-1.

The model allows the item difficulty of each question or statement to be based on the way in which an appropriate group of subjects actually responded to that question in practice. The model establishes the relative difficulty of each item stem in recording the development of an attitude from the lowest to the highest levels the instrument is able to record (Andrich, 1999; Bond & Fox, 2001;

Wright & Mok, 2000). It should also be noted that for the purposes of this analysis and in order for results to be meaningful, the scoring of some items was reversed.

4. Analysis and Results: initial validation process

By 'validation process' we refer to the accumulation of evidence to support validity arguments. Our psychometric analysis for this purpose was conducted within the Rasch measurement framework and therefore we follow the guidelines summarised by Wolfe and Smith Jr, (2007a, 2007b) based on Messick's (1988, 1989) validity definitions. Here we employ a measurement perspective on the issue of validity, drawing on the means provided by Rasch modelling, and interpreting our statistical results with qualitative evidence and judgement. This section deals with the procedures we employed for ensuring construct and communication validity, as well as the methods for checking for the applicability and appropriateness of our measure for different groups of teachers. The presentation and interpretation of the constructed measure is also enriched with qualitative data.

4.1. Construct Validity (Ensuring Unidimensionality)

In the Rasch context fit statistics indicate how accurately the data fit the model. Fit statistics are local indicators of the degree to which the data is cooperating with the model's requirements. Inconsistent data (e.g. misfit items or persons) may become a source of further inquiry. Fit statistics may also flag items to which responses are overly predictable (overfits), an indication that, in some way, they are dependent on the other items and might be the first choices for deletion (Bowles, 2003; Wright, 1994).

In this case, Rasch analysis showed acceptable fit of almost all the items suggesting that they could constitute a scale, i.e. they measure teachers' self reported pedagogic practices in mathematics classrooms in 6fFECs. A significant exception to this was one item which presented an INFIT meansquare of 2.4 (B27: "I encourage students to discuss the mistakes they make") and it was decided to delete it from the scale. The decision was not based on statistical results alone but also on the fact that some teachers said they found this item ambiguous. Re-calibration of the remaining items was performed and the results (fit statistics and items measures) are shown in Table 1 (item fit and measure).

Five possibly misfitting items (B6,B8,B22,B24,B26) are now highlighted within the results in Table 1 and contextually presented at the bottom. On theoretical and methodological grounds we decided not to exclude these items at this point. Gustafsson (1980) concluded that items which can be identified as misfitting should not be routinely excluded to obtain fit to the model, but instead other actions

should often be taken such as grouping of the items into homogeneous subsets. Wright (1994) also notes that fit statistics provide guidance, but not decisions. A substantive explanation of data identified as irregular is more important than the absolute magnitude of significance level of fit statistics. In sum, we follow Bohlig et al. (1998) who conclude that *"less than pleasing fit statistics say 'think again', not 'throw it out'"* (p. 607), and hence we seek explanations and interpretations for these high fit values.

Table 1: Measures and fit statistics for the items of the scale

Item	Entry No	Raw Score	Count	Measure	SE	Infit		Outfit	
						MNSQ	ZSTD	MNSQ	ZSTD
B1	1	355	110	.42	.10	0.8	-2	0.8	-1
B2	2	339	110	.57	.10	0.9	0	1.0	0
B3	3	431	110	-.44	.12	0.8	-1	0.9	-1
B4	4	435	110	-.50	.12	1.2	1	1.1	0
B5	5	452	110	-.83	.13	1.0	0	0.9	0
B6	6	456	110	-.84	.13	1.5	2	1.5	2
B7	7	411	110	-.18	.11	0.7	-2	0.7	-2
B8	8	405	110	-.11	.11	1.4	2	1.3	2
B9	9	434	110	-.49	.12	0.5	-3	0.5	-3
B10	10	426	110	-.43	.12	1.1	0	1.1	0
B11	11	295	110	.96	.09	1.0	0	1.0	0
B12	12	391	110	.05	.10	0.7	-2	0.7	-2
B13	13	367	105	.07	.11	1.2	1	1.3	1
B14	14	268	110	1.18	.10	1.2	1	1.2	1
B15	15	302	110	.90	.09	0.7	-2	0.7	-2
B16	16	348	110	.48	.10	0.7	-2	0.8	-2
B17	17	460	109	-1.06	.14	0.8	-1	0.8	-1
B18	18	405	110	-.15	.11	1.2	1	1.2	1
B19	19	426	110	-.43	.12	1.0	0	1.0	0
B20	20	317	110	.76	.09	0.8	-2	0.8	-2
B21	21	347	109	.45	.10	0.9	0	0.9	0
B22	22	398	110	-.03	.11	1.4	2	1.4	2
B23	23	382	109	.06	.11	1.2	1	1.2	1
B24	24	392	109	-.04	.11	1.5	3	1.5	3
B25	25	391	110	.05	.10	0.9	0	0.9	0
B26	26	444	109	-.77	.13	1.4	2	1.2	1
B28	27	359	110	.34	.10	1.1	0	1.1	0
			Mean:	.00	.11	1.0	-0.1	1.0	-0.1
			SD:	.58	.01	0.3	1.9	0.3	1.8

RMSE (Model) .11 Adj S.D. .57 Separation 5.13 Reliability .96
Fixed (all same) chi-square: 742.2 d.f.: 26 significance: .00
Random (normal) chi-square: 26.0 d.f.: 25 significance: .41

B6: I encourage students to work more slowly.
B8: I teach each topic from the beginning, assuming they know nothing.
B22: I find out which parts students already understand and don't teach those parts.
B24: I cover only the important ideas in a topic.
B26: I know exactly what maths the lesson will contain.

The items B8 and B22 clearly signal an approach to formative assessment: this is an important part of 'connectionist' and 'dialogical' teaching but may not be present in all student-centred teaching, which after all can include pure 'discovery' approaches. Similarly B6, B24, and B26 all have meaning to a connectionist teacher and may belong in the construct of connectionist teaching, and yet they may be interpreted otherwise by some teachers:

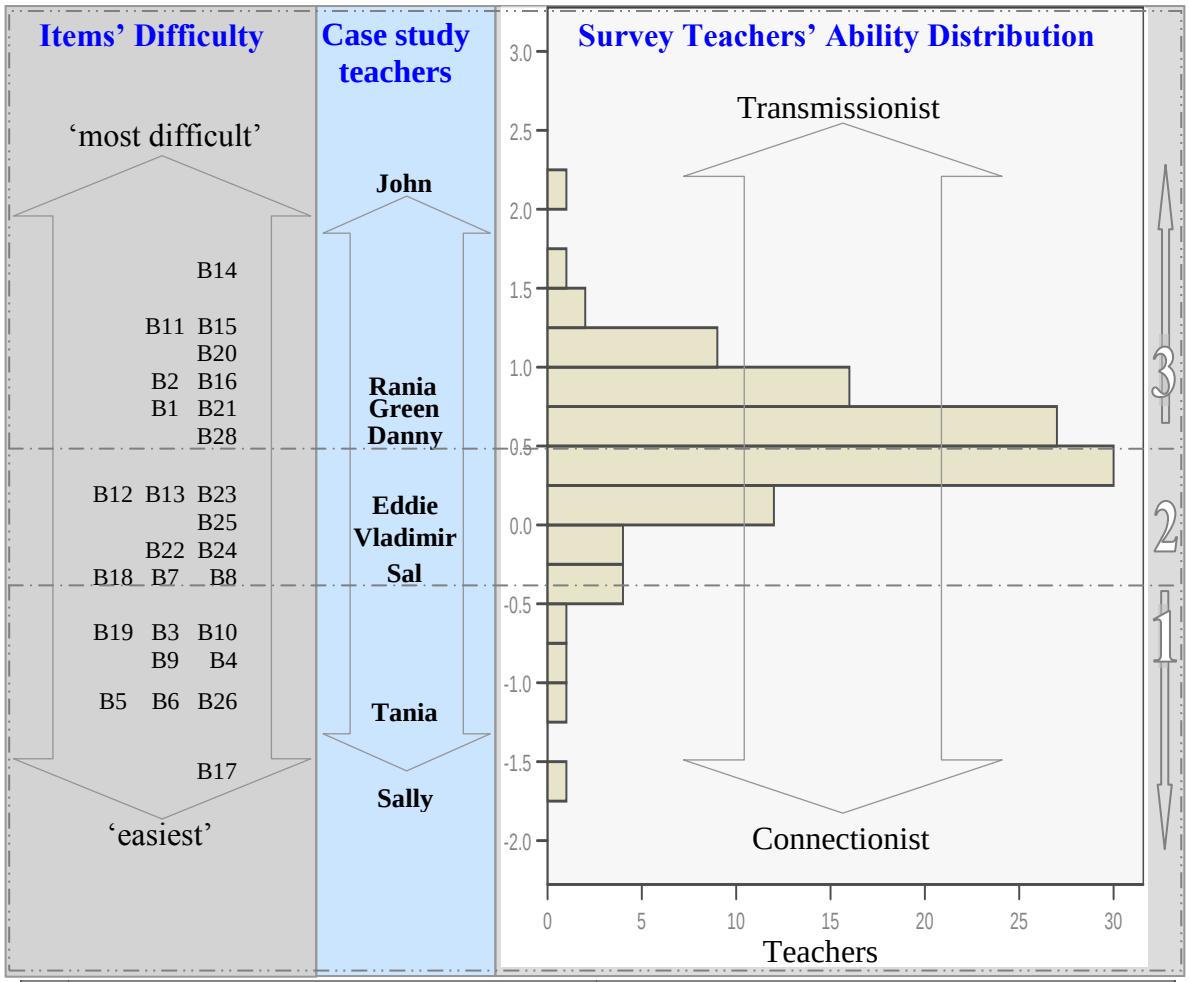
- B6, B24: 'working more slowly' and 'covering the most important ideas' (only) in some contexts may be seen as laissez-faire rather than encouraging more thoughtful work. The fact that these items fit less well than B20 ('I encourage students to work more quickly') may support this;
- B26 ('knowing exactly what maths the lesson will contain' similarly might be viewed as signalling a transmissionist (not connectionist) attribute, suggesting the lesson is controlled to ensure that the mathematics does not include non-standard methods, misconceptions or alternative approaches. But it may also be viewed by some teachers as an issue of teachers' subject mater knowledge, i.e. that they should know all the mathematics that 'might' arise.

Because these elements of connectionist practice were viewed as important by some connectionist teachers in our study, for the time being we decided not to exclude these items, as any construct-irrelevant variance they introduce might be balanced by construct validity arguments. We return to this issue when considering a reduced scale.

4.4. Description of the constructed measure

Figure 1 shows the resulting measurement scale. At the right side of the figure, the distribution of measures is shown (as a histogram). The higher the place of the "practice", the more teacher centered the pedagogy. Pedagogy that is mainly student-centered is at the bottom and pedagogy that is mainly 'teacher-centered' is at the top. On the left hand side of the figure the items that constitute the scale are presented, ranging from those easiest to report as frequent to the most "difficult" to report being frequent. For reversed items the opposite happens, e.g. B17[3].

[3] Note from here on for ease of interpretation we have usually added the term [don't] in square brackets for those reverse-coded items, so that high frequencies on all items imply "more transmissionist" practice. Thus the reverse-coded item "I draw links between topics…" becomes "I [don't] draw links between topics…"



B14	I tend to follow the textbook closely.	B24	I cover only the important ideas in a topic.
B11	I draw links between topics and move back and forth between topics.	B8	I teach each topic from the beginning, assuming they know nothing.
B15	Students discuss their ideas.	B18	Students work on substantial tasks that can be worked on at different levels.
B20	I encourage students to work more quickly.	B7	Students compare different methods for doing questions.
B2	Students work on their own, consulting a neighbour from time to time.	B10	I try to cover everything in a topic.
B16	Students work collaboratively in pairs.	B19	I tell students which questions to tackle.
B21	I go through only one method for doing each question.	B3	Students use only the methods I teach them.
B1	Students work through exercises.	B9	I teach the whole class at once.
B28	I jump between topics as the need arises.	B4	Students start with easy questions and work up to harder questions.
B13	I avoid students making mistakes by explaining things carefully first.	B26	I know exactly what maths the lesson will contain.
B23	I teach each student differently according to individual needs.	B5	Students choose which questions they tackle.
B12	Students work collaboratively in small groups.	B6	I encourage students to work more slowly.
B25	I teach each topic separately.	B17	Students invent their own methods.
B22	I find out which parts students already understand and don't teach those parts.	*	Items in highlighted cells are reverse coded for analysis

Figure 1: The 'teacher-centrism scale'

Three hierarchical levels of this 'teacher centerness' measure can be distinguished (as shown on the right side of Figure 1, by the dotted lines) based on both the statistical results and the qualitative analysis of the homogeneity of the item content, giving three levels. The statistics that support this hierarchical progression of our measure, based on the Rasch modelling, are the person and item

separation indices (mostly item separation in this case), which provide estimations of the persons or items on the measured variable (Wright & Masters, 1982). 'Separation' can be thought of as the number of statistically significant levels into which the sample of items and persons can be separated. For an instrument to be useful, separation should exceed 1.0, with higher values of separation representing greater spread of items and persons along a continuum (Green & Frantom, 2002). In our case item separation is more than 5 which could allow us to distinguish between 5 levels of groups of items, whereas person separation is 2.10 which allows us to develop three distinguishable levels of persons at most. A qualitative analysis of the items also resulted in just three categories which are separated and detailed in Table 2.

In the centre of Figure 1, the location of the pedagogies found in our case study colleges are tested to check for the validity and meaningfulness of our measures. Each pedagogy is identified with teachers we interviewed and observed teaching (see Williams, et al., in press). It should be noted that the two teachers who define the ends of the scale (maximum and minimum score) are within our case study sample. Apart from one case, all these teachers have 'acceptable' person fit. Table 2 gives some extracts from the interviews with four of the teachers to illustrate how their narratives match the measures obtained by their self-reported surveys (also reported here are their measure in the scale – TC – and their fit statistic). Some key points are underlined.

Table 2 illustrates how some of our case study teachers talk about their practices and at the same time their scores on the pedagogy measures as well as their denoted level. Sally points out how important formative assessment is to connectionism, for instance, thus justifying our inclusion of items B8 and B22 for instance. A more detailed study of two extreme cases is presented elsewhere (Williams et al., accepted).

Table 2: Qualitative justification of the validity of the pedagogy measurement scale

Teacher	Interview Extracts
Level 3 (N=56)	**Teacher centred, transmissionist, fast paced, exam orientated teaching**
John [TC = 2.08 logits INFIT =1.1]	**Highest End** "…I do tend to teach to the syllabus now…If it's not on I don't teach it. … but I do tend to say this is going to be on the exam, it's going to be worth X number of marks, that's why we're doing it." "It's old fashion methods, there's a bit of input from me at the front and then I try to get them working, practising questions as quickly as possible, …"
Green [TC= 0.77 logits INFIT=0.8]	**Lower End** **"**I follow the text fairly carefully…[ ] A lot of things like with the differentiation, we used to introduce incredibly carefully with the chord into a tangent and that sort of thing, I doubt I can do that. I should just announce this is it.
Level 2 (N= 46)	**Involves teacher practices from both ends of the spectrum in moderate frequencies** (e.g. small group work, but also emphasises 'careful explanations by the teacher)
Eddie [TC=0.32 logits INFIT=1.3]	"The purpose was to first of all go through the introduction lesson to introduce the new concepts of the average …. once they have grasped those concepts and seen it attached to that real life problem then the main aim of the lesson was then to look at similar problems." "They either do it individually or some of them like to work in pairs which I am happy for them to do. If you look around the room normally there is a cluster of pairs helping each other. I have to keep an eye on them and make sure that one is not doing all the work …."
Level 1 (N=8)	**Frequently student-centred, more connectionist practice**
Sally [TC=-1.62logits INFIT=0.9]	"… there's a sense that I've achieved the purpose…I've found out what they've come with and what they haven't come with so…we can work with that now" "…. from the teachers that I've met and talked to… it seems to me that one of the big differences is, I mean I don't sort of use textbooks… [ ]…I want to get students to think about the math, I want students to understand, I want students to connect ideas together, to see all those things that go together and I don't think a text book did that…[ ].

5. Further validity checks

5.1. Ensuring local independence

Given the structure of our sample and the fact that the same teachers appeared in different cases there is a need to check that the assumption of local independence is not violated. Therefore, analysis was also run considering only the 95 teachers of the sample once only (selecting randomly one class for each teacher who sent more than one survey back to us). The item parameters from this analysis were plotted against those from all cases as shown in Figure 2. This suggests that the threat of dependence is minimal. However, this raises an interesting issue for future study: how much can pedagogic practice be attributed to the teacher and how much variation is there between different classes?

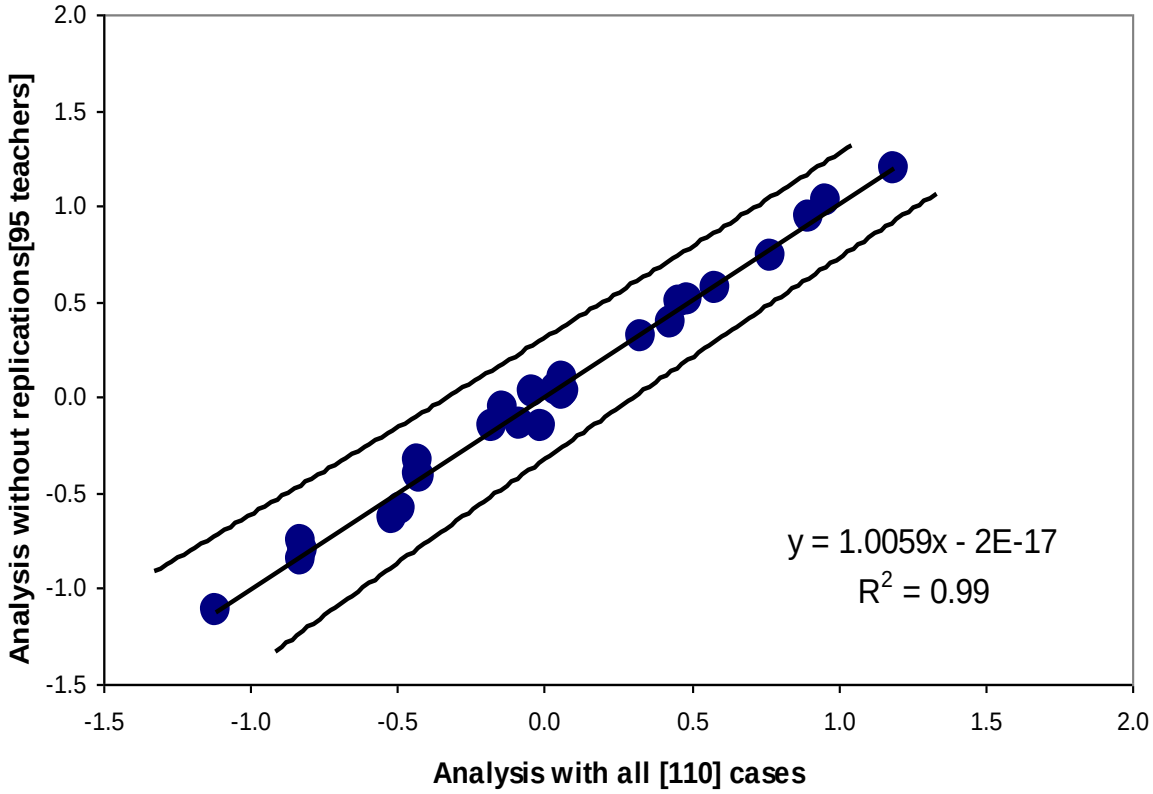


Figure 2: Item estimates for the two analyses (with the 95% confidence intervals)

5.2. Response format and 'listening' to the respondents

Rating scales and their response formats serve as tools with which the researcher communicates with the respondents. Lopez (1995; 1996) defines as 'communication validity' the extent to which the rating scale's categories perform as intended. Thus, category statistics are also examined for the appropriateness of the Likert scale used and its interpretation by the respondents, with the aid of Rasch analysis which provides the means for these checks (See Table 3).

Table 3: Category Statistics for the pedagogy measurement scale

Category label	Observed		Quality Control			Threshold /Step calibration		Thurstone threshold at
	Count*	%	Average Measure	Expected measure	OUTFIT Mnsq	Measure	S.E.	
1-Almost never	164	6	-.52	-.62	1.3	(none)		(Low)
2-Occasionally	436	15	-.11	-.14	1.0	-1.35	.09	-1.62
3-About half the time	571	19	.17	.28	.8	-.19	.05	-.50
4-Most of the time	1338	45	.69	.67	1.0	-.37	.04	.11
5-Almost always	466	16	1.06	1.04	1.0	1.91	.05	2.00

[*count refers to the total use of the response category among all respondents to all items, after recoding]

The most often used indices for this check are the average measure and the threshold (or step calibration). The average measure is approximately the average ability of the respondents observed in a particular category, averaged across all occurrences of the category, whereas the threshold is the location parameter of the boundary on the continuum between category k and category k-1 of a scale (Linacre, 2002). A well functioning scale should present ordered average measures, with acceptable

fit statistics, as happens with our case. It should also present ordered step calibrations, which seems problematic in our case, due to more frequent use of step 4 (i.e. 'most of the time'). Thurstone threshold (the measure at which the probability of being rated in this category or above equals that of being rated in any of the category below) should also be ordered, as here.

Categorization is crucial in designing any ordered-response scale (including the rating scale) and it has two important characteristics. First, while all categories of a scale should measure a common trait or property, each of them must also have its own well-defined boundaries, and the elements in a category should all share certain specific exclusive properties. Second, categories must be in an order, and numerical values generated from the categories must reflect the degrees or magnitudes of the trait. Factors that have been found to affect the categorization of a scale, among others, are the number of categories, their label and their position (Zhu, 2002; Zhu, Updyke, & Lewandowski, 1997). The label seemed to be the problem here (according to some of the teachers) regarding option 3 ("about half the time"). For some items 'about half the time' implies 'for about half of the lesson' in some cases, or 'in about half the lessons', in other: the problem comes when both interpretations are reasonable.

These results suggest that for future use of the instrument we should consider changing the 5-point Likert Scale to a 4-point one by collapsing adjacent categories, a common practice among the users of Rasch model, when rating scale diagnostics indicate that some categories were used infrequently or inconsistently by the respondents (Bond & Fox, 2001; Lopez, 1995; Lopez, 1996; Zhu, 2002; Zhu et al., 1997). The first checks in this direction are performed with our existing data.

Revising the rating scale

The functionality of a revised response format is checked within this section. Theoretically it makes more sense to collapse response categories 3 and 4, and teachers supported this 'common sense' view. This means that the original 5 point response format [12345] is transformed into a 4 – point format by joining together the responses 3 and 4 into the same category [hence the collapsing 12334] Analysis was performed again with this collapsing scheme. Results in Table 4 show that with this adjustment the problem of unordered step calibrations is overcome.

Table 4: Category Statistics for the revised 4-point Scale

Category label	Observed		Quality Control			Threshold /Step calibration		Thurstone threshold at
	Count*	%	Average Measure	Expected measure	OUTFIT Mnsq	Measure	S.E.	
1[Almost never]	164	6	-.62	-.81	1.3	(none)		(Low)
2 [Occasionally]	436	15	-.09	.00	.9	-1.37	.09	-1.79
3 [Most of the time]	1909	64	.72	.73	1.0	-1.11	.05	-.72
4 [Almost always]	466	16	1.46	1.40	1.0	2.48	.05	2.49

Table 4 shows that all measures that define the functionality of the proposed response format are now ordered. The less healthy result is the quite high Outfit value for response category 1 (as in previous table) but this is only because 6% occurrence is so small and should not concern us. However, we keep this category in our response format mainly because of the significance of the extremes to our modelling. The item parameters for the revised measurement scale are presented in Table 5.

Table 5: Measures and fit statistics for the items of the scale

Item	Entry No	Raw Score	Count	Measure	SE	Infit		Outfit	
						MNSQ	ZSTD	MNSQ	ZSTD
B1	1	316	110	.23	.16	0.8	0	0.8	-1
B2	2	299	110	.61	.15	0.9	0	0.9	0
B3	3	342	110	-.48	.17	0.7	-1	0.8	-1
B4	4	344	110	-.54	.17	1.2	1	1.3	1
B5	5	354	109	-.93	.18	1.0	0	1.0	0
B6	6	362	110	-1.08	.18	1.4	2	1.4	2
B7	7	327	110	-.06	.16	0.7	-1	0.7	-1
B8	8	328	110	-.08	.16	1.4	2	1.4	1
B9	9	347	110	-.63	.17	0.5	-4	0.5	-4
B10	10	339	109	-.48	.17	1.1	0	1.2	0
B11	11	265	110	1.25	.13	1.0	0	1.1	0
B12	12	324	110	.02	.16	0.8	-1	0.7	-1
B13	13	302	105	.13	.16	1.3	1	1.4	2
B14	14	248	109	1.50	.13	1.3	2	1.3	2
B15	15	275	110	1.08	.13	0.7	-2	0.7	-2
B16	16	300	110	.59	.15	0.7	-2	0.7	-2
B17	17	362	108	-1.29	.18	0.7	-2	0.8	-2
B18	18	329	109	-.19	.17	1.2	1	1.2	1
B19	19	343	110	-.60	.17	0.9	0	1.0	0
B20	20	281	110	.97	.14	0.7	-2	0.7	-2
B21	21	297	109	.56	.15	0.9	0	1.0	0
B22	22	326	110	-.03	.16	1.4	2	1.5	2
B23	23	311	109	.19	.16	1.2	1	1.2	1
B24	24	323	109	-.11	.17	1.4	2	1.5	2
B25	25	324	110	.02	.16	1.0	0	0.9	0
B26	26	356	109	-1.11	.18	1.3	1	1.2	1
B28	27	303	110	.46	.15	1.1	0	1.0	0
			Mean:	.00	.16	1.0	-0.1	1.0	-0.1
			SD:	.71	.01	0.3	1.9	0.3	1.8

RMSE (Model) .16 Adj S.D. .70 Separation 4.34 Reliability .95
Fixed (all same) chi-square: 596.4 d.f.: 26 significance: .00
Random (normal) chi-square: 26.1 d.f.: 25 significance: .40

As shown in Table 5 collapsing categories causes a small reduction in item separation (compared to the original 5-point response format in Table 1). There is, however a noticeable improvement of the fit of the 'problematic' items highlighted in Table 1 (notice that B26 is not a cause of concern anymore and 1.4 is the highest infit value for all the items).

If this revised instrument is going to be used as an alternative (for future applications) we need to check whether item and person parameters remain invariant between the measure with the collapsed categories and the original five-point format. The correlation of the item parameters was found to be very high (R^2=0.9861, p<0.01) and Figure 3 illustrates the correlation of person estimates when analysis is performed with the two different versions.

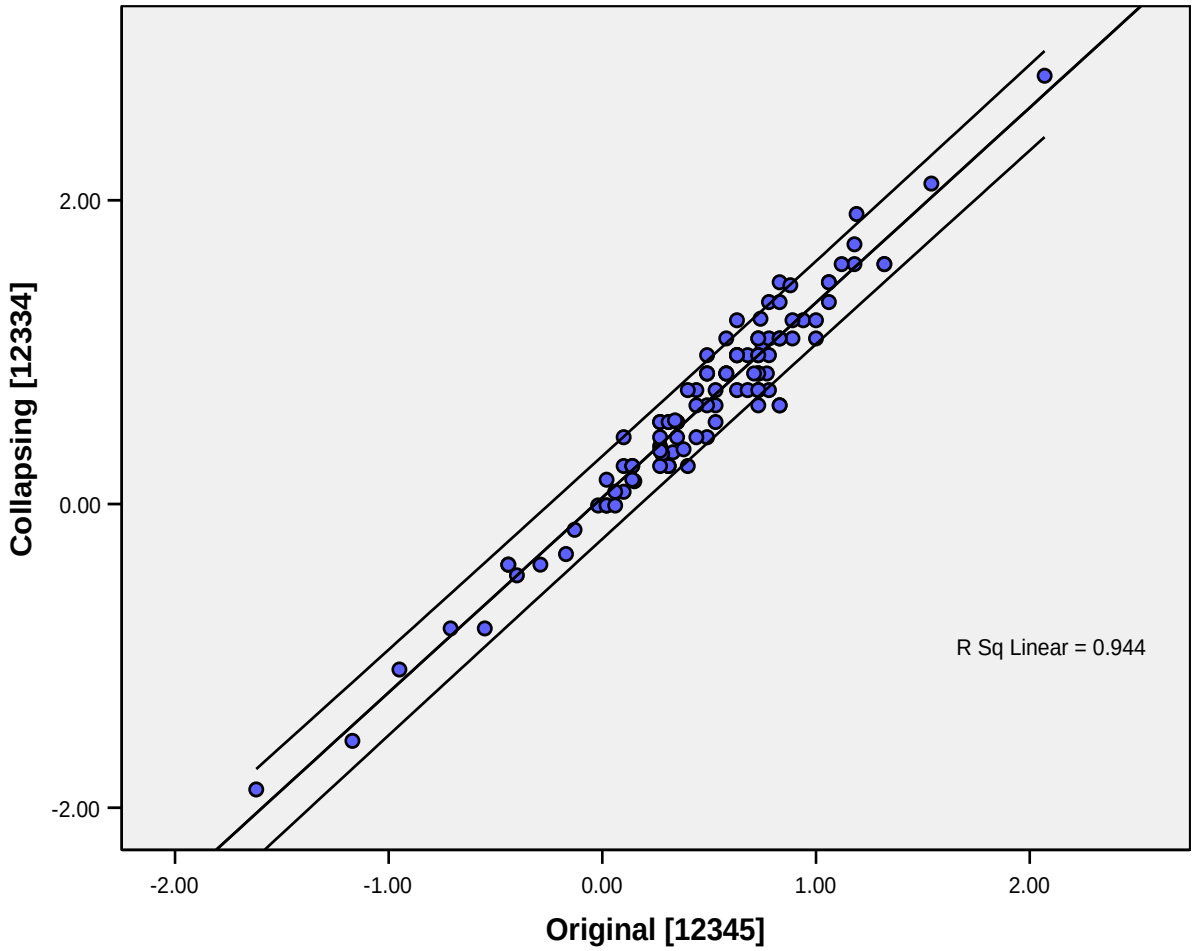


Figure 3: Comparison of person estimates of the original and revised measure

Statistically this is promising. However, future studies however will need to check the implicit assumption here, i.e. that respondents will choose the third category of four, who previously chose either the 3rd or 4th category when there were five!

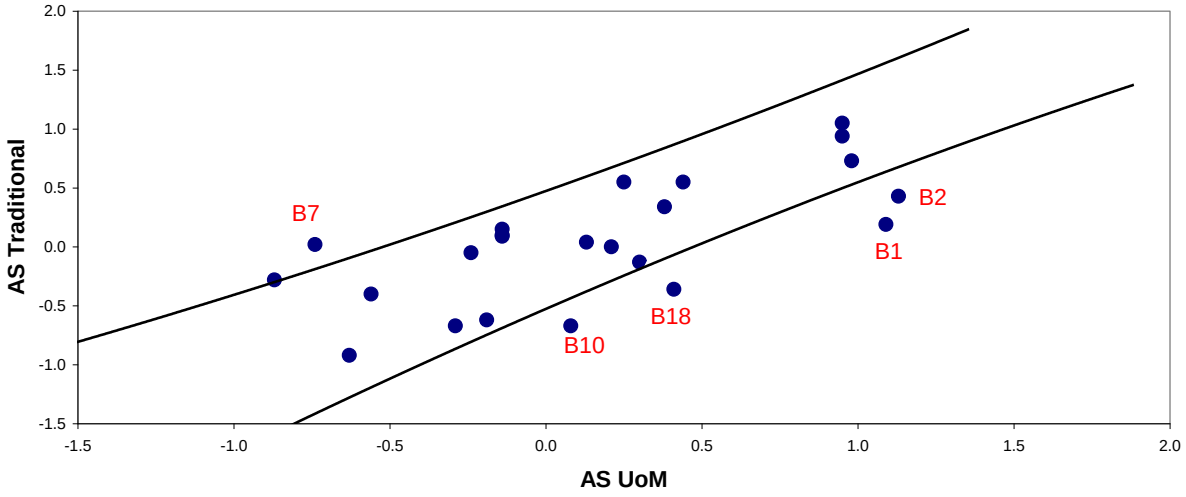
5.3. Validity across different groups

According to Wright and Masters (1982) when a variable is used with different groups of persons [or to measure the same persons on different occasions], it is essential that the identity of the variable be maintained from one occasion to the next (i.e. from group to group). Only if the item calibrations are invariant from group to group can meaningful comparisons of person measures be made. The groups we are interested to check here are teachers of UoM courses compared to teachers of Traditional AS

Maths courses. In particular it is important to compare groups with unbiased instruments, i.e. with items that are of equal 'difficulty' for equally 'able' subgroups.

A statistical way to inform this process is to check for Differential Item Functioning (DIF). DIF is an expression which describes a serious threat to the validity of items and tests used to measure an aptitude, ability or proficiency of members of different populations or groups. DIF measurement may be used to reduce this source of test invalidity and allows researchers to concentrate on the other explanations for group differences in test scores (Thissen, Steinberg, & Wainer, 1993).

There are different methods or techniques to check for DIF. In our case a t-test on the two estimates of difficulty parameters based on the two groups of teachers (see Figure 4, with the lines indicating the 95% confidence intervals in item estimates). The points that are labelled outside the confidence intervals in Figure 4 denote the items with high DIF when comparing the teachers of AS UoM with AS Trad classes.



B1: Students work through exercises
B2: Students work on their own, consulting a neighbour from time to time.
B7: Students [don't] compare different methods for doing questions.
B10: I try to cover everything in a topic.
B18: Students [don't] work on substantial tasks that can be worked on at different levels.

Figure 4: Comparison of item estimates for the two teacher groups

Given the direction of the DIF and the coding of these items it appears that some practices are more frequently reported within the Trad classes. In general, practices denoted by the item on the top (B7) are those that appeared to be more frequently reported in UoM classes (assuming they were not reversed), as opposed to the items on the bottom which are more frequently reported in ASTrad classes. The item on the top though was reverse-coded for analysis, therefore for interpretation purposes all items showed transmissionism to be more frequently reported in ASTrad classes.

Interestingly then, the items that function differently between the two groups of teachers are all in the same direction, and suggest that on these items the "uses" teachers score less highly, i.e. they report lower frequencies of transmissionist practices than the rest of their responses would indicate. The issue then is, for these items, whether they are in themselves 'biassed' or whether they reflect 'real' (i.e. interpretable and credible) differences between the two groups of teachers. Our view is that they are mostly interpretable and credible, because the "Uses" curriculum and assessment allow the teacher to 'engage the students in substantial tasks' in coursework (B18) and 'compare different methods' in modelling (B7), and because the texts in use have generally a greater variety of activities than just 'working through exercises' (B1). The DIF for B2 and B10 is less clear, though the fact that students on the UoM programme are generally weaker students may be relevant.

Generally we suggest that in this case DIF might be due to 'real' differences in the teaching practices that the curriculum affords, and that may suggest a nascent 'second dimension' of self-reported practice that the rest of the scale does not measure. Therefore the multi-dimensionality of the construct is worthy of further study in our view. By removing these items we may then build a measure of one dimension that is more focussed on practices that are less related to the actual differences in the two curricula.

Resolving DIF

The presence of DIF highlights some possible problems for subsequent use of the measures. There may be implications when comparing the two groups of teachers, resulting in limitations of the interpretation of substantive results. Even though the differences are marginal and the implications just hypothetical at this point, we ran the analysis again excluding the items that showed significant DIF. The result is shown in Figure 5.

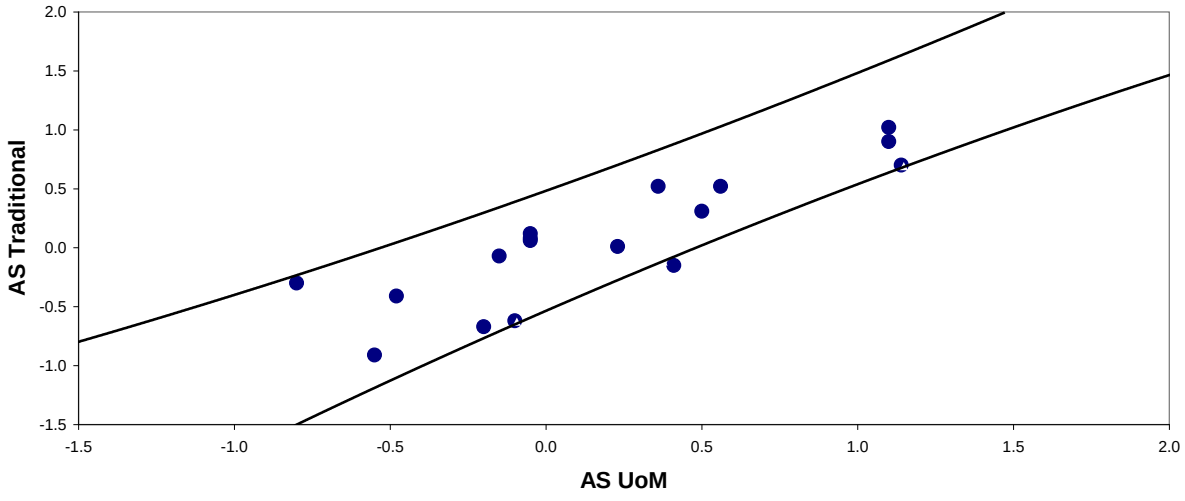


Figure 5: Item estimates for the two teacher groups after deleting items with DIF

Figure 5 shows that DIF is now not significant for the two groups of teachers and Figure 6 indicates that persons' measures remain stable between the two different analyses suggesting that little information on the pedagogy is lost in this reduced scale.

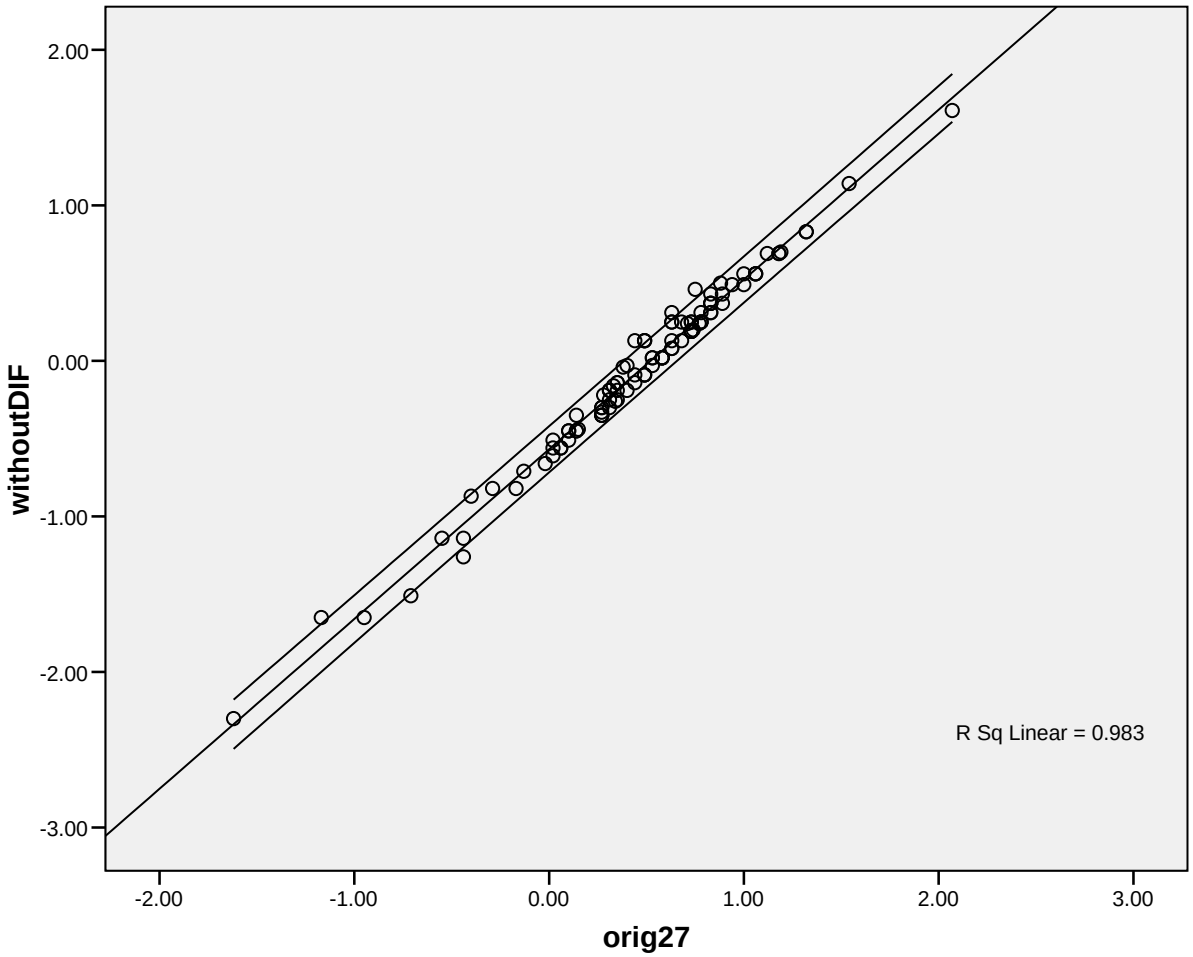


Figure 6: Comparison of person estimates of the measure without DIF items and the original

This provides an alternative measure. It is interesting to compare the teachers of different classes with these two measures. Table 6 shows that the mean measure for the UoM group has increased as expected from 0.47 to 0.54, but that this change of 0.07 logit is not significant (being of the order or 15% of s.e.). We can only speculate that this effect might have been greater if the pedagogy items made more reference to practices such as 'modelling', 'using technology', etc, that we guess might provide a second dimension of practice.

Table 6: Mean measures of the group of teachers for original measure and measure without DIF items

Groups	**Original Measure** Mean (SD)	**Measure without DIF-items** Mean (SD)
All cases (N=110)	0.4746 (0.52)	0.4964 (0.57)
AS Trad only (N=78)	0.4779 (0.54)	0.4808 (0.57)
UoM only (N=31)	0.4663 (0.47)	0.5356 (0.56)

6. Educational Importance of the Measure: Using the constructed measure

The instrument we reported here allowed us to measure with some confidence the student- or teacher-centredness of pedagogic practices. The measure has been used to model the impact of pedagogy on learning outcomes within our project. Here we will highlight some issues and possible implications, using one example of a model where the influence of pedagogy was found to be important.

The first implication of the measure of teacher-centrism comes from simple observation of the scores of pedagogic practices for classes we have observed and where we have interviewed the teachers involved. At this descriptive level it appeared that significantly different scores were reported for different classes - e.g. the extreme cases of Sally's and John's self-reported practice of the traditional course. We not only saw the classroom practices (before even constructing the instrument) but interviewed these teachers on a number of occasions so as to relate these observations with their intentions, beliefs and their professional identity. These cases have been intensively studied and reported (Davis & Williams, in press; Williams et al., in press). In sum, both teachers are regarded as highly effective by their College peers and senior managers, getting good results for their students in examinations. However, Sally is a committed 'connectionist' teacher who believes (and practices her beliefs) in working with small groups, problem-solving, and group- and class-discussions; for her, eliciting a variety of methods and misconceptions is important, and the aim is always to build understandings that under-pin procedures in mathematics, so that learners think 'like mathematicians'. Concepts are never, as far as we observed, simply declared. Rather, a lesson 'narrative' leads the students actively to put together the ingredients and connect them into the conceptual whole (Wake & Pampaka, in press).

On the other hand John emphasises what the students need for the exam, teaching is highly paced and there is often said not to be enough time for concepts, discussion and problem-solving or investigation. John sometimes says there is no time for a proof, and they do not need to know one, as the examiners won't ask for one. He himself is in control of the explanation-giving, while his students mainly have the task of practising exercises that involve procedures for getting the answers to exemplary, test-type questions. He says this is what his College management team and his students want, as it optimises their grades.

We also have interviewed and observed other teachers across the spectrum, and this has given us confidence in our interpretations, and that the interview quotes in Table 3 are reasonably characteristic.

The purpose of the constructed measure was to match pedagogy with the students' learning outcomes. A score of 'pedagogy' is therefore given for each student based on their teachers' responses to the teacher survey (or average pedagogy score when a student has 2 or more teachers for mathematics). Data from students were gathered at two data points (September – November 2006: N=1792 and May-July 2007, N=1049). Hence, the following results are based on the 750 students (475 AS Mathematics, 275 AS Use of Mathematics) for whom we have the matched the learning outcome data with the teachers' pedagogy measure.

At the descriptive level again (students' data) we found that some classes of students showed an overall increase or decrease in their disposition measures (i.e. disposition to go into higher education, disposition to study further mathematically demanding subjects and mathematical self efficacy) between the first two data points of our project. We report elsewhere the development and validation of these learning outcome measures (Pampaka et al., 2007; Wake & Pampaka, 2007; Williams et al., 2007). The survey data analysis employed Generalised Linear Modelling of these learning outcome variables over different time intervals (Hoffman, 2004; Hutchenson & Sofroniou, 1999). The following framework was used for the modelling process (Pampaka et al., 2007):

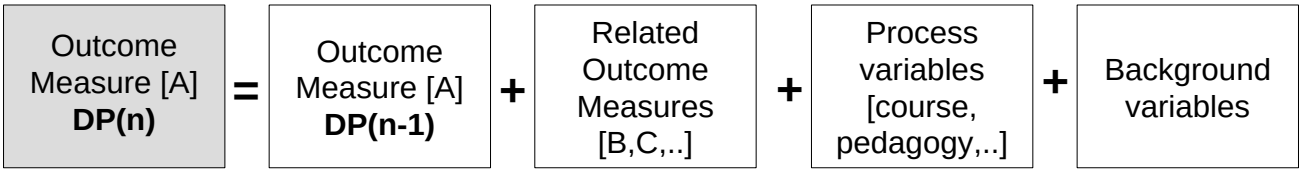


Figure 7: Process for modelling outcome variables at data point (n) regressed on data point (n - 1)

This leads to models of the effects of background variables and earlier 'input' measures (including outcomes from *previous* time intervals) and conditions (e.g. 'Traditional' Vs 'Experimental' programmes and pedagogy) on the outcome measure (disposition or attainment). An important result is that there is no significant additional effect of pedagogy on grades, dropout in any value added model.

However Table 7 shows a model for the value added to maths-disposition (i.e. disposition of the student to continue with mathematically demanding subjects at Data Point 2- DP2). It shows the statistically significant negative effects of pedagogy on students' dispositions to study mathematics further. The addition of background and other variables (i.e. GCSE grade and course) did not produce any other statistically significant effect, and also did not cause any change to the effect in the pedagogy measure.

Table 7: A Regression Model for HE maths-disposition at DP2 (MdispDP2)

	Coefficient B	s.e.	t	p
(Constant)	-0.91847	0.08661	-10.605	< 2e-16
MdispDP1	0.59801	0.03460	17.284	< 2e-16
MSE-DP2	0.34031	0.04706	7.232	1.25e-12
OtherSubjectsMathsDP2	0.14425	0.03073	4.695	3.21e-06
AveragePed	-0.24701	0.07966	-3.101	0.00201

F (4, 702) =127.6, p<0.001, R^2=0.421(Adjusted R^2=0.4177)

Overall we find that the pedagogy has a further negative effect, that is, transmissionist or 'teacher- and subject-centred teaching' is likely to further depress the students' maths-disposition. These models are still under investigation, but the result that pedagogy has a slope of approximately -0.25 is robust across different models including the effect of programme, and background variables related to ethnicity, class and gender (e.g. Table 8).

Table 8: A Regression Model for HE maths-disposition at DP2 (MdispDP2)

	Coefficient B	s.e.	t	p
(Constant)	-0.11448	0.19091	-0.600	0.54892
Ethnicity[4] [T.BLACK]	0.13523	0.24637	0.549	0.58326
Ethnicity[T.CHINESE]	-0.09018	0.43301	-0.208	0.83508
Ethnicity[T.OTHER]	-0.26626	0.29450	-0.904	0.36626
Ethnicity[T.WHITE]	-0.36373	0.18212	-1.997	0.04620
Language[T.ENGLISH]	-0.42708	0.19118	-2.234	0.02581
Language[T.OTHER]	0.41815	0.28278	1.479	0.13968
Course[T.UoM]	-0.77903	0.12870	-6.053	2.35e-09
MSE - DP2	0.46390	0.05344	8.681	< 2e-16
AveragePed	-0.26541	0.09521	-2.788	0.00546
OtherSubjectsMathsDP2	0.14801	0.03604	4.107	4.50e-05

F (10, 680) =24.9, p<0.001, R^2=0.268 (Adjusted R^2=0.257)

Is this slope educationally significant? It corresponds in effect to a difference of about 0.5 to 0.7 logits (as the spread from highly connectionist to the bulk of the teachers is about 2 to 3 logits), which is approximately the same as the difference between dispositions of students doing the two programmes (those following the traditional course are of course relatively highly disposed to do further mathematical study, as the Uses is a terminal course that does not allow students to progress to a full A-level, and these scores are consistently about 0.7 logits higher before and at the end of the course). This should be claimed as educationally significant, if dispositions to study more mathematics are regarded as important learning outcomes.

This effect, however, disappears when the most highly connectionist few classes are removed for the data, implying that most of this effect was caused in these few classes. This explains why we have focused on the extremes of the instrument in its development of shortened version.

[4] Reference categories: Ethnicity = Asian, Language= Bilingual. Also note that 'OtherSubjectMaths is a numeric variable that sums up the mathematicalness of the other courses the students attend.

7. A Short version of the instrument

We now make use of what we know to select and check for the validity of a shorter version of the instrument. The procedure involved an initial reduction to 15 items excluding the more misfiting and overfitting items and ensuring we had items from the whole spread of the scale. This analysis flagged some new items as misfiting which were deleted in a sequence of calibrations, always considering the spread. The remaining 11 item – instrument was the final short-version selected because additional reduction would begin to exclude items from the ends of the scale, and hence the new instrument would be insensitive to the extreme pedagogies. The results of the new instrument are summarised in Table 9, showing acceptable fit and good person and item separation.

Table 9: Measures and fit statistics for the items of the shortened version of the pedagogy instrument

Item description:	Measure	SE	Infit Mnsq	Outfit Mnsq
B3: Students use only the methods I teach them.	-0.67	0.14	0.9	0.9
B5: Students [don't] choose which questions they tackle.	-1.16	0.15	1.2	1
B7: Students [don't] compare different methods for doing questions.	-0.26	0.13	0.7	0.6
B11: I [don't] draw links between topics and move back and forth between topics.	1.35	0.11	1.2	1.2
B12: Students [don't] work collaboratively in small groups.	-0.01	0.13	0.8	0.8
B15: Students [don't] discuss their ideas.	1.19	0.11	0.8	0.8
B16: Students [don't] work collaboratively in pairs.	0.65	0.12	0.8	0.8
B17: Students [don't] invent their own methods.	-1.51	0.16	0.8	0.8
B19: I tell students which questions to tackle.	-0.64	0.14	1.3	1.3
B20: I encourage students to work more quickly.	1.06	0.11	1.4	1.4
B25: I teach each topic separately.	0	0.13	1.4	1.4
Person Summary Statistics: Separation 2.31 Reliability .84 **Item Summary Statistics**: Separation 6.92 Reliability .98 Category Statistics: OK				

The comparability of the new shorter version with the original measure of self-reported pedagogy was checked again, at both item and person level. Pearson correlation between the item estimates from the two different calibrations is almost 1 ($r=0.999$, $p<0.01$). The correlation of the person estimates by the two analyses is also very high ($r=0.892$, $p<0.01$).

8. Conclusion and discussion

We conclude, by way of substantive contribution to knowledge, that:

- It has been possible to measure the teacher-centrism of teaching practices through a teachers' self-report instrument with good psychometric performance. Future research might better collapse the rating scale to four or even three points and different versions might be preferred

(short and 'DIF-free') for different purposes, though they have little effect on person measures.

- Three levels were interpretable (based on an interpretation of the relevant clusters of items) but the middle level is essentially a mix of two extreme positions, and it is the extremes that are interesting in terms of effects, at least in modelling mathematics dispositions.
- There were no significant differences between practices self-reported on the two programmes, though there was some differential functioning suggesting that another dimension might be latent, and this would be consistent with what we observed (e.g. whole lessons on UoM devoted to coursework, where the teacher is a facilitator of the students' work).

Finally we can claim that pedagogy was found to have a significant effect on learner's dispositions, independent of programme, though most of the effect was due to a few connectionist classrooms.

The scale will need validating with other populations of teachers, but case study observations of practice and interviews of teachers support the validity of the scale's capacity to distinguish 'connectionist' and 'transmissionist' extremes. An interesting feature is the relatively low 'item difficulties' (i.e. practices described as less frequent) of many of the items that 'reform' teaching generally would regard as normative 'good' practice. It seems that this reform has not influenced teaching practices in this context very strongly, and this end of the scale is not doing much discriminating work in our sample. Yet, other researchers working with different, perhaps less exam-orientated contexts, may find these items and this part of the scale more significant. We suggest that the scales' psychometrics will be in need of recalibration for such new populations of teachers, of course.

Our modeling of student performance data suggest that one can build statistically significant regression models of student learning outcomes in which this measure of pedagogy is used as a process variable. As hypothesised, the model revealed a negative impact of transmissionist teaching practices on student disposition, though it is interesting that in fact the R-squared is quite low, and in fact most of the effect is in the extreme ends of the pedagogy scale. For this reason, and for practical reasons, we offered a smaller subscale which includes items at the extremes, cuts out DIF, but reduced discrimination in the centre of the scale: we suggest this might prove useful in similar studies in future.

Evidently there are many factors involved in students' dispositions to study mathematics, and so quite a significant difference in 'teacher centrism' is required to explain significant differences in students' dispositions. This is consistent with Swan's finding (Swan, 2006) with similar teachers that even when teachers report they have made significant shifts in pedagogy their students tend to rate this

change as relatively minor: they are used to quite significant differences in practices across their subjects of course, so the differences we report may seem major to us but not to the students.

Finally some methodological remarks. In our experience of using measurement we meet many common misconceptions about the nature of statistical analysis and 'reality': referees commonly tell us that a misfit of more than *x* (where x=1.2, 1.3, 1.4: take your pick) is 'unacceptable', or DIF should 'be eliminated', or 'models must fit the data if they are to be useful'. Our view is very different: statistics are a means to an end, and provide diagnostics in context that must always be judged alongside other evidence. Qualitative data and substantive expertise are the most helpful in this regard, and though this paper has shown our measurement methodology most prominently, we have taken care to reveal where judgment and qualitative data have been essential to the development. (For further details of qualitative case study analysis in this project see Williams et al., accepted).

We have therefore often had recourse to make decisions "on the balance of evidence/argument" and while we know this is not popular we insist that this is the only transparent way to report such work. In 'real' research projects such as this, one cannot await the perfect instrumentation and set of data: the bricoleur works with the best materials at hand.

Acknowledgment

The authors would like to acknowledge the support of the ESRC-TLRP programme of research into widening participation, funded by ESRC: RES-139-25-0241. We would also like to thank Malcolm Swan for providing his data to start this calibration work.

REFERENCES

Acton, S. F. (2003). What Is Good About Rasch Measurement? *Rasch Measurement Transactions, 16*(4), 902-903.

Andrich, D. (1999). Rating Scale Model. In G. N. Masters & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 110 - 121). Oxford: Pergamon.

Argyris, C., & Schon, D. (1978). *Organisational Learning: A theory of action perspective*. Reading, Mass: Addison Wesley.

Askew, M., Brown, M., Rhodes, V., Johnson, D., & Wiliam, D. (1997). *Effective Teachers of Numeracy (Final Report)*. London: King's College.

Bell, A. (1993a). Principles for the Design of Teaching. *Educational Studies in Mathematics, 24*(1), 5-34.

Bell, A. (1993b). Some experiments in Diagnostic Teaching. *Educational Studies in Mathematics, 24*, 115-137.

Bohlig, M., Fisher, W. P. J., Masters, G. N., & Bond, T. (1998). Content Validity and Misfitting Items. *Rasch Measurement Transactions, 12*(1), 607.

Bond, T., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ: Lawrence Erlbaum.

Bowles, R. (2003). Rejecting Best Items? *Rasch Measurement Transactions, 17*(1), 917.

Cuban, L. (1983). How did teachers teach, 1890–1980. *Theory Into Practice, 22*(3), 160-165.

Davis, P., & Williams, J. S. (in press). Hybridity of maths and peer talk: crazy maths In H. Mendick & e. al. (Eds.), *Mathematical relationships in Education: Identities and Participation*. London: Routledge.

Ernest, P. (1991). *The philosophy of mathematics education*. Basingstoke: Falmer.

Green, K. E., & Frantom, C. G. (2002). *Survey Development and Validation with the Rasch Model*. Paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing.

Gustafsson, J.-E. (1980). Testing and Obtaining Fit of Data to the Rasch Model. *British Journal of Mathematical and Statistical Psychology, 33*, 205 - 233.

Harwood, W. S., Hansen, J., & Lotter, C. (2006). Measuring teacher beliefs about inquiry: The development of a blended qualitative/quantitative instrument. *Journal of Science Education and Technology, 15*(1), 69-79.

Hiebert, J., Gallimore, R., Garnier, H., Bogard Givvin, K., Hollingsworth, H., Jacobs, J., et al. (2003). *Teaching Mathematics in Seven Countries: Results from the TIMSS 1999 Video Study*. US Department of Education: NCES: National Center for Educational Statistics

Hoffman, J. P. (2004). *Generalized Linear Models: An Applied Approach*: Pearson Education.

Horizon Research, I. (1996). *Local Systemic Change teacher questionnaire*. Chapel Hill, NC: Horizon Research Inc.

Hutchenson, G., & Sofroniou, N. (1999). *The Multivariate Social Scientist. Introductory Statistics Using Generalized Linear Models*. London: Sage.

Kember, D., & Gow, L. (1994). Orientations to teaching and their effect on the quality of student learning. . *Journal of Higher Education 65*(1), 58-74.

Linacre, J. M. (2002). Optimizing Rating Scale Category Effectiveness *Journal of Applied Measurement 3*(1), 85-106.

Linacre, J. M. (2003). A user's guide to FACETS: Rasch-Model Computer programs [software manual]. Chigago: Winsteps.com.

Lopez, W. A. (1995). Rating scales and shared meaning. *Rasch Measurement Transactions, 9*(2), 434.

Lopez, W. A. (1996). Communication Validity and Rating Scales. *Rasch Measurement Transactions, 10* (1), 482-483.

McCaffrey, D. F., Hamilton, L. S., Stecher, B. M., Klein, S. P., Bugliari, D., & Robyn, A. (2001). Interactions among Instructional Practices, Curriculum, and Student Achievement: The Case

of Standards-Based High School Mathematics. *Journal for Research in Mathematics Education, 32*(5), 493-517.

Messick, S. (1988). The Once and Future Issues of Validity: Assessing the Meaning and Consequences of Measurement. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 33-45). London: Lawrence Erlbaum Associates, Publishers.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (Third ed., pp. 13-103). USA: American Council of Education and the Oryx Press.

Mullis, I., Martin, M., Gonzales, E., Gregory, K., Garden, R., O'Connor, et al. (2000). *TIMSS 1999 - International Mathematics Report (Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade)*. Boston: The International Study Center & The International Association for the Evaluation of Educational Achievement

NCES. (2000). *Mathematics and Science in the Eighth Grade: Findings from the Third International Mathematics and Science Study*: US Department of Education Office of Educational Research and Improvement

OECD. (2003). *The PISA 2003 Assessment Framework - Mathematics, Reading, Science and Problem Solving Knowledge and Skills*

Pampaka, M., Williams, J., Hutcheson, G., Black, L., Davis, P., Hernandez-Martinez, P., et al. (2007). *Measuring the 'effectiveness' of Programme and pedagogy on maths disposition and self efficacy measures*. Paper presented at the Paper presentation at the Annual Conference of the British Educational Research Association (BERA 2007).

Roelofs, E., Visser, J., & Terwel, J. (2003). Preferences for various learning environments: Teachers' and parents' perceptions. *LEarning Environments Research, 6*(1), 77-110.

Ryan, J. T., & Williams, J. S. (2007). *Children's mathematics 4 - 15*. Milton Keynes: Open University Press.

Schon, D. (1990). *Educating the Reflective Practitioner*. Oxford: Jossey-Bass Publishers.

Schuh, K. L. (2004). Learner-centered principles in teacher-centered practices? *Teaching and Teacher Education, 20*(8), 833-846.

Swan, M. (2000). The Purpose of Mathematical activities and Pupils' Perceptions of them. *Research in Education, 63*, 199-223.

Swan, M. (2006). Designing and using research instruments to describe teh beliefs and practices of mathematics teachers. *Research in Education*(75), 58-70.

Swanson, C. B., & Stevenson, D. L. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP State Assessments. *Educational Evaluation and Policy Analysis, 24*(1), 1-27.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of Differential Item Functioning Using the Parameters of Item Response Models. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 67-114). London Lawrence Erlbaum Associates, Publishers.

Wake, G., & Pampaka, M. (2007). *Measuring Perceieved Self-Efficacy in Applying Mathematics.* Paper presented at the Proceedings of the Fifth Congress of the European Society for Research in Mathematics Education (CERME) (WK 13, pp. 2210-2219), Larnaca, Cyprus.

Wake, G., & Pampaka, M. (in press). The central role of the teacher - even in studen centred pedagogies. *Proceedings of the Joint Meeting of the 32nd Conference of the International Group for the Psychology of Mathematics Education and the XX North American Chapter. Morelia, Michoacán, México.*

Webster, B. J., & Fisher, D. L. (2003). School-level environment and student outcomes in mathematics. *LEarning Environments Research, 6*(3), 309-326.

Williams, J., Pampaka, M., Black, L., Davis, P., Hernandez-Martinez, P., & Wake, G. (2007). *Development and validation of two 'soft' outcome measures: Disposition to enter HE and disposition to study mathematically demanding subjects in HE.* Paper presented at the Paper presentation at the Annual Conference of the British Educational Research Association (BERA 2007)

Williams, J. S., Black, L., Davis, P., Hernandez-Martines, P., Hutchenson, G., Nicholson, S., et al. (2008). *TLRP Research Briefing No 38: Keeping open the door to mathematically demanding*

programmes in Further and Higher Education. School of Education, University of Manchester.

Williams, J. S., Black, L., Hernandez-Martinez, P., Davis, P., Pampaka, M., & Wake, G. (accepted). Repertoires of aspiration, narratives of identity, and cultural models of mathematics in practice. In M. César & K. Kumpulainen (Eds.), *Social Interactions in Multicultural Settings* Rotterdam: Sense. Publishers.

Williams, J. S., Black, L., Hernandez-Martinez, P., Davis, P., Pampaka, M., & Wake, G. (in press). Repertoires of aspiration, narratives of identity, and cultural models of mathematics in practice. In M. César & K. Kumpulainen (Eds.), *Social Interactions in Multicultural Settings* Rotterdam: Sense. Publishers.

Wolfe, E. W., & Smith Jr., E. V. (2007a). Instrument Development Tools and Activities for Measure Validation Using Rasch Models: Part I - Instrument Development Tools. *Journal of Applied Measurement, 8*(1), 97-123.

Wolfe, E. W., & Smith Jr., E. V. (2007b). Instrument Development Tools and Activities for Measure Validation Using Rasch Models: Part II - Validation Activities. *Journal of Applied Measurement, 8*(2), 204-234.

Wright, B. D. (1994). Data Analysis and Fit. *Rasch Measurement Transactions, 7*(4), 324.

Wright, B. D. (1999). Rasch Measurement Models. In G. N. Masters & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 268-281). Oxford: Pergamon.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

Wright, B. D., & Mok, M. (2000). Rasch Models Overview. *Journal of Applied Measurement, 1*(1), 83-106.

Zhu, W. (2002). A Confirmatory Study of Rasch-Based Optimal Categorization of a Rating Scale. *Journal of Applied Measurement, 3*(1), 1-15.

Zhu, W., Updyke, W. F., & Lewandowski, C. (1997). Post-hoc Rasch analysis of optimal categorization of an ordered-response scale. *Journal of Outcome Measurement,, 1*, 286-304.