SIÔN WYN EVANS

DIFFERENTIAL PERFORMANCE OF ITEMS IN MATHEMATICS ASSESSMENT MATERIALS FOR 7-YEAR-OLD PUPILS IN ENGLISH-MEDIUM AND WELSH-MEDIUM VERSIONS

ABSTRACT. This paper draws on data from the development of annual national mathematics assessment materials for 7-year-old pupils in Wales for use during the period 2000–2002. The materials were developed in both English and Welsh and were designed to be matched. The paper reports on item analyses which sought items that exhibited differential performance in relation to whether the materials were English medium or Welsh medium. The items that exhibited consistent differential item functioning in relation to language during pre-testing are reviewed in order to discuss the linguistic factors that could affect such behaviour.

KEY WORDS: assessment, differential item functioning, mathematics, English, language, Welsh

1. INTRODUCTION

Mathematics education can involve both learning and assessment, and both of these activities take place within language or, in the words of Durkin (1991), it "begins in language, it advances and stumbles because of language, and its outcomes are often assessed in language". Although language is always a factor in mathematics education, it has a particular significance in situations where there is more than one language involved, such as when assessing mathematical attainment by using tests that exist in more than one language as in the case of countries where two or more linguistic communities coexist or in the case of international surveys such as the Trends in International Mathematics and Science Study (TIMSS; see, for example, Mullis et al., 2004).

Between 1991 and 2001 there were statutory assessments in mathematics for 7-year-old pupils in Wales (i.e. those at the end-of-key-stage 1). The assessment materials were produced in both English and Welsh. Key stage 1 statutory assessments in Wales came to an end in 2001, although the 2002 materials had been developed and were sent to schools as optional materials. The materials for 2000–2002 were developed by the National Foundation for Educational Research (NFER) for Awdurdod Cymwysterau, Cwricwlwm ac Asesu Cymru (ACCAC, or in English 'Qualifications, Curriculum and Assessment Authority for Wales'). Each set of

Educational Studies in Mathematics (2006) 64: 145–168
DOI: 10.1007/s10649-006-1634-5

final assessment materials were the result of a two year development cycle that included two pre-tests.

During each pre-test there was an investigation of differential item functioning (DIF) in relation to the English-medium and Welsh-medium versions of the materials (referred to as 'language DIF' in this paper). The DIF analyses sought any differences in performance between the two groups of pupils on particular assessment items, taking the total scores of the pupils in the two groups into account. It was thus possible for a DIF analysis to indicate that one group was favoured even though the facility was less for that group. This paper reviews the items that exhibited language DIF during the development process in the period 1998–2001 when NFER was developing the key stage 1 assessment materials. Even though the period represents the development of only 3 years' assessment materials, it was the only time when the materials were developed specifically for Wales, unlike those produced up to 1999 that were for both Wales and England. The review of the items will consider the linguistic factors that could have contributed to their differential performance.

In the following section, a brief review of the relevant literature is given before moving on to the next section where there is information about the assessment materials and the samples used in pre-testing. The methodology of the present work will be indicated and then some of the items that exhibited language DIF will be discussed in relation to the possible linguistic factors that could have had an effect on such DIF. Finally, there will be a discussion of some of the issues that arise.

2. Language and mathematics assessment

2.1. *Language and mathematics*

A useful concept with which to discuss the issue of the use of language in mathematics education is that of a mathematics register (Pimm, 1987). In general, there will be a mathematics register related to any natural language used in learning or assessing mathematics, although the situation can be more complicated when the natural language of social interaction in a classroom is different from that of the teaching materials (Setati, 2004). Such registers "have to do with the social usage of particular words and expressions, ways of talking but also ways of meaning" (Pimm, 1987, p. 108). The natural language can be used orally or in its written form, or in combination as in a teaching situation. The learning of a mathematics register will involve learning "not just the use of technical terms, but also certain phrases and even characteristic modes of arguing that constitute a

register" (Pimm, 1987, p. 76). However, for young children there is a need to ensure that they have some of the basic building blocks of that register, such as the names of numbers and essential vocabulary. In England this is reflected in the publication of lists of such vocabulary for different ages of pupils (DfEE, 1999).

The issue of language in the mathematical materials seen by pupils was discussed by Shuard and Rothery (1984) where they made a distinction between 'mathematical English' and 'ordinary English'. They saw three types of relationships between mathematical and ordinary language as being relevant to issues of difficulty in understanding: words that have the same meaning in both mathematical and ordinary language, words that have a meaning in the mathematical language only and words that have different meanings in mathematical and ordinary usage. These difficulties can be linked with lexical word ambiguities within mathematical language, such as cases of homonymy, polysemy or homophony (Durkin and Shire, 1991).

2.2. *Differential item functioning*

Difficulty in interpreting the meaning of written text can be especially important in an assessment situation and is one aspect of a range of 'sources of difficulty' that can be associated with an assessment item (Pollitt and Ahmed, 1999). Since written text has the potential of being more of a 'source of difficulty' in one language as compared to another, it is important to minimize such effects in assessment instruments that are administered in different languages. This can involve the use of both statistical and judgemental techniques (Hambleton, 1993). One statistical technique is the use of DIF analyses, but the existence of DIF might not imply that there is bias in an item (see, for example, Camilli and Shepard, 1994). The unreliability of DIF analyses leads one to ask whether or not the findings of such analyses are consistent and interpretable. Hence, there is a need for a judgemental analysis to determine whether or not the DIF is a sign of construct-relevant item difficulty.

The occurrence of DIF across tests in different languages can be related to factors relating to the relevant language groups or to factors within those particular items. Recently, Emenogu and Childs (2005) considered the possible impact of curricular differences in addition to that of language on the differential performance of 13- and 16-year-old English-medium and French-medium pupils on measurement and geometry items in a mathematics test in Canada. They commented that their results were not conclusive and illustrated the complexity of the factors that contribute to performance on assessment items. There are many other possible factors that can be considered. The TIMSS international surveys consider five broad areas

that impact on pupil performance: curriculum, schools, teachers, classroom activities and pupils, so background questionnaires are used to collect information on such issues as the pupils' attitudes towards mathematics and the teachers' instructional practices (Chrostowski, 2004).

Based upon a study of verbal items from a psychometric test in Israel that was available in Hebrew and in Russian, the within-item sources of DIF across the languages were identified by Allalouf et al. (1999) as differences between the two language versions of an item in the following four categories: word difficulty, content, format and cultural relevance. The 'word difficulty' category related to cases where the two language versions were accurate translations of each other, but a word or expression was easier in one language. Differences in 'content' related to differences in meaning between the two language versions: such differences could be due to the use of a word that had more than one meaning in one of the languages and only one meaning in the other. Examples of differences in 'format' included a large difference between sentence lengths in the two language versions or a different order of clauses in a complex sentence or changes due to differences between grammatical constructions so that, for example, there is a change in the subject of a sentence. The 'cultural relevance' category referred to cases where the content was more familiar to one of the two linguistic groups.

Gierl and Khaliq (2000) made a study of DIF in mathematics and social studies achievement tests in Canada that were available in English and in French and came up with the following four sources of DIF: omissions or additions that affect meaning, differences in words or expressions inherent to language or culture, differences in words or expressions not inherent to language or culture and format differences. The authors compared their categorization with that of Allalouf et al. (1999) They considered 'omissions or additions that affect meaning' as being similar to difference in 'content'. The 'differences in words or expressions inherent to language or culture', they considered as corresponding to both differences in 'word difficulty' and 'cultural relevance'. The 'differences in words or expressions not inherent to language or culture', they considered as corresponding to 'content' differences. Their 'format' differences also included differences in punctuation and capitalization in addition to item structural differences such as the repetition of a word in both stem and elsewhere within the item in one language but not in the other.

2.3. *Mathematics assessment materials in Welsh and English*

In the schools of Wales, both Welsh and English are used in mathematics education. Some of the issues relating to the production of matched test

instruments in both Welsh and English were discussed by Wiliam (1994). He suggested that there are three Welsh linguistic factors that are relevant: literary and spoken Welsh differences, grammar and the technical lexicon. The requirements for tests are that pupils who are being taught through the medium of Welsh should find the language accessible and that the grammar is correct. The technical terminology has to be correct and in keeping with what pupils are familiar with at that stage in their school career. When discussing statutory tests for mathematics in both English and Welsh, Jones (1998) gave a slightly different list of linguistic issues: familiarity (of everyday words, of dialect words and of technical words) and grammar. The relationship between the factors mentioned by the Welsh authors and those of Allalouf et al. (1999) and Gierl and Khaliq (2000) will be discussed in Section 4.3.

2.4. *Mathematical vocabulary and pupil performance*

The suggestion is sometimes made that the counting systems of languages such as Chinese, Japanese and Korean is a factor in the high performance of speakers of those languages in arithmetic. For example, Park (2004) lists the simple pronunciation of numbers and the regularity of the number system amongst the possible factors contributing to the high achievement of Koreans in international comparisons of mathematical achievement such as TIMSS. Here 'regularity' refers to the existence of a rule for combining the number names of numbers from 1 to 10 to form the names of those from 11 onwards. The Korean number names are also consistent with the base ten written forms of the numbers.

The Welsh language has a regular base ten system of counting that is used in schools. This base ten system was developed from the previous mixed base ten and base twenty system around the beginning of the nineteenth century, and when teaching through the medium of Welsh began in the 1950s it was the base ten system that was used for arithmetic, whilst keeping the older system for such uses as ages and dates (Roberts, 2000). After 'ten', the Welsh names that are used for calculation purposes are 'un deg un' (one ten one), 'un deg dau' (one ten two) etc.

Dowker and Lloyd (2001) reported a difference in performance between Welsh-medium and English-medium pupils in relation to the names for numbers in both languages. In their study, 6- and 8-year-old pupils who were educated through the medium of Welsh performed better than those educated through the medium of English in reading and comparing two-digit numbers. The pupils were given a number comparison task where they were shown pairs of two-digit numbers and asked to read the numbers aloud before pointing to the biggest. The authors suggested that the difference

in performance was related to the difference in the degree of regularity between the Welsh and English number systems.

As to other differences between the mathematical vocabularies of different languages, Han and Ginsburg (2001) conducted two studies involving some words from the English and Chinese mathematical vocabularies. For their second study, they reported a strong correlation between Chinese junior high school pupils' performance on test items with mathematics words that, in their first study, had been rated as being clear in their meaning. They suggested that the relative clarity of mathematical terms in the Chinese language may have contributed to the Chinese-speaking pupils' understanding of mathematics and to superior mathematics performance. For example, the Chinese word for 'quadrilateral' means 'four-side-shape'.

Jones (1993) reported on the differential performance across English and Welsh versions of a mathematics test for 16-year-old pupils on an item assessing similar triangles where the Welsh-medium pupils performed significantly better. In terms of the differences between the English and Welsh mathematical vocabularies and their relationships to ordinary language, the English term 'similar' has a specific meaning in mathematical English and a more general one in everyday English, whereas the Welsh term 'cyflun' is used only in mathematical Welsh, being a purely technical term, and could have been cueing the pupils as to what was required more than the term 'similar' in English. However, there could have been other reasons as well, such as curricular ones.

2.5. *Concluding remarks*

Based on this brief review of some of the relevant literature the following two points can be noted:

- Difference in performance on mathematics assessment items between language groups when those items are in different languages can be related to either differences between the items or differences between the groups
- Some linguistic factors related to the mathematical vocabularies of different languages have been suggested as possible reasons for differences in performance.

As noted by Setati (2004) in relation to the development of mathematics registers for the official African languages in South Africa "research into the use of these registers and their effect on the mathematics, mathematics education and on the languages is crucial". Any discussion of the relationships between registers and mathematics performance needs to be evidence-based. Thus, in relation to the assessments addressed in this paper

one can ask the following two questions.

- What linguistic factors within the items could have played a role in the observed language DIF?
- Did the linguistic factors include differences between mathematics registers and, if so, how plausible is the idea that such differences could have had an effect on the differences in pupil performance?

3. Assessment materials, pre-testing and samples

3.1. *Assessment materials*

Key stage 1 pupils were statutorily assessed by means of either a paper-based test, that was taken by most of the pupils, or a teacher-administered task, that was used for pupils of lower ability in groups of between one and four. The paper-based test consisted of both oral questions, where the teacher read out the questions, and written questions, where all the information was given on the test paper. For the oral questions, there could be some stimulus material such as diagrams in the pupil answer booklet, but there were no written instructions. Thus, there were three types of items: oral items, written items and task items.

The items were developed in parallel in English and Welsh by a process of interactive development (Ruddock and Evans, 2000) where the demands of each language affected the wording of an item in the other language so that the versions in both languages were as close as possible whilst still having an acceptable wording in either language.

3.2. *Pre-testing*

The pre-testing was done by teachers who had the materials sent to them with instructions about how to conduct the tests and the tasks. They also received questionnaires to give feedback on the materials. Neither tests nor tasks were timed assessments and it was left to the teachers to give the pupils as much time as they saw fit.

Pre-test 1 for each cycle was for the purpose of item selection and used pupils from both Years 2 and 3, as the Year 2 pupils were not yet be familiar with all the material being tested at the time of pre-testing. Pre-test 2 for each cycle was for confirming the final form of the materials and setting cut scores and used only Year 2 pupils, although minor changes could still be made to the second pre-test versions of items for the final assessment materials used by schools. Details about the samples of pupils used and other information about the pre-tests, such as the comments from teachers

in the questionnaires, were presented to the body responsible for assessment in schools (ACCAC) in a series of twelve reports between 1999 and 2001, ending with those for the second pre-test for the 2002 materials (NFER, 2001a,b).

Various DIF analyses (see, for example, Camilli and Shepard, 1994) were performed during both pre-tests in each cycle. Whenever the size of the samples allowed, these consisted of DIF by gender, DIF by language and DIF by the language background of the pupils. Analyses where there were less than 50 in any of the groups need to be treated with caution and in most cases were not performed. Thus, no DIF by language background within the English-medium samples were performed.

For the 2000 development cycle, Mantel–Haenszel analyses were used to investigate DIF, but for the 2001 and 2002 cycles a logistic regression approach was used. As part of the development process, all the items that exhibited DIF at a significance level of at least 5% were listed and each one considered to see if there was an issue that could be addressed.

3.3. *Samples*

Details about the samples are summarized in Tables I and II. Efforts were always made to ensure that the total sample was representative as regards the types, geographical locations and the sizes of the schools. There was, for both paper-based test and task in each pre-test, a sample of English-medium pupils and a sample of Welsh-medium pupils, with any particular school supplying one or the other.

For the purposes of analysing the data the Welsh-medium and English-medium samples were each further divided into two sub-samples on the basis of language background. Language background classification in Wales is not a simple matter since there is a continuum of usage of the two languages (Baker, 1984). However, dividing the Welsh-medium pupils into those who, according to their teachers, spoke Welsh at home (W1) and those who did not (W2) allowed comparison, to some extent, of pupils for whom school could well be the only place where they used Welsh with those who were using it in other contexts. There was a similar division of the English-medium samples into those who were categorized, by their teachers, as being first language speakers (E1) and those for whom English was an additional language (E2). A large proportion of the pupils in the Welsh-medium samples came from homes where the language was English. Thus the Welsh-medium samples had a high proportion of pupils who were bilingual whereas the English-medium samples were mostly pupils who were monoglot.

TABLE I
Samples used in pre-testing paper-based tests

	Pupils Pre-Test 1									Pupils Pre-Test 2								
				Welsh medium			English medium						Welsh medium			English medium		
Year	Schools	Boys	Girls	Total	W1	W2	Total	E1	E2	Schools	Boys	Girls	Total	W1	W2	Total	E1	E2
2000	35	310	333	254	n/a[a]	n/a[a]	389	358[b]	24[b]	38	416	443	304	103	174	555	531	24
2001	37	306	339	250	90	160	395	446[b]	25[b]	48	365	360	289	131	158	481	449	32
2002	54	423	404	370	187	181	458	446	12	47	298	354	157	79	78	496	462	34

[a]This information was not available.
[b]These figures were based on both the English-medium samples used for item analysis and additional samples drawn in order to obtain more E2 pupils.

TABLE II
Samples used in pre-testing tasks

	Pupils Pre-Test 1									Pupils Pre-Test 2								
				Welsh medium			English medium						Welsh medium			English medium		
Year	Schools	Boys	Girls	Total	W1	W2	Total	E1	E2	Schools	Boys	Girls	Total	W1	W2	Total	E1	E2
2000	89	191	143	82	32	50	252	251	1	85	181	132	43	9	34	270	261	8
2001	66	125	90	30	11	19	185	167	18	92	144	103	48	16	32	199	188	11
2002	55	96	78	54	27	27	120	113	4	59	95	58	47	14	33	106	104	2

4. METHODOLOGY

4.1. *Selection of items for discussion*

DIF analyses are generally used during test development for such purposes as weeding out problematic items, but they can also be used to investigate test performance and responses to items (Shimizu and Zumbo, 2005). In this paper, DIF analyses that were used in the test development process are re-visited in order to collect together possible within-item sources for the DIF across the two language versions.

The unreliability of DIF analyses needs to borne in mind since one is dealing with difference measures and the behaviour of a single item rather than a more stable collection of many items (Camilli and Shepard, 1994). Thus, the pattern of language DIF of only those items that were pre-tested twice is considered, i.e. those items that appeared in a first pre-test and, often in a revised form, in a second pre-test.

4.2. *Some aspects of Welsh grammar*

Before discussing linguistic issues in Welsh-medium items, it might be useful to indicate some of the grammatical issues that can be important in the case of a Welsh-medium item. One issue is that of 'mutations' where the initial consonants of some words are replaced by other consonants. Such changes are dependent upon the grammatical role of a word and on the previous word, so that 'ci' (dog) can be 'gi' as in 'ei gi' (his dog), 'chi' as in 'ei chi' (her dog), or 'nghi' as in 'fy nghi' (my dog).

Another issue is the fact that all nouns are either masculine or feminine and many adjectives will change according to the noun referred to. This also happens with the numbers 'two', 'three' and 'four'. Thus 'two dogs' is 'dau gi' ('ci' is masculine) and 'two cats' is 'dwy gath' ('cath' is feminine). Both 'ci' and 'cath' have mutated following the number 'two'.

4.3. *Linguistic analysis of items*

The factors discussed by the Welsh authors can be related to the work of Allalouf et al. (1999) and Gierl and Khaliq (2000). Wiliam's (1994) reference to differences between literary and spoken Welsh is related to that made by Jones (1998) to dialect. The form of Welsh used in tests tends to be more literary, and hence more neutral, in relation to variations in dialect. Both the issue of dialects and that of the familiarity of everyday words (Jones, 1998) correspond to the 'word difficulty' category of Allalouf et al. (1999). Technical words, that both Wiliam and Jones referred to as

a separate category, can also fit in under the 'word difficulty' heading. However, words (whether everyday or technical, literary or dialect) can also contribute to difference in 'content' when more than one meaning is present. The grammatical differences between Welsh and English can lead to differences in 'format', to differences in 'word difficulty' or differences in 'content'. For example, the mutated form 'nghi' can be considered more difficult than 'ci' and the mutated form 'chi' also means 'you' (second person plural).

The items exhibiting consistent language of DIF will be considered in relation to two sets of factors: differences between the Welsh- and English-medium items and possible differences between the Welsh- and English-medium pupils. The linguistic analysis will be in terms of categories that are an adaptation of the within-item sources of DIF used by Allalouf et al. (1999) and Gierl and Khaliq (2000) taking into account the work of Wiliam (1994) and Jones (1998).

The 'cultural relevance' category is not used since the items were based on the national curriculum common to both language groups and the contexts of the items, when there were any, were based on what was relevant for the age group. A new category, 'mathematical vocabulary' has been added thus changing 'word difficulty' to 'everyday word difficulty'. The use of the 'mathematical vocabulary' category recognises the different role to other words and expressions that such vocabulary can play within an item in that it has the potential of being an intrinsic part of the mathematics being assessed. Thus we have the following four categories: *format*, *content*, *everyday words or expressions* and *mathematical vocabulary*. These categories can also be adapted for use with task items in that the references to words and expressions are now to spoken rather than written language. The issue of format can be considered in relation to any printed material that the pupils see.

5. Items exhibiting language DIF

5.1. *The items*

There were, in total, 15 oral items, 90 written items and 71 task items that were pre-tested twice. Of these, seven paper-based items, all of which were written items, and two task items exhibited language DIF consistently in two pre-tests. Details about DIF in relation to these are summarized in Table III where P1–P7 are the written items and T1 and T2 the task items. There were only a few cases where these items exhibited differences in performance according to the language background of the Welsh-medium

TABLE III
Items exhibiting consistent language DIF

		Pre-test 1				Pre-test 2			
		Language		Background		Language		Background	
Year	Item	Favoured	Level of significance	Favoured	Level of significance	Favoured	Level of significance	Favoured	Level of significance
2000	P1	English	0.1%	n/a	n/a	English	0.1%	no DIF	n/a
2000	P2	English	1%	n/a	n/a	English	0.1%	no DIF	n/a
2000	P3	Welsh	0.1%	n/a	n/a	Welsh	5%	W1	0.1%
2001	P4	Welsh	5%	no DIF	n/a	Welsh	5%	no DIF	n/a
2001	P5	English	0.1%	W1	5%	English	0.1%	no DIF	n/a
2002	P6	Welsh	0.1%	W1	5%	Welsh	1%	no DIF	n/a
2002	P7	English	0.1%	no DIF	n/a	English	0.1%	W2	1%
2000	T1	English	5%	n/a	n/a	English	5%	n/a	n/a
2002	T2	Welsh	5%	n/a	n/a	Welsh	5%	n/a	n/a

pupils and there was no consistent pattern. Table IV lists the texts or scripts of these items in their second pre-test forms.

5.2. *Factors outside the items*

The possibility that factors related to differences between the two linguistic groups played a role in the observed DIF exists for all the nine items. For example, the fact that many of the Welsh-medium pupils were learning mathematics through their second language could have had an effect on knowing the number names up to ten in T1. There could have been a difference between English- and Welsh-medium classrooms in relation to the familiarity of pupils with particular topics within the curriculum such as tally charts in P3 or polygons in P5. There is also the possibility of differences in emphasis given to particular types of question such as word problems as in P1/2. There could also be differences in the acquaintance of pupils with particular ways of presenting questions such as in the case of P4 where there were comments in the teacher questionnaires about the fact that the item had not told the pupils where to draw the line by giving them the starting point on the paper.

As far as the intended curriculum is concerned, this is the same for both English- and Welsh-medium schools (ACCAC, 2000). The statutory assessments can be assumed to have played some role in ensuring that the implemented curriculum was close to the intended one since the assessments reflected the content of the national curriculum. However, there could still have been differences between the curriculum exposure in English- and Welsh-medium classrooms. Such differences could be related to differences in the textbooks and other teaching materials that were being used.

As regards pedagogy, there could be more emphasis on learning terminology within Welsh-medium classrooms. The emphasis in the list of standardized terminology published by the body responsible for curriculum and assessment in schools (ACCAC) is on Welsh-medium technical terms and not English ones (ACCAC, 1998). A similar a state of affairs was mentioned by Emenogu and Childs (2005) as existing in Canada in that French-language schools tend to place more emphasis on terminology. However, in the case of the nine items discussed here, although Welsh-medium pupils performed better on the ‘right angle’ item (P6) English-medium pupils performed better on the ‘pentagon’ item (P5).

5.3. *Factors within the items*

The results of the linguistic analyses of the individual items are summarized in Table V. Explanations as to how decisions were made as to which

TABLE IV
Texts or scripts of items exhibiting consistent language DIF

Item	English	Welsh
P1/2	5 Pencils fill 1 box.	Mae 5 pensel yn llenwi 1 bocs.
	Jill has 23 pencils.	Mae 23 pensel gan Jill.
	How many boxes can she fill?	Sawl bocs fydd Jill yn medru ei lenwi?
	How many pencils will be left over?	Sawl pensel fydd ar ôl?
P3	This chart shows how children come to school.	Mae'r siart yn dangos sut mae plant yn teithio i'r ysgol.
	Tally chart	siart cyfrif
	car bus taxi walk	car bws tacsi cerdded
	Use the chart to complete the graph.	Defnyddiwch y siart i gwblhau'r graff.
P4	Use a ruler to draw a line 9cm long.	Defnyddiwch bren mesur i dynnu llinell 9cm o hyd.
P5	Put a tick (√) inside all three pentagons.	Ticiwch (√) y tu mewn i'r tri siâp pentagon.
P6	Find the shape with two right angles.	Mae dwy ongl sgwâr gan un o'r siapiau hyn.
	Put a tick (√) in this shape.	Ticiwch (√) y siâp hwnnw.
P7	These are the temperatures for four days.	Dyma'r tymheredd ar gyfer pedwar diwrnod.
	Monday Tuesday Wednesday Thursday	Llun Mawrth Mercher Iau
	On which day was the temperature highest?	Ar ba ddiwrnod oedd y tymheredd uchaf?
T1	I am going to tell each of you a number.	Yr wyf am roi rhif i bob un ohonoch chi.
	You should write it on one of your cards.	Dylech chi ei sgrifennu ar un o'ch cardiau.
	Your number is [1^1, 5^2, 3^3, 7^4]	Eich rhif ydi [1^1, 5^2, 3^3, 7^4]
	Now I am going to tell you another number.	Yr wyf yn mynd i roi rhif arall i chi.
	You should write it on one of your other cards.	Dylech chi ei sgrifennu ar un arall o'ch cardiau.
	Your next number is [10, 5, 6, 9]^1, [7, 4, 9, 1]^2, [8, 10, 2, 6]^3, [2, 8, 3, 4]^4	Eich rhif nesaf ydi [10, 5, 6, 9]^1, [7, 4, 9, 1]^2, [8, 10, 2, 6]^3, [2, 8, 3, 4]^4
T2	There are [18 pears^1, 16 apples^2, 17 bananas^3, 19 oranges^4] in your basket. I want you to write the number [18^1, 16^2, 17^3, 19^4] here.	Mae [18 gellygen^1, 16 afal^2, 17 banana^3, 19 oren^4] yn eich basged. Mae arnaf eisiau i chi sgrifennu'r rhif [18^1, 16^2, 17^3, 19^4] yn y fan hyn.

Note. 1: First pupil's number; 2: second pupil's number; 3: third pupil's number; 4: fourth pupil's number.

TABLE V
Linguistic factors present in the items

Item	Format	Content	Everyday words or expressions	Mathematical vocabulary
P1/2	Sentences 1 and 2: Welsh has 'pensel' (pencil) near the beginning [W]		Welsh has two forms for 'pencil' depending on dialect: 'pensel' and 'pensil'. [E]	
	Sentence 3: Welsh uses Jill and English uses a pronoun [E/W]		Welsh has two ways of asking 'how many': 'sawl' and 'faint'. [E]	
P3	Capitalization of chart titles: Tally chart/siart cyfrif [E/W]	Come/teithio (travel) [E]	defnyddiwch/use [E]	Chart/siart [E/W]
		This chart/'r siart (the chart) [E/W]	i gwblhau/to complete (The unmutated form is 'cwblhau'.) [E]	Tally chart/siart cyfrif (counting chart) [W]
				Graph/graff [E/W]
P4		'O hyd' in Welsh means both 'of length' and 'always'. [E]	defnyddiwch/use [E]	Ruler/pren mesur (measuring wood)
			'Bren' is the mutated form of 'pren' (wood). [E]	'Ruler' has more than one meaning in English. [W]

(*Continued on next page*)

TABLE V
(*Continued*)

Item	Format	Content	Everyday words or expressions	Mathematical vocabulary
P5		English contains 'all' and Welsh does not. [E]		Pentagon [E/W]
		Welsh contains 'siâp' (shape) and English does not [E/W]		
P6		Find the shape with two right angles./Mae dwy ongl sgwâr gan un o'r siapiau hyn. (There are two right angles with one of these shapes.) [E]		Right angle/ongl sgwâr (square angle) [W]
		Put a tick in/Ticiwch (Tick) [E/W]		'dwy' is the feminine form of 'dau' (two) [E]
P7	Labelling of thermometers: Welsh has omitted 'dydd' (day) from 'dydd Llun' (Monday) etc. [E]		Temperature highest/tymheredd uchaf 'Tymheredd uchaf' can also be read as 'highest temperature'. [W]	Temperature/tymheredd In everyday speech 'gwres' (technically 'heat') is used for 'temperature'. [E]
T1		One of your other cards/un arall o'ch cardiau (another one of your cards) [E/W]		Number/rhif [E/W]
				Names of numbers [E]
T2			Names of fruits [E]	number/rhif [E/W]
				Names of numbers [W]

Note. E: possibly favouring English-medium pupils; W: possibly favouring Welsh-medium pupils.

language group could be favoured are given below in the discussions on each factor, together with examples from some of the items. Two items are analysed in greater detail since these allow a discussion of the possible effects of mathematics registers on performance.

5.3.1. *Format*
In relation to the order or repetition of words, the language that exhibited greater internal consistency within an item was considered to be favoured. In the Welsh version of P1/2 the noun 'pensel' (pencil) came before the other nouns 'bocs' (box) and 'Jill' in both sentences of the introduction. However, it was with the English version that there was better performance. The Welsh version of P1 included the repetition of the proper name, Jill, rather than using the relevant pronoun. Since all nouns in Welsh are either masculine or feminine, the use of a pronoun has the potential of being ambiguous. In this case, neither English nor Welsh version was considered favoured.

5.3.2. *Content*
As regards a word or expression with more than one meaning, the assumption was made that the language with only one meaning was favoured. However, in P4 the use of 'o hyd' did not seem to have affected the performance of the Welsh-medium pupils, probably because the context did not trigger any other meaning and they were familiar with the wording used.

Decisions about which language is favoured in cases where there is an extra word or a different word or expression were based on which language version could have helped pupils get the correct answer. There was a reference to 'all' in the English version of P5 that was not present in the Welsh one. This could have stressed the importance of finding all the pentagons. The Welsh version had 'siâp' (shape) inserted between 'tri' (three) and 'pentagon' so that there would not be a mutation of the key word from 'pentagon' to 'phentagon'. The first pre-test version referred to 'four pentagons' so there was no need for the extra word 'siâp' since there is no mutation in 'pedwar pentagon'. The insertion of 'siâp' would not be expected to affect performance unless pupils stopped reading when they reached it.

5.3.3. *Everyday words or expressions*
As regards the existence of alternative words or expressions, the assumption was made that the language with no alternatives was favoured. Thus, Welsh-medium pupils could have been disadvantaged in P1/2. However, there was a picture of pencils filling a box and the form 'sawl' is in common use.

In the case of Welsh, one can assume that a mutated form of a word might be more unfamiliar than the unmutated one. Otherwise, decisions

about which language has the more difficult word or expression can be rather subjective. In P3 and P4 'defnyddiwch' was noted as being difficult by Welsh-medium teachers and in P4 'gwblhau' by Welsh-medium teachers and 'complete' by English-medium teachers. However, both of these items favoured Welsh-medium pupils.

5.3.4. *Mathematical vocabulary*
All words or expressions that are part of mathematical vocabulary are listed. Terms that are either the same or very similar in both languages: 'chart'/'siart' and 'graph'/'graff' in P3, 'pentagon' in P5 and 'shape'/'siâp' in P6 were considered to have a similar effect in both languages. Apart from 'shape'/'siâp' that are both part of everyday English and Welsh, the others are probably learnt in the classroom.

5.3.5. *Item P6: Right angle*
The item showed pupils six polygons and asked them to tick the one that had two right angles. That shape was a pentagon with the two right angles formed by two vertical lines and one horizontal line. The four factors listed in Table V will now be discussed in turn.

The second pre-test version of the first sentence was different in both languages. The English sentence was noted as favouring English-medium pupils since it tells the pupils what to do. The Welsh sentence corresponded to the English sentence used in the first pre-test: "One of these shapes has two right angles." Performance was better on the Welsh version in both pre-tests regardless of the initial sentence, so it is likely that the introductory sentence did not have a crucial role in the differential performance.

The difference between being instructed to tick the shape and to place the tick inside it was not regarded as affecting differential performance since any unambiguous indication of the correct shape would have received credit.

Since the feminine form of the number 'two' (dwy) was used rather than the usual masculine one used in counting (dau), this was noted as favouring the English-medium pupils. The feminine form was used because the word for 'angle' (ongl) is feminine in Welsh. However, both 'dau' and 'dwy' are similar to each other and the Welsh-medium pupils did not seem to be disadvantaged.

The Welsh term for 'right angle' is 'ongl sgwâr' (square angle) and has more clarity of meaning than the English term. The terms can be compared in relation to the differences between the English and Welsh mathematics registers and their relationships to ordinary language. Both English and Welsh terms contain the word for 'angle' or 'ongl' which probably play the same role in the two languages in that they are not part of the ordinary

vocabulary of key stage 1 pupils but are used for particular concepts in mathematics lessons. The English term 'right angle' contains the word 'right' that is polysemous in both mathematical and everyday English and the whole term needs to be learned as a unit for this particular concept: separating it into its constituent words could well lead to confusion. Pimm (1987) mentions an example of a pupil referring to 'right angled' and 'left angled' triangles because of the positions of the right angles. On the other hand, the Welsh term 'ongl sgwâr' contains the word 'sgwâr', that does have other everyday meanings such as in the case of 'sgwâr y dref' or 'town square', but its mathematical meaning is unambiguous and separating the whole term into its constituent words can aid understanding.

However, caution needs to be exercised in relation to the possible effect of the 'clarity' of the term on pupil performance. There was an item assessing right angles in each of the three final written papers for the years 2000–2002 and, of these, P6 was the only one that exhibited consistent language DIF in both pre-tests. One of the others did not exhibit language DIF at all and the other favoured Welsh-medium pupils in the first pre-test and, after some re-wording, English-medium pupils in the second pre-test.

5.3.6. *Item T2: Two-digit numbers*

Pupils were shown a sheet of paper with pictures of fruits in baskets. Each pupil was told by the teacher how many fruits there were in his or her basket and asked to write the number on a work sheet. The three factors listed in Table V will now be discussed in turn.

The word 'gellygen' ('pear') could have been considered unfamiliar for some pupils. As always in a task situation, the teachers could have used another name or pointed to the picture to make the item understandable.

Although there are two Welsh terms corresponding to the English 'number': 'rhif' for the arithmetical value and 'nifer' for the total count as in 'number of apples', this was not an issue in T2 where 'rhif' was used.

As discussed in Section 2.4, Welsh has a more regular oral counting system than English (Roberts, 2000). The present task required pupils to write down the two-digit numbers that they heard. The Welsh-medium pupils heard the digits in the same order as they needed to be written, 'un deg wyth' (one ten eight), whereas the English-medium pupils heard them in the reverse order, 'eighteen' (eight ten). Since the pupils only hear the names of the numbers, there was a possibility of ambiguity in relation to numbers such as 'eighteen' and 'eighty' in English. However, there was a general instruction covering all tasks for teachers to discuss ambiguous or unexpected responses with pupils and to record an assessment on the basis of that discussion.

There was only one other item assessing names of two-digit numbers that was pre-tested twice. This was an oral item on a paper-based test and it favoured Welsh-medium pupils in one pre-test and did not exhibit DIF in the other. However, there were two task items in one year's first pre-test that both favoured Welsh-medium pupils.

6. Discussion

6.1. *Methodology*

An attempt was made to deal with the unreliability of DIF analyses by considering the occurrence of DIF within more than one sample. However, there was a problem in doing this with the available data from pre-testing in that the contents of the assessment materials were not the same in the two pre-tests and also some of the items themselves had often undergone changes. In spite of this, it did allow the consideration of the factors that might affect performance.

The categories of analysis used by other authors proved a useful starting point, but an important shortcoming of the present work is that there was no systematic study of pupils' work in answering the questions. Such a study could have helped in choosing between some of the possible factors that could have been causing DIF. In general, test development has the potential to be used to further the understanding of factors affecting performance at item level by inserting a research question with the materials, such as by varying a factor within a particular item.

6.2. *Results*

As regards the possible factors that could affect the occurrence of language DIF, it seems that non-linguistic ones can hardly ever be discounted and, in some cases, could to be the only plausible explanation. This does highlight the need to have sufficient information about the differences between language groups in relation to such issues as curriculum and pedagogy. As discussed by Emenogu and Childs (2005), curriculum differences might be related to differences in performance at item level even within the same country.

Each of the items discussed in this paper contained a number of possible linguistic factors that could affect pupil performance. Deciding which factor played the dominant role is not clear cut as was reported by Gierl and Khaliq (2000) who discussed a case where one translator predicted that a mathematics item would favour English-medium students and another

that it would favour French-medium students. Each person focussed on a particular aspect of the wording of the item which turned out to favour the English-medium students.

In relation to the possibility that words from the Welsh and English mathematics registers had different effects on pupil performance, both the Welsh for 'right angle' and the Welsh names of two-digit numbers could have helped Welsh-medium pupils, but there are other contending explanations, such as possible differences between Welsh- and English-medium classrooms in the materials that are used or the emphasis on learning mathematical vocabulary. However, given the results presented here and that of others such as Han and Ginsburg (2001) in relation to Chinese mathematical vocabulary or Dowker and Lloyd (2001) in relation to regular counting names in Welsh, intrinsic qualities of mathematical words and expressions do deserve serious consideration.

In their description of the question answering process, Pollitt and Ahmed (1999) discuss the role of words in examination questions in provoking specific schemas and the need for students to activate appropriate schemas in order to answer the questions. It was the different effects of the use of 'similar' in English and 'cyflun' in Welsh that was one possible explanation for the different performances reported by Jones (1993). In relation to the results of this paper, both 'ongl sgwâr' and the Welsh number names have structures that are related to the concepts to which they refer.

Occurrences of DIF need to be judged as to whether they are related to constructs that are relevant or irrelevant to the construct that the test is to measure (Camilli and Shepard, 1994). It is possible to argue for the testing of pupils' knowledge of the names of numbers or of the mathematical vocabulary that is relevant for their age group. However, in order that meaningful cross-language comparisons of attainment can be made, there is a need to understand the effects that differences in the mathematics registers of different languages might have on pupil performance. Mathematical vocabulary can play a different role to other vocabulary in assessment items in that it can be a part of the mathematics being assessed. Where there is a possibility that it helps one language group more than another there is a need to lessen that effect. For example, in the case of the possible different cueing properties of 'similar' and 'cyflun' mentioned above, there has been the use of 'mathematically similar' in English rather than simply 'similar'.

To conclude, the purpose of the paper was to have an overview of possible language issues in the case of assessment materials for 7-year-olds in Wales by comparing the performance of English- and Welsh-medium pupils during pre-testing. It is in the nature of test development to be concerned with creating test instruments that do the job they were intended to

do rather than answering fundamental questions but, during the process, issues do become apparent and it is important to raise these in order to guide further research, so that there is a greater understanding of the role of language in the assessing of attainment in mathematics.

ACKNOWLEDGMENTS

The first version of the paper was written when the author was employed by NFER. The data used was from the development of the 2000–2002 KS1 mathematics assessment materials for ACCAC. The author is grateful for the comments from various reviewers during different stages of writing the paper.

REFERENCES

ACCAC: 1998, *Y Termiadur Ysgol: Termau wedi'u safoni ar gyfer ysgolion Cymru/Standardized terminology for the schools of Wales*, ACCAC, Cardiff, pp. v–xiii.

ACCAC: 2000, *Mathemateg yn y Cwricwlwm Cenedlaethol yng Nghymru/Mathematics in the National Curriculum in Wales*, ACCAC, Cardiff.

Allalouf, A., Hambleton, R., and Sireci, S.: 1999, 'Identifying the causes of translation DIF on verbal items', *Journal of Educational Measurement* 36, 185–198.

Baker, C.: 1984, 'Language background classification', *Journal of Multilingual and Multicultural Development* 5, 43–56.

Camilli, G. and Shepard, L.A.: 1994, *Methods for Identifying Biased Test Items*, Sage, Thousand Oaks, CA.

Chrostowski, S.J.: 2004, 'Developing the TIMSS 2003 background questionnaires', in M.O. Martin, I.V.S. Mullis and S.J. Chrostowski (eds.), *TIMSS 2003 Technical Report*, TIMSS & PIRLS International Study Center, Chestnut Hill, MA, pp. 66–91.

DfEE: 1999: *The National Numeracy Strategy: Mathematical Vocabulary*, DfEE, Sudbury.

Dowker, A. and Lloyd, D.: 2001, 'Linguistic influences on mathematics learning: A study of Welsh bilingual and English monolingual children in Wales', Paper presented at the British Psychological Society Developmental and Education Sections Joint Conference, 2001.

Durkin, K.: 1991, 'Language in mathematical education: An introduction', in K. Durkin and B. Shire (eds.), *Language in Mathematical Education: Research and Practice*, Open University Press, Milton Keynes, pp. 1–3.

Durkin, K. and Shire, B.: 1991, 'Lexical ambiguity in mathematical contexts', in K. Durkin and B. Shire (eds.), *Language in Mathematical Education: Research and practice*, Open University Press, Milton Keynes, pp. 71–84.

Emenogu, B.C. and Childs, R.A.: 2005, 'Curriculum, translation and differential functioning of measurement and geometry items', *Canadian Journal of Education* 28(1/2), 128–146.

Gierl, M.J. and Khaliq, S.N.: 2000, 'Identifying sources of differential item functioning on translated achievement tests: A confirmatory analysis', Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Hambleton, R.K.: 1993, 'Translating achievement tests for use in cross-national studies', *European Journal of Psychological Assessment* 9, 57–68.
Han, Y. and Ginsburg, H.P.: 2001, 'Chinese and English mathematics language: The relation between linguistic clarity and mathematics performance', *Mathematical Thinking and Learning* 3(2/3), 201–220.
Jones, D.: 1993, 'Words with a similar meaning', *Mathematics Teaching* 145, 14–15.
Jones, D.: 1998, 'National curriculum tests for mathematics in English and Welsh: Creating matched assessments', *Assessment in Education* 5, 193–211.
Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., and Chrostowski, S.J.: 2004, *TIMSS 2003 International Mathematics Report*, TIMSS & PIRLS International Study Center, Chestnut Hill, MA.
NFER: 2001a, *Key Stage 1 Levels 2–3 Mathematics Test for Wales in 2002: Report on the Second Pre-Test*, National Foundation for Educational Research, Slough, unpublished.
NFER: 2001b, *Key Stage 1 Level 1 Mathematics Task for Wales in 2002: Report on the second pre-test*, National Foundation for Educational Research, Slough, unpublished.
Park, K.: 2004, 'Factors contributing to Korean students' high achievement in mathematics', Paper presented at the 10th International Conference on Mathematical Education.
Pimm, D.: 1987, *Speaking Mathematically: Communication in Mathematics Classrooms*, Routledge, London.
Pollitt, A. and Ahmed, A.: 1999, 'A new model of the question answering process', Paper presented to the International Association for Educational Assessment, Bled, Slovenia.
Ruddock, G. and Evans, S.W.: 2000, 'Developing tests in two languages (or two variants of the same language)', Paper presented at the 26th Annual International Association for Educational Assessment Conference.
Roberts, G.: 2000, 'Bilingualism and number in Wales', *International Journal of Bilingual Education and Bilingualism* 3, 44–56.
Setati, M.: 2004, 'Mathematics education and language: Policy, research and practice in multilingual South Africa', in C. Keitel, J. Adler and R. Vithal (eds.), *Mathematics Education Research in South Africa: Challenges and Possibilities*, HSRC, Pretoria, pp. 73–109.
Shimizu, Y. and Zumbo, B.D.: 2005, 'A logistic regression for differential item functioning primer', *Japan Language Testing Association Journal* 7, 110–124.
Shuard, H. and Rothery, A.: 1984, *Children Reading Mathematics*, John Murray, London.
Wiliam, D.: 1994, 'Creating matched National Curriculum assessments in English and Welsh: Test translation and parallel development', *The Curriculum Journal* 5, 17–29.

6 Saint John's Crescent
Treganna, CAERDYDD/CARDIFF
CF5 1NX, Cymru/Wales
U.K.
Tel: +44-2902-222350
E-mail: swevans@tiscali.co.uk