Cross-country Comparisons of Inattentive, Hyperactive and Impulsive Behaviour in School-Based Samples of Young Children

Dr Christine Merrell,
Durham University, England

Professor Irene Styles,
The University of Western Australia

Paul Jones,
Durham University, England

Peter Tymms,
Durham University, England

Helen Wildy
The University of Western Australia, Australia

Keywords: Inattention, hyperactivity, impulsivity, ADHD, cross-country comparison

Abstract

This paper uses the Rasch measurement model to analyse data collected on children's attention, activity and impulsiveness at the end of their first year at school by teachers in England, Scotland and Australia. The analysis offers insights into differences in teachers' perceptions of children's behaviour between countries and changes with age. The analysis of large school-based samples from three countries (approximately 2,500 children per country) indicated that the Scottish teachers perceived the behaviour of their pupils as more problematic than teachers in Australia and England. The reasons for this are not clear but it raises issues for researchers wishing to compare behaviour characteristics internationally. Possible ways forward are discussed.

Introduction

To what extent do teachers of pupils at the end of their first year of school in England, Scotland and Australia vary in their ratings of pupils' behaviour in relation to inattention, hyperactivity and impulsivity and do these behaviours change with age? This paper uses the Rasch measurement model to analyse data to attempt to answer these questions. The analysis of large school-based samples from three countries (approximately 2,500 children per country) builds upon other analyses that have used the Rasch measurement model.

Attention Deficit Hyperactivity Disorder (ADHD) is characterised by inattentive, hyperactive and impulsive behaviours and for a diagnosis tight criteria must be met. This paper does not address the issue of diagnosis nor does it seek to study pupils with a diagnosis. Rather it looks at the manifestation of characteristics associated with ADHD as observed by teachers. The characteristics were derived from the fourth version of the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) (American Psychiatric Association, 1994). It identifies three sub-types of ADHD; 'Combined', where an individual displays symptoms of inattention, hyperactivity and impulsivity, 'Predominantly Inattentive' and

ISSN 1838-0689 online

www.education.monash.edu.au/irecejournal/

'Predominantly Hyperactive/Impulsive'. The DSM-IV criteria are used internationally and so it is important to investigate whether perceptions of children's behaviour are the same in different countries. Teachers' ratings are used in this study and whilst it is acknowledged that teachers are not clinicians, they are professionals in a position where they do identify children with behavioural problems and make an important contribution in referrals to and discussions with educational psychologists and child psychiatrists.

The underlying cause of ADHD, and whether or not it is a single disorder or a collective term for multiple disorders, has been the subject of debate which is reflected in the evolving diagnostic criteria used in past versions of the DSM. The DSM-II (American Psychiatric Association, 1968) referred to the disorder as 'Hyperkinetic Reaction of Childhood', with the main defining feature being hyperactivity. Later, Douglas (1972) argued that children labelled as being hyperactive also exhibited problems of impulsiveness and inattention, leading the disorder to be re-named in the DSM-III as 'Attention Deficit Disorder with or without Hyperactivity' (ADDH and ADD respectively). In their discussion of the scientific basis and educational implications of diagnosing ADHD using DSM criteria, McBurnett, Lahey and Pfiffner (1993) described some of the criticisms voiced by researchers over the complexity of the diagnostic criteria in the DSM-III and the validity of the 'Attention Deficit Disorder without Hyperactivity' classification. These concerns influenced the development of the diagnostic criteria to be included in a revised edition (DSM III-R) (American Psychiatric Association, 1987) which presented a single scale of fourteen items representing the symptoms of Attention Deficit Hyperactivity Disorder (ADHD). Still the debate concerning the true definition and the root cause of the disorder continued but the DSM-IV settled on three subtypes (Lahey et al., 1994), which are defined as follows: Combined (when six or more criteria from the nine relating to inattention and six or more criteria from the nine relating to hyperactivity/impulsivity are met); Predominantly Inattentive (when six or more criteria relating to inattention are met); and Predominantly Hyperactive/Impulsive (when six or more criteria relating to hyperactivity/impulsivity are met). For a diagnosis of ADHD, an individual should have displayed the symptoms at a severe and persistent level for at least the preceding six months in more than one context, and onset should have been before age 6.

Research emerging since the publication of the DSM-IV is consistent with the view that the cause of ADHD is complex and multi-factorial (Sergeant, Geurts, Huijbregts, Scheres, and Oosterlaan, 2003) and that it is associated with executive function domains (Willcutt, Doyle, Nigg, Faraone, and Pennington, 2005). Behavioural inhibition enables the processing of information by the executive functions to prevent individuals from reacting to a stimulus too rapidly. Barkley, (1994 and 1997) suggested that behavioural inhibition in individuals with ADHD is impaired, leading to impaired executive functions, causing an individual to be hyperactive and impulsive. These individuals are also likely to appear inattentive. However, impaired executive functions do not account for all cases of ADHD, and Wilcutt et al. (2005) and Sonuga-Barke (2005) suggested that motivational development could also be important. It has also been suggested that the symptoms of inattention associated with the Predominantly Inattentive subtype could have a different root cause to the symptoms of inattention seen in the Combined or Predominantly Impulsive/Hyperactive subtypes (Barkley, 1994; Lahey, Pelham, Loney, Lee and Wilcutt, 2005; Milich, Balentyne and Lynam, 2001). Analysis of the diagnostic criteria reveals that they fall into two dimensions with inattention forming one dimension and hyperactivity/impulsivity the other, suggesting the presence of two separate disorders (Anastopoulos, Barkley and Shelton, 1995; Merrell and Tymms, 2005; Milich, Balentyne, and Lynam, 2001; Smith and Johnson, 2000). Further, there is evidence to suggest that hyperactivity could be divided into physical and verbal hyperactivity (Merrell and Tymms, 2005) and that impulsivity is not a unitary construct (Evenden, 1999).

The fifth version of the DSM, which is currently under development, proposes retaining the three sub-types of ADHD listed in the DSM-IV and adding a further Inattentive (Restrictive) presentation (American Psychiatric Association, 2010(a)). The two categories for inattentive behaviour address the criticism that the DSM-IV did not accurately allow for purely inattentive individuals (American Psychiatric Association, 2010(b)).

The prevalence of individuals diagnosed with ADHD varies depending on factors such as age, the reliability of diagnostic criteria and diagnostic practices. There are cultural differences in the level of activity and inattention that are regarded as problematic (Sonuga-Barke, Minocha, Taylor and Sandberg, 1993). In their study of teachers' ratings of the behavioural deviance of native 'English' and West Indian children living in inner city areas, Rutter et al. (1974) found that the behaviour of over 40% of West Indian children compared with less than 20% of native 'English' children was rated as 'deviant' by teachers. It might be argued that children in different groups behave differently and one way to measure whether or not ethnic bias does exist in the diagnosis of ADHD is to compare subjective teacher assessments with objective measures of the pupil's actual behaviour. Sonuga-Barke et al. (1993) did just this in their investigation of the relationship between subjective ratings of hyperactivity and attention in groups of children classified as being of Asian or English origin, attending primary schools. They concluded that teachers appeared to over estimate the Asian children's levels of activity relative to those of the English children.

This paper reports the use of the Rasch measurement model (Andrich, 1978; Bond and Fox, 2001; Rasch, 1960/80) to analyse data collected on children's behaviour by teachers in England, Scotland and Australia to investigate the presence of the three components of inattention, hyperactivity and impulsivity, and differences in perceived behaviour. It builds upon other analyses that have used the Rasch measurement model to analyse the severity and principal components of inattention, hyperactivity and impulsivity (Smith and Johnson, 2000; Merrell and Tymms, 2005; Young, Levy, Martin and Hay, 2009).

Measures and Sample

Data for this study were collected from schools in England, Scotland and Australia who used the Performance Indicators in Primary Schools (PIPS) monitoring system run by the Centre for Evaluation and Monitoring (CEM) at Durham University, UK and The University of Western Australia. The system, which began in England in 1992, monitors the progress of children as they move through primary schools (Tymms, 1999). It has expanded to include several thousand schools in England and also schools in Scotland and Australia. The schools assess their pupils on a regular basis and the data are returned to the centres for analysis. Participation is voluntary but schools pay an annual fee and complete a form which states that they have satisfied themselves that parents/guardians have been given sufficient information about the purpose of the assessment and that anonymous pupil and school-level data will be used for research purposes. Schools receive standardised feedback about the attainment and progress of their pupils and as a result of this process, CEM holds a large, longitudinal dataset. The samples of schools in England and Scotland have been found to be nationally representative. The schools in Australia were mainly located in Western Australia and not necessarily representative of the west or of Australia as a whole.

The PIPS monitoring system begins with a baseline assessment of children on entry to school (the PIPS BLA). This is a computer-delivered assessment of early reading, phonological awareness and mathematics that is administered by a teacher working with one pupil at a time. The computer program is adaptive and uses a series of stopping rules so that when items within a section become too difficult for a child, the assessment moves to the next section. The assessment is repeated at the end of the first year of school, at which point it also includes an assessment of pupils' behaviour which schools can choose to complete in addition to the main part of the assessment. For this section, teachers are asked to complete a rating scale which is based on the DSM-IV diagnostic criteria for ADHD. The teacher is asked to rate each pupil against eighteen criteria (nine relating to inattention, nine relating to hyperactivity and impulsivity). Unlike the DSM-IV, each item has ten response categories. In practice, teachers are presented with a sliding scale labelled *Never* to *Always* on which they can locate a child. For more information about the content and psychometric properties of the PIPS BLA and its behaviour rating scale, see Merrell and Tymms (2001), Merrell and Tymms (2004), Tymms (1999), Wildy and Styles (2009). A full list of items in the behaviour rating scale can be found in Table 1.

The sample consisted of pupils who had completed their first year of school in England (14,272 pupils), Scotland (2,588 pupils) and Australia (10,064 pupils). Not all schools completed the behavioural ratings but only data from schools where whole cohorts were assessed were included in the samples. To obtain more equally sized groups, the scores from all pupils from Scotland were analysed, and a similar number of pupils selected randomly from each of the English and Australian samples. Ages ranged from about four to seven years at the time of the assessment, depending on national policies and practices for the age at which pupils started school.

Analyses and Results

The analysis of the psychometric properties of the PIPS BLA behaviour rating scale was carried out using the Rasch measurement model (see for example Andrich, 1978; Rasch, 1960/80). The Rasch measurement model creates a uni-dimensional equal-interval scale from a set of items and is used to check that the items can indeed be used to create that scale. It is also used to see if the scale that holds across different groups of people. It can be used to investigate the properties of items and individuals taking the assessment; they are both located on the same scale. In the same way as a ruler can measure the distance between a set of points, by using Rasch Measurement the difficulty of items in terms of 'distance' from each other on a scale can be seen. The units of the scale are referred to as 'logits' and each item has a logit value. Each person taking the assessment also has a logit value which shows their ability on the scale. The more positive the logit value for an item, the harder it is to score on that item. The more negative the logit value, the easier it is to score on that item. Therefore when we are investigating items on the behaviour rating scale, a high logit value indicates a high score, i.e. the behaviour is displayed infrequently in the population. The higher the logit value for a person, in this case, the more behavioural problems they exhibit. It is also possible to find the mean location on the scale for groups of pupils such as boys and girls and make comparisons. In the context of this paper, comparisons were also made between the mean locations of the pupils of different ages and gender in the three countries. The software RUMM2030 (Andrich, Sheridan and Luo, 2010) was used.

Before comparing the behaviour of the different groups of pupils, we use the Rasch model to ensure that the scale is measuring the construct consistently and reliably. The steps in the analysis of the scale followed those set out in Andrich and Styles (2004), namely:

1. Thresholds; As described above, teachers were asked to rate the behaviour of their pupils against a set of criteria using a ten-point scale. A threshold is the point on the logit scale that signifies a change from one point on that ten-point scale to the next. Sometimes the thresholds are "disordered" and this would indicate that unexpectedly the thresholds did not increase monotonically along the scale. Probably the teachers had difficulty discrimination between all points on the scale.
2. Item fit; The Rasch model assumes that a scale is measuring a single construct and that the items in the scale fit well within it. Therefore we need to check that the items do fit the model and whether any are not congruent with the construct under investigation.
3. Differential item functioning (DIF); Do the items in the scale have the same relative difficulty across groups of interest such as gender and country groups? For example, whilst the mean score of the individuals in one group might be higher than a second group (maybe boys score more highly than girls), we would still expect individual items within the assessment which are most frequently met by one group to also be most frequently met by the second group. We do not expect the item difficulty to vary across groups (such variation is called 'item bias').
4. Item and person distributions; are the items targeted to the sample of persons used for the analysis? Are the items placed evenly along the logit scale in terms of their difficulty values or are there gaps in some areas of the scale and too many items of equivalent difficulty in other areas of the scale? Are the scores of the individuals who have been assessed distributed as expected?
5. Item dependencies; are the residual correlations amongst items highly correlated with each other? This may provide evidence of possible sub-scales.
6. Reliability; if all the above are acceptable, how reliable are the measures obtainable with the scale? The Person Separation Index (equivalent of Cronbach's alpha) with and without extreme locations gives an indication of this.
7. Item locations; are the items ordered in difficulty (or intensity) approximately as expected?

Item Thresholds

Nine items had reversed thresholds, indicating that their response categories were not operating as required. In other words, the teachers experienced difficulty discriminating between points on the scale. A further four or five items had thresholds which were very close together. The threshold probability curves showed that it was particularly the middle categories which were disordered. On the basis of these results, a decision was taken to rescore all items using just five categories by collapsing each adjacent pair of categories (thus 0 and 1 were scored as 0, 2 and 3 as 1, 4 and 5 as 2, etc). We note here that rescoring post hoc is not recommended, but for research purposes rescoring can give an indication of whether this might work better. In future, new data using the reduced numbers of categories would need to be collected to confirm the rescoring. After rescoring, there were no further reversed thresholds. Examples of the Category Characteristics curves (CCCs) for one item before and after rescoring are shown in Figures 1 and 2.

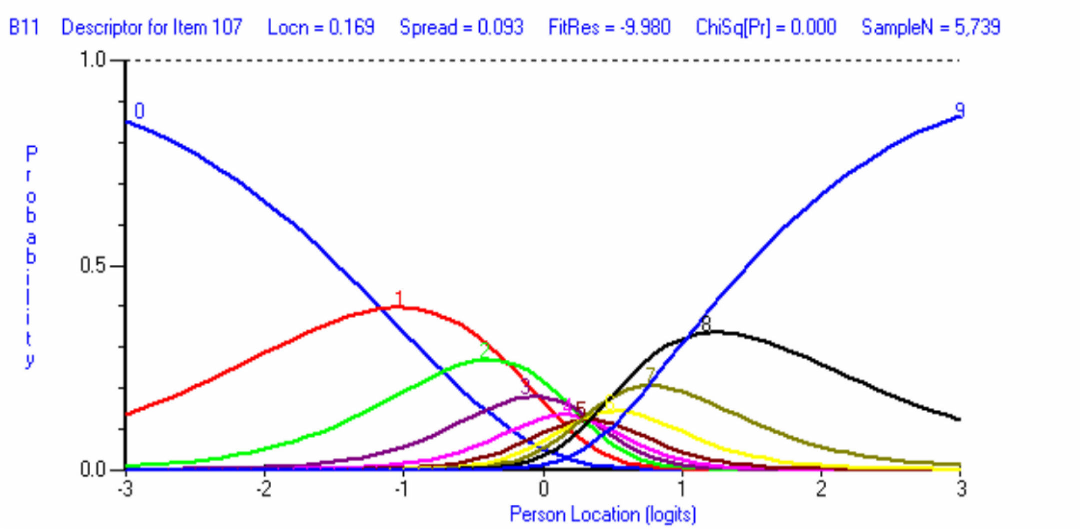

Figure 1: Category Characteristic curves for Item B11 with 10 response categories

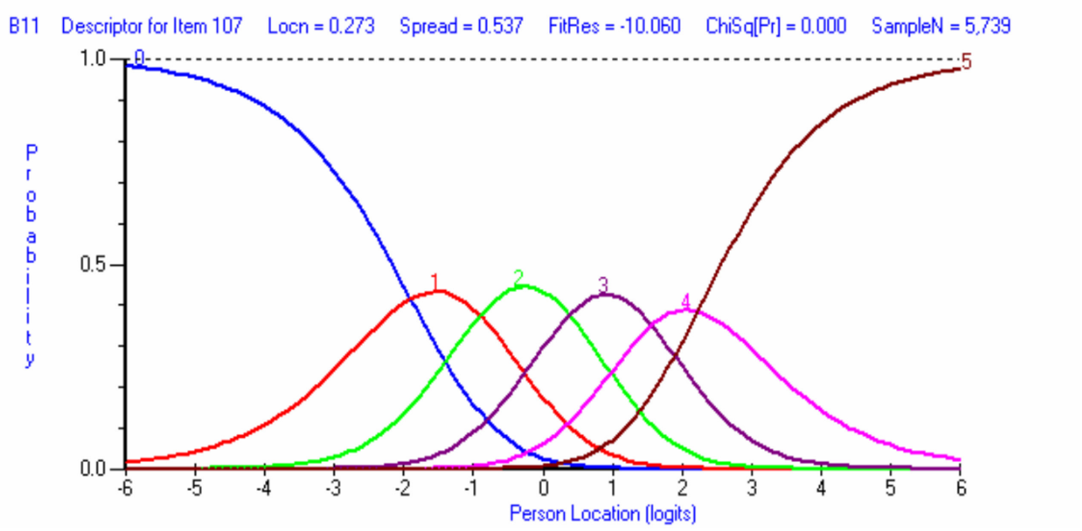

Figure 2: Category Characteristic curves for item B11 with the number of response categories were reduced to five

The analysis of thresholds indicated that the teachers did not distinguish meaningfully amongst the 10 response categories. Instead of the reliability (an index of the precision of measurement) decreasing when the number of categories was reduced to five categories, it increased a little from 0.934 to 0.948, providing supporting evidence that reducing the categories improved the quality of the measures.

Item fit

It is important to see how well the items fit the Rasch model and whether any items are mis-fitting. The rescored items were considered. Three tests of fit were used here. The first was the item-trait interaction test of fit; a chi-square test between the expected and observed scores on each item. The chi square test compares proportions of expected with actual values within discrete categories and so, in this case, we created six step-interval categories of person locations on the test as a whole. The second was a log residual test of fit which explores the consistency of the pattern of person responses for each item, i.e. how closely the person responses fit the model. The third was a graphical test of fit which looked at how well observed score values fit the theoretical item characteristic curve (ICC). The number of children used in the calculation of the chi square statistics was reduced from the full sample size of 6,090 to 1,800, based on the number of categories and the number of items in the scale (Andrich and Styles, 2009). Six items showed some misfit according to the chi square test and the log residual test identified two items as being potentially problematic, however, when the ICCs for these items were inspected, their fit looked acceptable, though it is clear the scale is a complex one. Figure 3 shows the ICC for the least well-fitting item which tends to be slightly under-discriminating. Based on these results, a decision was taken to retain all items.

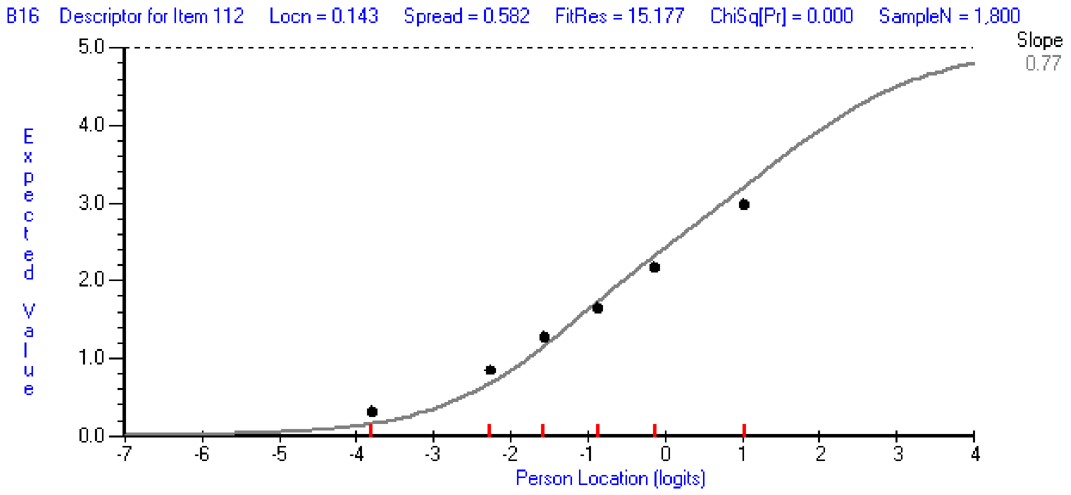

Figure 3: ICC of least well-fitting item

Bias as identified by Differential Item Functioning (DIF)
Two items displayed some DIF according to country ('Makes careless mistakes in school work or other activities' and 'Is often 'on the go' as if driven by a motor') and four according to gender ('Often runs about excessively in situations in which it is inappropriate', 'Has difficulty in playing quietly', 'Is often 'on the go' as if driven by a motor' and 'Talks excessively'). The sample sizes of first the gender and then the country groups were made equal (by random selection of students) for the DIF analyses.

For exploratory purposes, the items which exhibited DIF items were split, one at a time, on a group basis, splitting the item with most DIF first, then reanalysing the data until no further significant DIF remained. This means that the difficulty value of an item was estimated for one group and then separately for another group, which has the advantage of being able to retain all items in the analysis, though of course the split items are no longer available to compare children across groups. It was necessary to split 'Often runs about excessively', 'Talks excessively', 'Is often "on the go"', and 'Has difficulty playing quietly' (in that order) according to gender group, but necessary only to split 'Makes careless mistakes' according to country group before DIF was judged of no further significance. It is interesting to note that the DIF by gender only occurred in items concerned with hyperactive behaviour. Boys were likely to be scored higher on 'often runs about excessively', 'often on the go' and 'difficulty playing quietly', and girls higher on 'talks excessively' even though they had the same total scores. It was noted that boys tended to be scored higher than girls on these items across the range of total person locations. These results suggest that these behaviours are viewed by teachers in qualitatively different ways for girls and boys. The Australian sample's scores were higher on the item 'makes careless mistakes in school work and other activities' than expected. This suggests that the way in which teachers in Australia interpret and apply this criterion is qualitatively different from teachers in England and Scotland.

Because it is not feasible to routinely take account of DIF in this way and because the analysis adjusts item and person locations to take account of DIF, the scale was retained in its original form for the later substantive analysis where group mean locations were compared.

Targeting (item and person distributions)
The distribution of items and persons is shown in Figure 4.

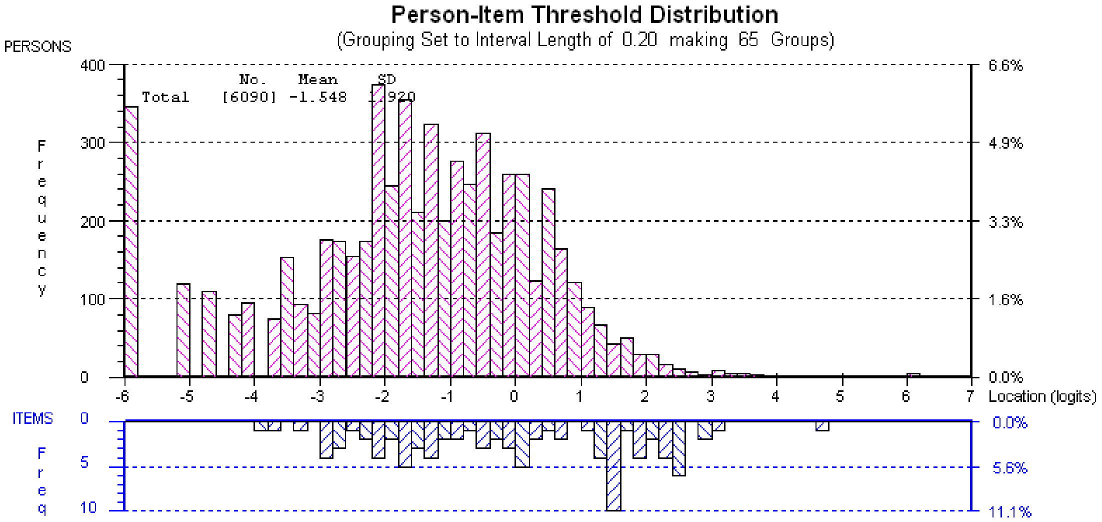

Figure 4: Distributions of person and item threshold locations for the Behaviour scale

A large number of children were given the lowest possible rating (never) for all items. At the other end of the scale, very few children exceeded the highest threshold location. The purpose of this scale is to identify children with severe behavioural problems that are likely to impact on their later outcomes and so it is acceptable that those who have few or no difficulties are not being measured as reliably as the children with behavioural difficulties. Except for one threshold at the highest end, the item thresholds were located fairly evenly along the continuum.

Item dependencies
Once what is common amongst items was removed and the correlations amongst the residuals examined, several pairs of items showed dependencies on each other (residual correlations are >0.3, in this case). They indicate items where there is redundancy because the pairs of items are measuring the same aspects to a significant extent. In this analysis, the principal component analysis of these residuals' correlations suggested the presence of two sub-scales comprised of the first eight items (all related to inattention) and the last eight items (all related to hyperactivity and impulsivity), respectively. The two middle items in the scale ('Forgetful in daily activities' and

'Fidgets with hands or feet or squirms in seat') showed no significant relationship with either subscale. In view of these findings, a sub-scale analysis was carried out. (It should be noted that principal components analysis is the same as factor analysis. Its purpose is this analysis is to identify the common factor that explains most of the residual variance rather than construct variables. For a more detailed explanation, see www.rasch.org).

The Person Separation Indices, with and without extreme scores, were 0.948 and 0.955, respectively, indicating high reliability, however, in the presence of sub-scales and item dependencies, these values will be inflated. Note that the Person Separation Index shows the reliability of the person estimate. In other words, a high person separation index suggests that we can be confident that individuals with, say, high scores on the scale do have higher measures than those with lower scores.

Item locations

Table 1 shows the item locations on the logit scale in order of intensity from least to most intense. Those items at the least intense end of the continuum were items that the teachers found easiest to rate their children as exhibiting. Items such as 'Distracted by extraneous stimuli' and 'Difficulty sustaining attention in tasks and play activities' were relatively easy for teachers to rate as being observed, whereas teachers tend to rate severe (i.e. 'always') levels of behaviour such as 'Often runs about excessively in situations where inappropriate' and 'Is often "on the go" as if driven by a motor' as being less frequently observed. The order of items indicated that, in general, items related to inattention were more frequently observed than those related to hyperactivity and impulsivity. This echoes the findings of a similar analysis of young children in the UK by Merrell and Tymms (2005).

Table 1: Item locations for the full behaviour scale

Item	Item content	Location (logits)	Std error
8	Distracted by extraneous stimuli	-1.001	0.017
2	Difficulty sustaining attention in tasks and play activities	-0.564	0.017
10	Fidgets with hands or feet or squirms in seat	-0.341	0.016
6	Reluctant to engage in tasks which require sustained mental activity	-0.232	0.017
5	Difficulty organising tasks and activities	-0.174	0.017
4	Does not follow through instructions, fails to finish work	-0.095	0.017
9	Forgetful in daily activities	-0.06	0.017
1	Makes careless mistakes in school work or other activities	-0.038	0.019
3	Does not seem to listen when spoken to directly	-0.003	0.017
15	Talks excessively	0.025	0.016
16	Blurts out answers before questions have been completed	0.143	0.016
17	Has difficulty awaiting turn	0.151	0.017
18	Interrupts or intrudes on others	0.180	0.017
13	Has difficulty playing quietly	0.193	0.017
11	Leaves seat in classroom or other situations where expected	0.273	0.017
7	Looses equipment necessary for activities	0.373	0.017
14	Is often "on the go" as if driven by a motor	0.525	0.017
12	Often runs about excessively in situations where inappropriate	0.647	0.017

Sub-scale analysis

When the scale was divided into two sub-scales; inattention and hyperactivity/impulsivity, the Person Separation dropped from 0.948 for the full 18-item scale to 0.750 confirming the presence of these two sub-scales, which were included in subsequent analyses. It suggests that a proportion of children (in this case, 17% showing a significant difference in locations on the two scales at a 1% level of confidence) would be represented better by a 'profile' of two separate measures rather than a single measure on the scale as a whole.

The concept of unidimensionality in the Rasch paradigm is a question of the level of scale that is useful for a particular purpose. Thus, for an overall indication of behavioural problems the measures on the full scale may be used, but for more detailed diagnostic purposes, a profile of measures on the two subscales may be more useful.

Comparisons of groups

After creating person measures, one way Analysis of Variance (ANOVA) was used to compare mean person locations on the full behaviour scale and for each of its subscales, (inattention and hyperactivity/impulsivity) across gender, age and country groups. Age groups were created for the following categories as shown in Table 2.

Table 2: Age categories (end of year) and numbers of pupils in each (small numbers in brackets)

Category	Age range in years	Number of pupils: England	Number of pupils: Scotland	Number of pupils: Australia
1	Less than 4.49	(7)	-	-
2	4.50-4.99	512	(4)	(17)
3	5.00-5.49	893	445	453
4	5.50-5.99	524	1029	788
5	6.00-6.49	48	507	641
6	6.50-6.99	(5)	(18)	115
7	More than 7.00	(1)	(4)	(11)
Total		1990	2077	2025

The ANOVA table is shown below.

ISSN 1838-0689 online

Table 3: ANOVA for the full behaviour scale

Source	Type III sum of squares	df	Mean squares	F statistic	P value (significance)
Corrected model	1660.654	35	47.447	13.800	0.000
Intercept	345.546	1	345.546	100.500	0.000
Country	21.261	2	10.630	3.092	0.045
Gender	38.535	1	38.535	11.208	0.001
Age group	217.935	6	36.322	10.564	0.000
Country by Gender	31.386	2	15.693	4.564	0.010
Country by Age Group	53.295	10	5.330	1.550	0.115
Gender by Age group	23.315	6	3.886	1.130	0.342
Country by Gender by Age group	36.248	8	4.531	1.318	0.229
Error	20581.425	5986	3.438		
Total	366647.416	6022			
Corrected total	22242.079	6021			

There was a statistically significant main effect for age groups, with behavioural problems becoming less frequent with increasing age, as expected (p<0.000). There was also a statistically significant main effect for gender with girls being rated with fewer (less severe) behavioural problems on average than boys (p<0.001). The differences in locations amongst countries are not directly comparable because pupils begin school at different ages in the three countries: it is better to compare the performance across similar age groups for each country. The interaction between country and age group was not statistically significant (p<0.110). There were also no significant interactions between age and country or age and gender groups, but there was a statistically significant interaction between country and gender groups (p<0.010). The differences in mean locations between girls and boys in Australia and Scotland were similar (0.774 and 0.728 logits, respectively) and significantly less than the difference in means between girls and boys in England (1.788 logits).

Figure 5 illustrates the relationship between age and behaviour for the three countries. Age groups with low pupil numbers (shown in brackets in Table 2) are excluded. The mean behaviour score and 95% confidence interval is shown for each age category.

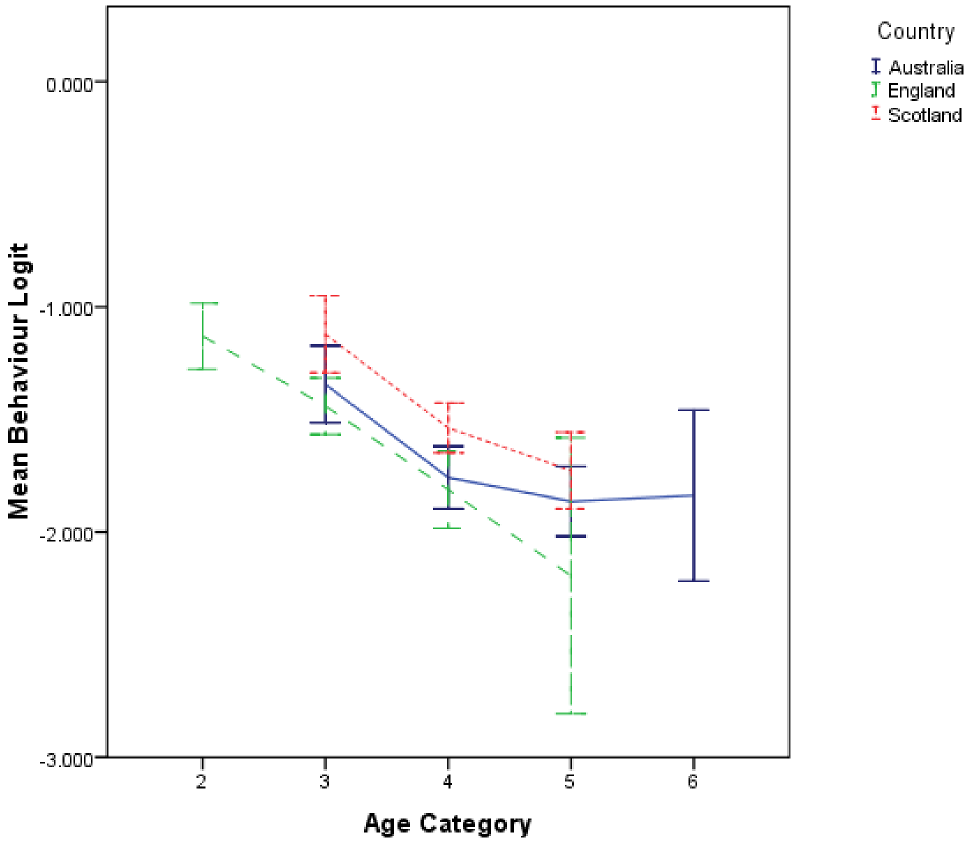

Figure 5: Mean locations on the full Behaviour scale by country and age group

Figure 5 illustrates the trend of decreasing behavioural problems as children got older. Children's executive functions continue to develop through childhood, tending to be fully developed by the age of 7 and many young children exhibit inattentive, hyperactive and impulsive behaviour as a result of immaturity. It is to be expected that some children will continue to exhibit problems as they grow older and some may be diagnosed with ADHD.

Another interesting aspect of Figure 5 is that children in Scotland have higher average ratings for behavioural problems than pupils from England or Australia across almost all age groups.

As found with the full behaviour scale, for the inattention subscale there was a statistically significant interaction between country and gender groups (F=4.953, p<0.007). Here, the largest difference in means between boys and girls was in Australia (1.00 logits), followed by Scotland (0.891 logits), then England (0.375 logits), and girls in all three cases were rated with less severe levels of behavioural problems than boys. There was also a statistically significant main effect for age group (F=12.757, p<0.001). The pattern shown in Figure 5 was very similar to the pattern for the inattentive sub-scale and the hyperactivity/impulsivity sub-scale.

Discussion

Before addressing the question about teachers' ratings of their pupils' behaviour at the end of the first year at school, the psychometric properties of the rating scale were investigated. The analysis of the item thresholds of the PIPS behaviour rating scale suggested that teachers did not meaningfully discriminate between the behaviour of children on a scale that used ten response categories and the reliability of the scale improved when the number of response categories was reduced to five. This has implications for the future development of the this particular behaviour rating scale as well as wider application for rating scales used by teachers. Five categories appear to be a useful number of categories with respect to frequency of behaviour; more than five categories do not contribute useful information. Subsequent analyses in the paper used the rescored items.

The 18 criteria of the full behaviour rating scale were considered to form a good single scale. However a principal components analysis of residual correlations indicated the presence of two sub-scales (inattention and hyperactivity/impulsivity). This supports previous research of separate sub-types of ADHD including the theoretical model currently used by the DSM-IV (American Psychiatric Association, 1994) for the diagnostic criteria. It also suggests that, for some children, a profile of scores on the two subscales would be more informative than a total score on all the items of the full scale.

The order of severity of symptoms was interesting. The correlation between the item logits of this paper and those of a previous study by Merrell and Tymms (2005) was 0.91. Merrell and Tymms used the same items (from the PIPS Baseline Assessment) with children at the end of their first year at school in England only and teachers rated their pupils as meeting a criterion at a severe and persistent level or not, using a yes/no scale. Although the items were identical, the scoring systems for the two studies were different and the item difficulties reported in this paper were based on children from Australia, England and Scotland. This suggests stability of teachers' ratings of young children with respect to the order of severity of behaviours and also similarities whether using a dichotomous or five-point scale.

The trends in decreasing severity of symptoms with increasing age shown in all three country samples suggest that many young children may mature and no longer experience behavioural problems of inattention, hyperactivity and impulsivity. This is consistent with the theoretical position of executive functions developing throughout childhood but needs further investigation with the collection of more longitudinal data.

Conclusion

This paper adds to previous research by comparing teachers' ratings of school-based samples of young children across countries. The teachers in Scotland rated inattentive, hyperactive and impulsive behaviour to be more severe than in Australia or England for children of the same age. This has important implications for the development and interpretation of behavioural rating scales as used by teachers and for the use of the DSM-IV criteria internationally. The DSM criteria are used around the world and considered to be an international standard. Prevalence rates of ADHD characteristics in different countries are reported in studies, which include school-based populations, and are also of research interest (Baumgaertel, Wolraich and Dietrich, 1995; Gaub and Carlson, 1997;Gomez, Harvey, Quick, Scharer and Harris, 1999; Wolraich, Hannah, Baumgaertel and Feurer, 1998). The different ratings between the countries analysed in this study could have a number of causes. One is the teachers' perceptions and another is the children's behaviour. Further research is needed to distinguish between these two possibilities either by using objective measures using instruments which collect data on children's gross motor movements or perhaps by asking teachers in different countries to rate the same videoed behaviour alongside teachers' ratings on the basis of their observations.

The confirmation of two distinct components; inattention and hyperactivity/impulsivity, suggest that, in some instances, it is useful to consider a profile of a child rather than a total score from the Behaviour Rating Scale when implementing interventions to manage behaviour. This informs the types of interventions required to remediate behavioural problems which can be different for each component. Children who are predominantly inattentive may benefit from different remediation programmes from those who are predominantly hyperactive and impulsive. Indeed, those children who display impulsive behaviour in the form of blurting out answers to questions before they have been completed may be actively engaged in learning compared with inattentive children who are easily distracted by extraneous stimuli (a symptom of inattention). Well designed evaluations of the effectiveness of school-based remediation strategies for inattentive, hyperactive and impulsive behaviour are few and far between (Taylor et al., 2009).

More research into the effectiveness of interventions for predominantly inattentive children and predominantly hyperactive/impulsive children is clearly needed. Importantly, in the light of the differences found between teachers' rating in difference countries found in this study, the possibility of extrapolating both the prevalence rates of behavioural problems and ways to remediate them from one culture to another should be considered cautiously.

References

American Psychiatric Association (1968). *Diagnostic and Statistical Manual of Mental Disorders*. Washington D.C.: American Psychiatric Association.

American Psychiatric Association (1980). *Diagnostic and Statistical Manual of Mental Disorders*. Washington D.C.: American Psychiatric Association.

American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders*. Washington D.C.: American Psychiatric Association.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*. Washington D.C.: American Psychiatric Association.

American Psychiatric Association (2010a). *Proposed revisions to the Diagnostic Criteria for Attention Deficit/Hyperactivity Disorder.* http://www.dsm5.org/ProposedRevisions/Pages/proposedrevision.aspx?rid=383# Website accessed 06/06/10.

American Psychiatric Association (2010(b)). DSM-5 Options Being Considered for ADHD, February 2, 2010. http://www.dsm5.org/Proposed%20Revision%20Attachments/APA%20Options%20for%20ADHD.pdf Website accessed on 06/06/10.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43* (4), 561-574.

Andrich, D., Sheridan, B. and Luo, G. (2010). *RUMM2030: Rasch unidimensional models for measurement.* RUMM Laboratory: Perth, Western Australia.

Andrich, D. and Styles, I. (2004). *Psychometric properties of the Australian Early Development Inventory (AEDI)*: Perth, Western Australia: Institute of Child Health Research.

Andrich, D. and Styles, I. (2009). Distractors with information in multiple choice items. In Smith, E.V. and Stone, G. E. (Eds*). Criterion referenced testing: practice analysis to score reporting using Rasch meaurement models.* Maple Grove, Minnesota: JAM Press.

Anastopoulos, A. D., Barkley, R. A., and Shelton, T.L. (1995). The History and Diagnosis of Attention Deficit/Hyperactivity Disorder. In P. Cooper and K. Kent Ideus (Eds.), *Attention Deficit/Hyperactivity Disorder: Educational, Medical and Cultural Issues* (pp. 4-15). The Association of Workers for Children with Emotional and Behavioural Difficulties.

Barkley, R. A. (1994). Delayed responding and attention deficit hyperactivity disorder: Toward a unified theory. Disruptive behavior disorders in children: *Essays in honour of Herbert Quay*. (Ed.) D. K. Routh. New York: Plenum: 11 - 57.

Barkley, R. A. (1997). Behavioural Inhibition, Sustained Attention, and Executive Functions: Constructing a Unifying Theory of ADHD. *Psychological Bulletin* 121(1): 65-94.

Baumgaertel, A., Wolraich, M. and Dietrich, M. (1995). Comparison of diagnostic criteria for attention deficit disorders in a German elementary school sample. *Journal of the American Academy of Child and Adolescent Psychiatry* 34: 629 - 638.

Bond, T. G., and Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Douglas, V. I. (1972). Stop, look and listen: The problem of sustained attention and impulse control in hyperactive and normal children. *Canadian Journal of Behavioural Science*. 4: 259 - 282.

Evenden, J. L. (1999). Varieties of Impulsivity. *Psychopharmacology* 146 (4), 1432-2072.

Gaub, M. and Carlson, C. L. (1997). Behavioural Characteristics of DSM-IV ADHD Subtypes in a School-Based Population. *Journal of Abnormal Child Psychology* 25(2): 103 - 111.

Gomez, R., Harvey, J., Quick, C., Scharer, I. and Harris, G (1999). DSM-IV AD/HD: Confirmatory Factor Models, Prevalence, and Gender and Age Differences Based on Parent and Teacher Ratings of Australian Primary School Children. *Journal of Child Psychology and Psychiatry* 40(2): 265-274.

Lahey, B. B., Applegate, B., McBurnett, K., Biederman, J., Greemhill, L., Hynd, G.W., Barkley, R.A., Newcorn, J., Jensen, P., Richters, J., Garfinkel, B., Kerdyk, L., Frick, P.J., Ollendick, T., Perez, D., Hart, E.L., Waldman, I. and Shaffer, D. (1994). DSM-IV Field Trials for Attention Deficit Hyperactivity Disorder in Children and Adolescents. *The American Journal of Psychiatry* 151(11): 1673 - 1685.

Lahey, B. B., Pelham, W.E., Loney, J., Lee, S.S. and Willcutt, E., (2005). Instability of the DSM-IV Subtypes of ADHD From Preschool Through Elementary School. *Arch Gen Psychiatry 62*, 896-902.

McBurnett, K., Lahey, B. B., and Pfiffner, L.B. (1993). Diagnosis of Attention deficit disorders in DSM-IV: Scientific basis and implications for education, *Exceptional Children* **60**, 108–117.

Merrell, C. and Tymms, P. (2001). Inattention, hyperactivity and impulsiveness: Their impact on academic achievement and progress, *British Journal of Educational Psychology*, 71, p43 – 56.

Merrell, C. and Tymms, P. (2004). Large-Scale Teacher Assessment Of Behavioural Problems In Young Children. Paper presented at the International Association of Educational Assessment Annual Conference, Philadelphia, USA, June 2004.

Merrell, C. and Tymms, P. (2005). Rasch Analysis of Inattentive, Hyperactive And Impulsive Behaviour In Young Children And The Link With Academic Achievement. *Journal of Applied Measurement* 6 (1), 1-18.

Milich, R., Balentyne, A. C. and Lynam, D. R. (2001). ADHD Combined Type and ADHD Predominantly Inattentive Type Are Distinct and Unrelated Disorders. *Clinical Psychology: Science and Practice*, 8(4), 463 – 488.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press.

Rutter, M., Yule, W., Berger, M., Yule, B., Morton, J. and Bagley, C. (1974). Children of West Indian Imigrants -1. Rates of behavioural deviance and of psychiatric disorder. *Journal of Child Psychology and Psychiatry*. 15, 241-262.

Sergeant, J. A., Geurts, H., Huijbregts, S., Scheres, A., Oosterlaan, J. (2003). The top and bottom of ADHD: A neuropsychological perspective. *Neurosci. Biobehav. Rev.* 27, 583–592.
Smith, E. V. and Johnson, B. D. (2000). Attention Deficit Hyperactivity Disorder: Scaling and Standard Setting using Rasch Measurement. *Journal of Applied Measurement* 1 (1), 3-24.

Sonuga-Barke, E. J., Minocha, K., Taylor, E. A., and Sandberg, S. (1993). Inter-ethnic bias in teachers' ratings of childhood hyperactivity. *British Journal of Developmental Psychology*, 11, 187–200.

Sonuga-Barke, E. J. S. (2005). Causal Models of Attention-Deficit/Hyperactivity Disorder: From Common Simple Deficits to Multiple Developmental Pathways. *Biol. Psychiatry* 57, 1231 – 1238.

Taylor, E., Kendall, T., Asherson, P., Bailey, S., Bretherton, K., Brown, A., Costigan, E., Duncan, A., Harpin, V., Hollis, C., Keen, D., Lewis, A., Mavranezouli, I., Merrell, C., Mulligan, D., Perez, A., Pettinari, C., Ryan, N., Salt, N., Sayal, K., Sheppard, L., Stockton, S., Taylor, C., Thorley, G., Turner, J., Tymms, P., Wolpert, M., Wong, I., and Young, S. (2009). *Attention deficit hyperactivity disorder: Diagnosis and management of ADHD in children, young people and adults*. National Clinical Practice Guideline Number 72, National Collaborating Centre for Mental Health, Commissioned by the National Institute for Health and Clinical Excellence.

Tymms, P. (1999). *Baseline Assessment and Monitoring in Primary Schools.*
London: David Fulton.

Wildy, H. and Styles, I. (2009). Measuring what students know and can do on entering school. *Journal of Early Childhood,* 33(4), 43-52.

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. E. and Pennington, B. F. (2005). Validity of the Executive Function Theory of Attention Deficit/Hyperactivity Disorder: A Meta Analytic Review. *Biol. Psychiatry* 57, 1336-1446.

Wolraich, M. L., Hannah, J. N., Baumgaertel, A. and Feurer, I,D (1998).
Examination of DSM-IV criteria for attention deficit hyperactivity disorder in a county-wide sample. *Journal of Developmental and Behavioural Paediatrics* 19(3): 162-168.

Young, D.J. Levy, F., Martin, N.C., and Hay, D.A. (2009). Attention Deficit Hyperactivity Disorder: A Rasch Analysis of the SWAN Rating Scale. *Child Psychiatry Hum. Dev.* 40, 543 – 599.

Authors

Dr Christine Merrell, Director of Research, Centre for Evaluation and Monitoring, Durham University
Mountjoy Research Centre Rowan Block, Stockton Road, Durham DH1 3UZ
United Kingdom
Tel., +44 191 3344226
Christine.Merrell@cem.dur.ac.uk

Professor Irene Styles, Faculty of Education, The University of Western Australia (M428), 35 Stirling Highway, CRAWLEY WA 6009, Australia.
Irene.styles@uwa.edu.au

Dr Paul Jones, Programme Manager, Centre for Evaluation and Monitoring, Durham University, Mountjoy Research Centre Rowan Block, Stockton Road, Durham, DH1 3UZ, England.
Paul.Jones@cem.dur.ac.uk

Professor Peter Tymms, Head of Department, School of Education, Durham University, Leazes Road, Durham, England.
P.B.Tymms@dur.ac.uk

Professor Helen Wildy
Dean, Faculty of Education, The University of Western Australia (M428), 35 Stirling Highway, CRAWLEY WA 6009, Australia.
Helen.Wildy@uwa.edu.au