7-2004

Developers, Users and Consumers Beware: Warnings about the design and use of psycho-behavioural rating inventories and analyses of data derived from them

Ken Rowe
ACER

Kathy Rowe
Royal Children's Hospital, Melbourne

Developers, Users and Consumers Beware: *Warnings about the design and use of psycho-behavioural rating inventories and analyses of data derived from them*[1]

Kenneth J. Rowe

Australian Council for Educational Research

Dr Rowe is a Principal Research Fellow at ACER. Dr Rowe's post-graduate training in research methodology and design, educational psychology, assessment, measurement, psychometrics, and advanced statistical modelling, was undertaken at the University of London, and his doctoral research at the University of Melbourne. Dr Rowe's substantive research interests and expertise include: 'authentic' educational and psychological assessment; multilevel, 'value-added', educational/organisational performance indicators, achievement target-setting and benchmarking; teacher and school effectiveness; differential gender effects of schooling in the context of teaching and learning; the impact of externalising behaviour problems on students' learning outcomes in literacy and numeracy; and the educational/epidemiological implications of Attention-Deficit/ Hyperactivity Disorder (AD/HD) and Chronic Fatigue Syndrome (CFS) in children and adolescents.

Katherine S. Rowe

Department of General Paediatrics, Royal Children's Hospital, Melbourne

Dr Katherine Rowe is a Consultant Physician in the Department of General Paediatrics, and in the Centre for Adolescent Health at Melbourne's Royal Children's Hospital. In addition to her medical qualifications (MB.BS, MD), she has a Masters degree in Public Health (MPH), and a post-graduate Diploma in Child Development and Health and Welfare Education (DipEd) from the University of London, Institute of Education. Dr Rowe has also held academic appointments in the Department of Paediatrics at the University of Melbourne, consisting of teaching, clinical and research responsibilities, as well as the co-ordination of the undergraduate medical students' training program in paediatrics (1986-1997). Throughout these appointments, Kathy has developed extensive clinical and research experience in the management of children and adolescents with disabilities, behavioral and learning difficulties (including those with ADD and AD/HD), ear, nose and throat problems, as well as those with Chronic Fatigue Syndrome.

Abstract

Psycho-behavioural rating inventories are used routinely by psychologists and psychiatrists as assessment instruments to assist with the evaluation and 'diagnosis' of children and adolescents. They are also used in epidemiological studies to obtain normative/prevalence estimates of children/adolescents with psycho-behavioural 'problems'. Advantages entailed in their use include ease of administration and the convenience of obtaining estimates of normative behaviours from large numbers of informants. However, serious decisions are frequently made on the basis of 'measures' obtained from such instruments, including the labelling of a child as 'pathologic', subsequent referral to intervention therapy services, and prescription of medication by a physician. This workshop highlights key methodological issues endemic to the design and use of psycho-behavioural rating inventories, and the analyses of data derived from them. With specific reference to the assessment of ***inattentive*** behaviours, the workshop provides evidence indicating that traditional psychometric methodologies employed to construct 'scales' (typically from ordinal, item-response formats) and to report 'norms' that ignore the sampling, measurement, distributional and structural properties of the derived data, have long since passed their 'use-by-date'. Also demonstrated is that claims of *validity* and *reliability* employing these traditional methodologies can no longer be justified. Using data obtained from the administration of psycho-behavioural rating inventories in several large-scale research projects, these issues are illustrated and discussed in terms their substantive implications.

The outcomes of more robust methodologies are presented that stress the need to revise the design of child/adolescent psycho-behavioural rating inventories, and point to the adoption of more rigorous approaches to *measurement* and analyses of the related data.

1.0 Introductory comments

Psycho-behavioural rating inventories are used routinely by psychologists and psychiatrists as assessment tools to assist with the evaluation and 'diagnosis' of children and adolescents. They are also used in epidemiological studies to obtain normative/prevalence estimates of children and adolescents with psycho-behavioural 'problems', as well as for estimating effect magnitudes of the overlap between externalizing behaviour problems and educational under-achievement. Advantages entailed in their use include ease of administration and the convenience of obtaining estimates of normative behaviours from large numbers of informants. Nonetheless, serious decisions are often made on the basis of 'measures' obtained from such instruments, including the labelling of a child as 'pathologic', subsequent referral to intervention therapy services, and prescription of medication by a physician – all of which have potential impacts on students' cognitive, affective

[1]Enquiries related to this paper should be directed to: Dr Ken Rowe, Research Director (Learning Processes & Contexts), Australian Council for Educational Research, 19 Prospect Hill Road (Private Bag 55), Camberwell, VIC 3124; Australia; Email: rowek@acer.edu.au; OR to Dr Katherine Rowe, Consulting Physician, Department of General Paediatrics, Royal Children's Hospital, Melbourne, VIC 3052, Australia; Email: kathy.rowe@.rch.org.au.

and social/behavioural progress, especially in educational contexts.

Since behaviour at school and at home affects students' opportunities for learning and development, an enduring concern of teachers, parents and health professionals is the extent to which such maladaptive, externalising behaviours (particularly *inattentiveness*) adversely affect their learning outcomes. Students whose behaviours are regarded as inattentive, disruptive or maladjusted have been shown to be at risk of poor educational attainment.[2] Moreover, in addition to the consequences for an individual, behaviour problems in the classroom diminish educational opportunities for other students and contribute to teacher stress (Barkley & Pfiffner 1995a,b; Hinshaw & Nigg 1994; Brenner, Sörbom & Wallius 1985). Thus, in the context of clinical practice, as well as in psychosocial, epidemiological and educational, research, the measurement of child/student behaviour is of crucial importance.

The measurement of behaviour, however, is problematic. While it is possible to observe and estimate the frequency and saliency of specific behaviours by direct, objective means (see Rowley 1976, 1989), such approaches typically ignore the context in which behaviour takes place and fail to account for the possibility that some behaviours may be appropriate in certain circumstances and at certain stages of socio-behavioural development but inappropriate in others. Systematic observation techniques, particularly in school settings, are time-consuming and not practical options for screening large numbers of students.

In practice, child/student behaviour is assessed most frequently by means of rating inventories or 'checklists' completed by teachers, parents or clinicians (see Figures 2a, 2b and Figure 4). Typically, these multiple item inventories require response ratings in: (a) dichotomous categories (e.g., 'present'/'absent'; coded: '1' and '0', respectively); and/or (b) in Likert-type, ordered, polytomous categories of monotonically-increasing salience or frequency – coded: 0, 1, 2, 3, 4, etc (Likert 1932). Among the best known and widely used inventories in psychosocial, educational and epidemiological research include: the *Achenbach System of Empirically Based Assessment-ASEBA* (Achenbach & Rescorla 2001),[3] *Conners' Rating Scales* (Conners 1969, 1973, 1978, 1990a,b, 1994), *The Children's Attention and Adjustment Survey* (Lambert, Hartsough & Sandoval 1990), the *Rutter B(2) Scale* (Rutter 1967), the *Behaviour Problems Management System* (Galvin & Singleton 1984), and the *Rowe Behavioral Rating Inventories-RBRI* (Rowe & Rowe 1997a,b, 1999).

Although ratings on such instruments are essentially subjective, in the case of teachers and parents, this subjectivity is an asset since raters make 'in-context' judgments about behaviour against normative expectations and experience. Moreover, in addition to their convenience, behavioural rating scales for use by parents and teachers are indispensable, since child behaviours '...are almost always manifest in natural settings such as home and school, but might not be evident in laboratory or clinical environs. Parents' and teachers' judgments regarding the frequency, severity and appropriateness of children's behaviour are therefore essential for accurate detection and diagnosis...' (Edelbrock & Rancurello 1985, p. 429). The importance of teachers' roles in identifying, describing and defining child/student behaviour has long ago been expressed by Bower (1970, p. 94) as follows:

> The myth still exists that someone, somewhere, somehow knows how to assess behavior and/or mental health as positive or negative, good or bad, healthy or non-healthy, independently of the school context in which the individual is living and functioning. I strongly suspect that teachers, by focusing on the child's observable behavior in school, are closer to an operational reality of mental health than one can come up with in a sedentary examination.

2.0 Design problems endemic to typical behavioural rating inventories

Design problems endemic to typical behavioural rating inventories are at least twofold. First, for large-scale educational and epidemiological studies, a key disadvantage is their length. For obvious logistic reasons, inventories of thirty or more items with multiple response categories take considerable time to complete (e.g., Achenbach's CBCL/6-18 has 121 major items, and a further 26 'context/background' items). Completion of such inventories by teachers for all students in a class, for example, can be an arduous task and increase the likelihood of inaccuracies. Moreover, for longitudinal studies designed to investigate change in behaviour over time, it is necessary to use an inventory that is applicable to a wide age range. Inventories that have been designed to identify behaviours for specific age groups are not suitable for such purposes.

[2]For comprehensive reviews of this literature, see: Cantwell and Baker (1991); Elkins and Izard (1992); Hinshaw (1992a,b, 1994); Rowe (1991); Rowe and Rowe (1992a,b,b 1999); Singh, Ollendick and Singh (2002).

[3]The ASEBA comprises: the *Child Behavior Checklist for Ages 6 to 18* (CBCL/6-18), *Youth Self-Report* (YSR) and the *Teacher's Report Form* (TRF).

Second, a major disadvantage of most existing behavioural rating inventories is the use of items that focus exclusively on maladaptive rather than adaptive behaviours (e.g., Achenbach & Rescorla 2001; Conners 1969, 1973, 1994; Quay & Peterson 1975; Rutter 1967). Two examples are given in Figures 2a and 2b.[4] On the one hand this is not surprising given that such instruments are mostly constructed from the 'pathologic' (or negative) nomenclature contained in published manuals of diagnostic criteria for mental and behavioural disorders such as *DSM-II, DSM-III, DSM-III-R, DSM-IV* (APA 1968, 1980, 1987, 1994) and *ICD-9, ICD-10* (WHO 1978, 1992, 1996). In pointing to limitations entailed in the exclusive use of negatively-anchored items typical of most behavioural rating instruments, we have argued elsewhere:

> Emphasis on negative nomenclature is at the expense of a more balanced assessment and increases the risk of prejudicial searches for 'pathology', regardless of its presence or absence (Rowe & Rowe 1992a, p. 350).

Nor are such instruments independent of socio-cultural differences (Yao, Solanto & Wender 1988). For example, in a normative study of Achenbach's CBCL/6-18, Hensley (1988) found a consistent tendency by Australian parents to rate their child's behaviour as 'problematic' – significantly more so than their North American counterparts. Similar findings have been reported in comparative and normative studies of parent and teacher ratings (e.g., Glow 1978; Goyette, Conners & Ulrich 1978; Rowe & Rowe 1993a, 1997c; Verhulst & Akkerhuis 1989).

Item Nos. and Description		Response categories and coding			
Item No.	Item description	Not at all	Just a little	Pretty much	Very much
1	Restless and overactive	0	1	2	3
2	Excitable, impulsive	0	1	2	3
4	Fails to finish things he/she starts	0	1	2	3
5	Constantly fidgeting	0	1	2	3
6	Inattentive, easily distracted	0	1	2	3

Figure 2a Items from the *Inattentive/Overactive* sub-scale of Conners 10-item *Abbreviated Parent–Teacher Questionnaire* – ATPQ (n = 6923; α = 0.840)

Item Nos. and Description		Response categories and coding		
Item No.	Item description	Not true	Somewhat or sometimes true	Very true or often true
4	Fails to finish things he/she starts	0	1	2
8	Can't concentrate, can't pay attention for long	0	1	2
10	Can't sit still, restless, or hyperactive	0	1	2
41	Impulsive or acts without thinking	0	1	2
78	Inattentive or easily distracted	0	1	2

Figure 2b Items from the *Inattention and Hyperactivity-Impulsivity* sub-scale of Achenbach's *Child Behavior Checklist* – CBCL/6–18 (n = 6923; α = 0.777)

Apart from the negatively anchored wording, an interesting feature of the Conners' and Achenbach 5-item scales given in Figures 2a and 2b is the similarity of the constituent item nomenclature. However, the dissimilarity in the response formats – from a 4-category response (Conners' ATPQ) to a 3-category response (Achenbach's CBCL/6-18) – has had a notable effect on reducing the 'reliability' estimate (i.e., from α = 0.840 to = 0.777, respectively).[5]

More than 26 years ago Sandoval (1977) criticised the use of rating scales exclusively employing negatively worded items on the grounds that they are highly susceptible to rater bias and response sets such as 'reverse halo effects' or 'reverse generosity errors'. In a comparative study of format effects in rating scales of 'hyperactivity', Sandoval (1981) subsequently demonstrated that for positively worded items, raters are more willing to use the extreme rating categories for a given item, thus increasing the dispersion and discrimination of the ratings. In contrast, an inspection of the marginal distributions for negatively worded items show that they tend to be

[4]The data from which Cronbach's (1951) α 'reliability' coefficients for these scales have been computed derive from studies reported by Rowe and Rowe (1993c, 1995, 1997c, 1999).

[5]For comparative purposes, but with some reservations, the conventional estimate of 'internal consistency', namely, Cronbach's (1951) *standardised item alpha* (α), is given here. There are two major problems with the use of α: (1) the magnitude of α is a direct function of the number of items in a scale, regardless of their individual and shared error variance, and (2) α estimates of 'reliability' are lower-bound estimates, based on negatively-biased and inappropriate Pearson product-moment correlations among the constituent items – the data from which consist of responses in ordinal categories (see discussion in #3.0 below). For detailed expositions of the limitations of Cronbach's alpha in such circumstances, see McDonald (1981), Miller (1995) and Raykov (1997, 1998). For example, McDonald shows that: 'Proposals to regard coefficient alpha as a coefficient measuring homogeneity, internal consistency, or generalisability, do not appear to be well founded' (1981, p. 100). Similarly, Miller demonstrates '...the failure of α to meet certain basic criteria as an index of test homogeneity' (1995, p. 255).

highly skewed and leptokurtic – as illustrated in Figure 2c below.

Q1	Frequency	Percentage	Bar Chart
0	3875	56.0	
1	2325	33.6	
2	587	8.5	
3	136	2.0	
Q2	**Frequency**	**Percentage**	**Bar Chart**
0	2989	43.2	
1	2941	42.5	
2	803	11.6	
3	190	2.7	
Q4	**Frequency**	**Percentage**	**Bar Chart**
0	3968	57.3	
1	2193	31.7	
2	528	7.6	
3	234	3.4	
Q5	**Frequency**	**Percentage**	**Bar Chart**
0	4392	63.4	
1	1740	25.1	
2	536	7.7	
3	255	3.7	
Q6	**Frequency**	**Percentage**	**Bar Chart**
0	3607	52.1	
1	2519	36.4	
2	584	8.4	
3	213	3.1	

Figure 2c. Univariate distributions for the 5-item *Inattentive/Overactive* sub-scale of Conners' 10-item *Abbreviated Parent-Teacher Questionnaire* – ATPQ: Parent ratings for 6923 children aged 5–16 years

Methodological and data-analytic problems

3.1 Factor analytic (FA) pathologies '

By far the most popular methodological means of defining and 'measuring' emotional and behavioural domains are via *exploratory* factor-analytic (FA) approaches to 'determine' the underlying dimensionality of multiple-item rating inventories administered by parents, teachers or clinicians. From Hinshaw's (1987) comprehensive review of 60 FA studies published between 1970 and 1986, it is interesting to note that all used *exploratory* FA approaches, and that 56 (93%) used orthogonal methods of factor extraction and rotation (mostly principal components analysis or principal factor solutions – both with varimax rotation). Such approaches are problematic on at least three grounds.

First, in the case of exploratory (unrestricted) methods of FA, the solutions are arbitrary, data-driven, hypothesis-generating, and invariably result in theory conflation (Jöreskog 1981; Rowe 1989, 2004). Second, orthogonal methods of factor extraction and rotation assume that the derived factors are uncorrelated or independent – by definition (Harman 1976). Since all items are allowed to load on more than one factor, the resulting correlated error variance alone is sufficient to yield shared variance across factors. Although the construction of uni-dimensional scales is highly desirable from a measurement perspective, given the considerable literature concerning the non-independent and overlapping dimensions of child behaviour and psychopathology, orthogonal FA methods are difficult to justify either

substantively or empirically. At best, such procedures yield discrepant findings that are all-too-frequently ignored or interpreted as 'statistical artifact'. At worst, such procedures yield mis-specified and misleading estimates that contain large proportions of measurement error, with crucial implications for substantive interpretations of findings from subsequent statistical modelling.

Third, and perhaps most serious of all, such methods are invariably applied to item responses in dichotomous or 3 to 5-point Likert-type ordinal categories, and rely on the computation of Pearson product-moment (PP-M) inter-item correlation matrices – estimated by default in most omnibus statistical packages. What is overlooked in such instances is that the assumptions underlying PPM correlations (i.e., normal distribution and homogeneity of variance) are always violated (see: Jöreskog 1994; Rowe 2002, 2004a; Rowe & Rowe 1992a, 1997c, 1999). Indeed, failure to take account of the measurement and distributional properties of response variables in factor analysis, amounts to what Hendrickson and Jones (1987) refer to as 'an undisciplined romp through a correlation matrix' (p. 105). Consistent with the insights of Scarr (1985), we have suggested elsewhere: 'Given the almost universal application of these procedures, it could be argued that current claims to substantive knowledge about dimensions of child psychopathology may be little more than the products of methodological and statistical artifact' (Rowe & Rowe 1992a, p. 351). Whereas there is evidence for awareness of this problem among some researchers in child psychology and psychiatry, it is rare, and warnings about such violations have remained patently unheeded. For example, Morris, Bergan and Fulginity (1991, pp. 373-374) attempted to alert their fellow researchers in the following terms:

> Traditional factor analytic procedures assume that manifest indicators are normally distributed continuous variables. Test items are generally dichotomous or polytomous variables that reflect no more than an ordinal scale. Thus, a normal distribution cannot be assumed. Traditional practice has been to ignore the requirement of continuous normally distributed variables and to factor analyze test items. The result of this approach is biased estimates of model parameters.

A number of approaches are now available that provide ways to carry out confirmatory factor analyses with ordinal data and obtain unbiased estimates of model parameters. Applications of these techniques with clinical assessment instruments are largely lacking. Thus, the state of affairs that exists at present is that little attempt has been made to establish the construct validity of large numbers of clinical assessment instruments that are used with children. … Of particular concern is the issue of the validity of using existing assessment instruments for referral, diagnosis, treatment selection, forensic evaluations, and the evaluation of treatment outcome.

Further, from Jöreskog (1994, p. 383), the special features of ordinal variables are worth noting:

> Observations on an ordinal variable are assumed to represent responses to a set of ordered categories, such as a five-category Likert scale. It is only assumed that a person who responds in one category has more of a characteristic than a person who responds in a lower category. *Ordinal variables are not continuous variables and should not be treated as if they are. Ordinal variables do not have origins or units of measurement. Means, variances, and covariances of ordinal variables have no meaning* (our emphasis).
>
> It is common practice to treat scores 1, 2, 3, 4, representing the ordered categories of an ordinal variable as numbers on an interval scale and use a covariance matrix computed in the usual way to estimate a structural equation model. What is so bad with this is not so much that the distribution is non-normal; more importantly the distribution is not continuous: there are only four distinct values in the distribution. The use ordinal variables in structural equation models (SEM) requires other techniques than those which are used for continuous variables.

It should also be noted that, in general, SEM techniques (including both exploratory and confirmatory factor analysis) assume that the observed data are quantitative variables measured, at least approximately, on an interval scale, and whose distributions are approximately multi-normal. In most psychosocial research applications, however, the observed variables are typically non-normal and/or of mixed response types: categorical, ordinal (Likert-type ratings) and continuous. Under such circumstances, the use of ordinary product-moment correlations is not appropriate (Brown 1989; Healy & Goldstein 1976). Instead, *tetrachoric* (dichotomous with dichotomous) *polychoric* (ordinal with ordinal)[6] and *polyserial* correlations (ordinal with continuous) should be computed, and

[6]Unlike the product-moment correlation which is a measure of association (or standardised co-variation) between the 'scores' for two continuous variables, the polychoric correlation is an estimate of joint variation '..in the latent bivariate normal distribution representing the two ordinal variables' (Jöreskog & Sörbom, 1988, pp. 1–9). For further technical details related to the estimation of polychoric correlations, see Jöreskog (1994), Olsson (1979), Poon and Lee (1987).

the correct asymptotic covariance matrix of such correlations should be analyzed by the method of Weighted Least Squares (WLS), using PRELIS (Jöreskog & Sörbom 2003a), for example. Failure to do otherwise can lead to gross errors in correlation estimates, distorted parameter estimates, and incorrect goodness-of-fit measures and standard errors (Huba & Harlow 1987; Jöreskog & Sörbom 2003b).

Hence, when the data on observed items/indicators are non-normal and non-continuous (e.g., dichotomous, ordinal/polytomous categories), the use of product-moment correlations is inappropriate (Jöreskog 1990, 1994), yielding large negative biases in their estimates (Carroll 1961; Jöreskog & Sörbom 1979, 1988; Lord & Novick 1968). An illustration of the negative bias entailed by the use of PP-M correlation estimates compared with their polychoric counterparts is given in Tables 3a and 3b – using the five items from the *Inattentive/Overactive* sub-scale of Conners' 10-item *Abbreviated Parent–Teacher Questionnaire* (ATPQ) given in Figure 2a. In this case, compared with the polychoric correlations, the PPM correlations are negatively biased by 0.1 (on average).

Table 3a Lower Triangular Matrix of PPM Inter-correlations Among Conners' Inatten/OA Items

Items	Q1	Q3	Q4	Q5	Q6
Q1	1				
Q2	0.621	1			
Q4	0.408	0.360	1		
Q5	0.597	0.484	0.481	1	
Q6	0.466	0.415	0.659	0.546	1

Table 3b Lower Triangular Matrix of Polychoric Inter-correlations Among Conners' Inatten/OA Items

Items	Q1	Q2	Q4	Q5	Q6
Q1	1				
Q2	0.734	1			
Q4	0.497	0.424	1		
Q5	0.697	0.585	0.575	1	
Q6	0.563	0.492	0.769	0.656	1

In brief, as a consequence of the typical inappropriate use of PP-M correlation estimates for dichotomous or ordinal variables, instead of their consistently less biased *tetrachoric* or *polychoric* counterparts, respectively, substantial negative bias (i.e., under-estimates) in the inter-item correlations and subsequent factor parameters is unwittingly introduced.

These moribund approaches, that have long-since passed their 'use-by-date', lead to at least two major problems when modelling relationships among composite scale scores, or to compare the magnitudes of their interdependent effects. First, the unit-weight addition of indicator variables in the formation of the scale scores ignores the possibility that indicators typically contribute differentially to the measurement of composite/scale 'scores'. Second, the unit-weight addition of indicators may invalidate the composite score if one or more of the indicators 'measure' a construct other than the one under consideration. Behavioural rating developers and researchers who continue to use 'data-fishing' methods that fail to account for the measurement, distributional and structural properties of the obtained data (typically consisting of raw, un-weighted response scores on Likert-type item/indicators), run the risk of generating biased and misleading estimates (Hendrickson & Jones 1987; Morris, Bergan & Fulginity 1991; Rowe 2002, 2004a; Rowe & Rowe 1992a,b, 1997c, 1999; Table 3a).

During the past 25 years, these problems have been minimised somewhat by the use of confirmatory factor analysis (see Bentler 1980; Bollen 1989; Jöreskog 1981, 1990; McDonald 1978, 1985; Muthén 1989). The advantages of confirmatory factor analysis (CFA) methods over exploratory factor analysis (EFA) approaches for such purposes are well documented and need not be reiterated here, but for relevant discussions, see Bollen (1989), Gorsuch (1983), Marsh (1987, 1994), Marsh and Grayson (1994), Rowe (1989, 2002, 2004a), Rowe and Rowe (1992a, 1997, 1999), Scott Long (1983), and Stevens (1995). In brief, the advantages include: '...the ability to formulate, define specifically, and test an a *priori* model; the ability to selectively specify or estimate particular model parameters; and the opportunity to directly test and compare the relative goodness of fit of competing models' (Stevens 1995, p. 217). CFA models allow for unequal contributions of indicators towards the measurement of latent variables (e.g., *inattentiveness*) and the models will fit only when the indicator variables associated with any one latent variable are valid indicators of that latent variable. Further, when the number of indicator variables becomes large, parameter estimation and model fit statistics are unstable unless the sample size is also large.

3.2 Scale 'score' 'pathologies'

A further problem in applied research relates to the widespread use of scale 'scores' derived from behavioural rating inventories for the purposes of classification and diagnosis. Typically, scale 'scores' are computed as *factor scores* (from factor analysis), or worse, as simple, unit-weighted, additive indices (or counts) of their indicators, regardless of either the measurement or distributional properties of the constituent indicators, or their relative contribution to the scale 'score'. Illustrations of the distributional characteristics of unit-weighted scale 'scores' from two behavioural rating inventories are provided in Figures 3a and 3b next page.

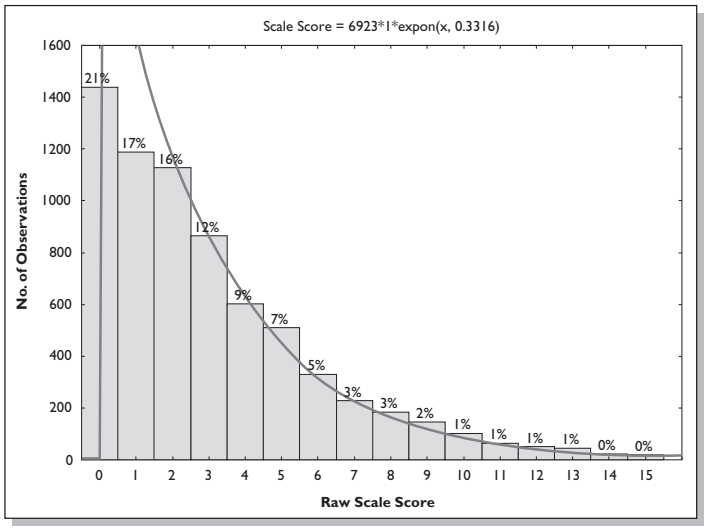

Figure 3a. Distribution of raw scale 'scores' from the five *Inatten/OA* scale items from Conners' *Abbreviated Parent-Teacher Questionnaire* – ATPQ: Parent ratings for 6923 children aged 5-16 years (Min-Max: 0-15)

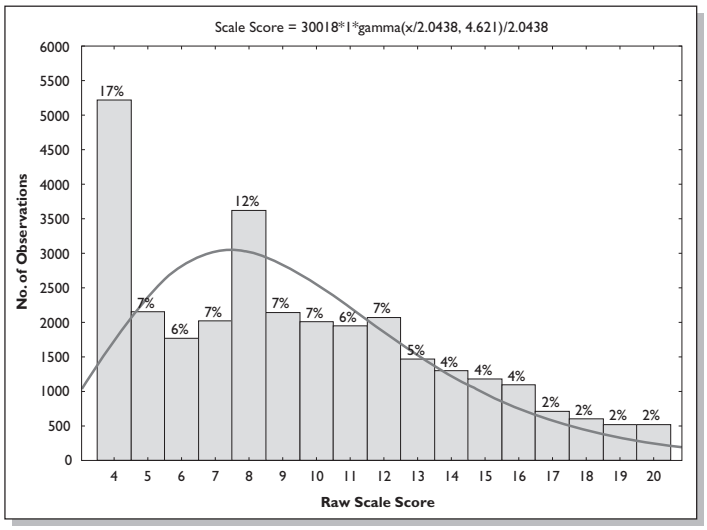

Figure 3b. Distribution of raw scale scores from the four *Attentive-Inattentive* scale items of the *RBRI 12-Item Teacher Form:* Teacher ratings for 30,018 children – aged 5-16 years (Min-Max: 4–20)

From Figures 3a and 3b, it is clear that the distributions of the raw scale 'scores' are non-Normal. That is, the score distribution for Conners' *Inatten/OA* scale (Figure 3a) indicates that the 'best-fit' to the data is described by a negative exponential function, whereas the distribution for the RBRI *Attentive-Inattentive* scale scores (Figure 3b) is best described by a gamma function.[7] All too frequently, these 'scores' are then treated (inappropriately and incorrectly) as Normally-distributed continuous variables in omnibus applications of the general linear model, which further assume that both the constituent item indicators and the computed scale 'scores' are 'measured' without error (Rowe 1989). In such cases, it is well established that the use of standard Normal deviate estimates to describe the distribution of scale 'scores' is misleading (see: Johnson, Kotz & Balarkrishann 1994, 1995; Kendall & Stuart 1963).

Due to the inherent complexity of behavioural disorders in childhood, Ullmann *et al.* (1985) have argued that the common use of a single 'cutoff' score on a rating scale to diagnose deviance is inappropriate and misleading [e.g., a score of 15 on Conners' ATPQ to 'diagnose' *Attention-Deficit/Hyperactivity Disorder* (AD/HD). Although it is customary to select two standard deviations from the mean for these purposes, such selections are arbitrary and can be modified depending on whether one wishes to minimise false positives or false negatives. This approach has been aptly illustrated by Szatmari, Offord and Boyle (1989) in their review of eleven studies reporting prevalence rates of AD/HD. Four of these studies employed diagnostic 'cutoff' scores of 1.0, 1.5 or 2.0 standard deviations from the mean, in the absence of substantive criteria for doing so, 'resulting in the identification of different numbers and types of cases' (Szatmari *et al.* 1989, p. 221). For example, reported prevalence rates for AD/HD vary from less than 1% (Rutter, Tizard & Whitmore 1970), 14.3% (Trites, Dugas, Lynch & Ferguson 1979), to as high as 20% (Shaywitz & Shaywitz 1991), depending on: (1) the methods of data collection, (2) the sampling characteristics of the populations targeted, and (3) the arbitrary determination of deviance criteria.

Further, variability in measurement and 'cutoff' scores, together with sampling differences, lead to substantial

[7]Despite the problems associated with computing simple additive 'scale scores' discussed here, the advantages of employing bi-polar item nomenclature formats (as used in the RBRIs) is evident from Figure 3b – especially in terms of discrimination. Note that the special design features of the RBRI are outlined in #4.0.

differences in prevalence estimates. In the context of predictive or explanatory research, there is little rational justification for identifying, *a priori*, a fixed proportion of the child population as 'AD/HD', for example, particularly when such a dimension is more meaningfully viewed as a continuum, both in quantitative and qualitative terms. Despite the utility and obvious convenience of rating scales, especially for large-scale survey research, the psychometric limitations endemic to their common design, construction and use seem to be largely unrecognised by most developers, users and researchers.

In sum, 'cut-off' scores based on commonly used statistical criteria (i.e., + 1 ≤ SD ≤ + 2) are arbitrary since they are dependent on the properties of the 'measures' used, as well as on sampling variability across studies. Such arbitrariness leads to substantial differences in prevalence estimates that may or may not reflect actual problems. Moreover, given that all behavioural 'measures' computed in this way are highly skewed (as illustrated in Figure 3), statistical criteria of these kind are difficult to justify due to the inevitable violation of the assumptions of *normality* of distribution and *homogeneity* of variance.

3.3 Measurement and scale construction problems

It is important to stress that the foundation of ALL responsible data analysis and statistical modelling is ***good measurement*** and the ***minimization of measurement error variance***– otherwise, what is generated are serious 'garbage-in' 'garbage-out' problems that unjustifiably conflate theory and yield misestimated parameters (at best) and misleading 'findings' (at worst). This is especially the case for analyses of data obtained from behavioural rating inventories (Rowe & Rowe 1992a, 1997c, 1999), as well as for agencies and/or health professionals wishing to use data to identify performance indicators of 'health' or 'pathology', particularly for intervention and policy purposes (see: Rowe 2001, 2004b; Rowe & Lievesley 2002). It should also be noted that measurement error problems are **seriously compounded** with 'contextual' or 'compositional' variables that are aggregated from the characteristics of level-1 units (i, e.g., students) within level-2 units (j, e.g., classes or schools), because the measurement error inherent in the level-1 variables is averaged across the level 1 units in each level-2 unit, or higher (see Rowe 2004b). Moreover, there is additional sampling error whenever $nj < Ni$ – which is always the case.[8]

It is now well established that factor-analytic (FA) and Classical Test Theory (CTT) approaches to measurement and scale construction in psychosocial inquiry do not even meet the three **basic 'requirements'** of ***measurement***, namely: (1) the need to focus on only ***one way*** in which objects or persons differ in terms of an attribute of interest; (2) the need for a ***unit of measurement*** (so that equal numerical differences represent equal amounts); and (3) ***objectivity*** (freedom from the characteristics of the instrument and of the person(s) undertaking the measurement).[9] Further, it has been demonstrated that FA and CTT approaches are **not** commensurate with modern measurement theory and practice (see especially: Embretson 1996; Embretson & Hershberger 1999; Masters & Keeves 1999; Wilson & Engelhard 2000; Wright 1999). Key reasons for this are beyond the scope of this paper. Nevertheless, in brief, scale score meaning via CTT and FA approaches is merely inferred from norm-referenced 'standards'. That is, the scores *per se* have no meaning for what an assessed individual does or can do; moreover, such scores are *sample-dependent*.

By contrast, from item-response approaches to measurement (better known as *Item Response Theory* models – IRT), the scale scores are *sample-independent* and score meaning can be referenced directly to the constituent items – from which a linear scale can be constructed and described qualitatively (e.g., Masters 2001a,b; Masters, Meiers & Rowe 2003; Stephanou 2000). Following the seminal work of Thorndike (1904), Thurstone (1926) and Guttman (1944), the 'requirements' of objective measurement in the psychosocial sciences have been promulgated by the Danish mathematician Georg Rasch (1960), who laid the foundations of what has become known as *modern measurement theory*, or *Rasch measurement*. The advantages of this approach to measurement are noted in more detail later in #5.0.

4.0 Improving the design of psycho-behavioural rating inventories

Due mainly due to the poor design, low reliability and lack of predictive validity

[8]Note that Fuller (1987) provides a comprehensive account of methods for dealing with measurement errors in linear models, and Goldstein (1995, chp. 10) extends some of those procedures to the multilevel modeling case.

[9]For comprehensive treatments and applications of modern measurement theory (including Rasch measurement), see: Embretson and Hershberger (1999), Masters (1982, 1988), Masters and Keeves (1999), Masters and Wright (1997), Rasch (1960, 1977), Stephanou (2000, 2002), Wilson & Engelhard (2000), Wright (1999), Wright and Mok (2000). For excellent introductory overviews, see Masters (2001a,b). For an application to psycho-behavioral rating inventories, see Smith and Johnson (2000).

of the Conners' and Achenbach's negatively-worded types (as illustrated earlier in Figures 2a and 2b), the *Rowe Behavioral Rating Inventories* (RBRIs)[10] were developed from empirical research applications to obtain valid and reliable 'in-context' measures of child/student *externalising* behaviours for use in clinical settings, as well as in educational, psycho-behavioural and epidemiological research. Since the rationale for the development and use of the RBRIs has been comprehensively documented and demonstrated by Rowe (1991, 1997a) and by Rowe and Rowe (1992b,c, 1993c, 1994b, 1995, 1997c, 1999), the need for reiteration here is not required. However, for illustrative purposes, Figure 4 records the constituent 4 items of the *Attentive-inattentive* scale from the RBRI *12-item Teacher Form*.

Three features of the design and item content of the *RBRIs* should be noted. First, following the semantic bipolar format advocated and used by Kysel, Varlaam, Stoll and Sammons (1983), the RBRI items allow for assessments both adaptive and maladaptive behaviours (i.e., *health* and *pathology*). Second, the item nomenclature has been formulated on the basis of extensive cross-validations of parent and teacher descriptions of typical child/student externalizing behaviours at home and at school, and in three domains: *sociable-irritable/antisocial, attentive-inattentive*, and *settled-restless*. Third, the items are applicable to a wide age range, having been developed from comprehensive trialing and application among large samples of children/students (n > 180,000) in the age range of 5 to 16 years.

Teachers, for each of the following paired behavioral statements, please mark a cross over the dot (e.g., ⊗) which is **nearest** the statement that **best** describes the TYPICAL behavior of THIS student at school

1. Cannot concentrate on any particular task; easily distracted	O O O O O	Can concentrate on any task; not easily distracted
2. Perseveres in the face of difficult or challenging tasks	O O O O O	Lacks perseverance; is impatient with difficult or challenging tasks
7. Persistent, sustained attention span	O O O O O	Easily frustrated; short attention span
10. Purposeful activity	O O O O O	Aimless; impulsive activity

Figure 4. *Attentive-inattentive* items from the *RBRI 12-item Teacher Form* (n = 30,018; α = 0.926)

How the *RBRI* forms should be scored depends on the purposes for which they are to be used, but a major advantage of the bipolar item format is that alternative methods for item scoring may be used.[11] That is, in studies concerned with the measurement of *maladaptive* behaviours, Item Nos. 2, 7 and 10 shown in Figure 1c may be scored 1 to 5 (from left to right) on the five-point ordinal scale, with scoring reversed for Item No. 1. In such cases, a low score on each item reflects positive adjustment and a high score, poor adjustment. In studies concerned with the effects of *adaptive* behaviours, the items may be scored such that high scores are reflective of positive adjustment.

5.0 Improving the measurement properties of behavioural rating inventories

At this point, a brief discussion of the utility of fitting behavioural rating data to item-response measurement models that meet the basic requirements of *objective measurement* is helpful. In particular, what is highlighted here is the utility of *Rasch measurement* in constructing scales by calibrating item indicators with dichotomous and/or polytomous response categories – typical of behavioural rating inventories. For relevant work in this area, see references cited in footnote 9.

[10]The RBRI inventories consist of two major rating forms: (1) a 16-item Teacher Form, and (2) a 20-item Parent Form, for use in clinical settings with children in the age range of 5 to 16 years. Both forms are supported by an accompanying interactive computer software package, RBRI Profile® (Rowe & Rowe, 1997a) and a User's Manual (Rowe & Rowe, 1997b). Similar information is provided for two shorter versions of these forms, namely the 12-item Teacher Form and the 16-item Parent Form – devised specifically for use in the large-scale monitoring and epidemiological research.

To date, the research applications include epidemiological studies of the relationship between the ingestion of synthetic food dyes and behavioral change in pediatric populations (Rowe KS, 1988, 1996; Rowe KS & Briggs, 1992, 1993; Rowe & Rowe, 1994a,b), and in studies of factors affecting student literacy and numeracy achievement (Crévola & Hill, 1998a; Hill et al. 1993, 1996; Hill & Rowe, 1996, 1998; Rowe, 1991, 1997; Rowe, Fullarton et al., 2003; Rowe & Hill, 1998; Rowe & Rowe 1992b,c, 1993, 1995, 1997c, 1999). In these studies, the inventories have been validated for dye-challenge and for monitoring the comorbidity of externalizing behaviors and academic under-achievement.

The psychometric and normative properties of the *RBRIs* are based on cross-validated and replicated samples of teacher ratings for 33,433 school-aged children in five age cohorts (5–6, 7–8, 9–11, 12–13, 14–16 years) and parent ratings on 16,569 children across the same age cohorts. Data on concurrent parent and teacher ratings have been obtained for 9566 children. Specific details of the samples, data properties and related research applications are available in the RBRI User's Manual (Rowe & Rowe, 1997b) and in Rowe and Rowe (1999).

[11]A further advantage of employing a bipolar item format is that it minimises the occurrence of 'negative halo effects' by minimising the risk of prejudicial searches for 'pathology'.

The work of Rasch and those who have followed has impacted radically on the *theory* of *measurement*, and especially on applications in educational and psychological assessment (psychometrics). In brief, the Rasch approach to the *measurement* of a *latent* or *composite* variable – derived from responses to multiple items/indicators in dichotomous or polytomous categories – is that it allows for *scale construction* by calibrating jointly the location of each item **and** respondent on an empirical scale of increasing attribute (e.g. *performance, extroversion, attentiveness, attitude*, etc.). Fitting indicator-response data to Rasch's logistic model yields an unbounded **logit scale**[12] (with interval properties) that allows any pair of items (and person pairs) to be compared in terms of the magnitude of the interval difference between their locations on the scale. An illustration of this feature is the 'Person-item map' provided in Figure 5 [print-out from *ACER-QUEST*, Adams and Khoo (1999)] that not only facilitates the setting of 'cut-scores', or 'pass marks' on assessments, but also, 'benchmarks' and/or performance standards, for example.

A particular advantage of Rasch-calibrated scales is that empirical, evidence-based evaluations can be made of the extent to which each item or indicator contributes to the measurement of the latent variable being constructed (i.e., *differential item/indicator functioning* in terms of measurement accuracy). A further advantage is that a scale so constructed allows detailed **descriptions** of performance levels or standards to be made in both quantitative and qualitative terms (e.g., Masters 2001a,b; Masters, Meiers & Rowe 2003; Stephanou 2000, 2002). The properties of Rasch-calibrated scales are such that items from separate assessment sources/occasions of the same kind (e.g., performance standards) can be ***equated*** and located on a common measurement scale – provided that some indicators and/or respondents (cases) overlap, or are linked from one assessment to another. These procedures are known as *common-item equating* and *common-case equating*, respectively.

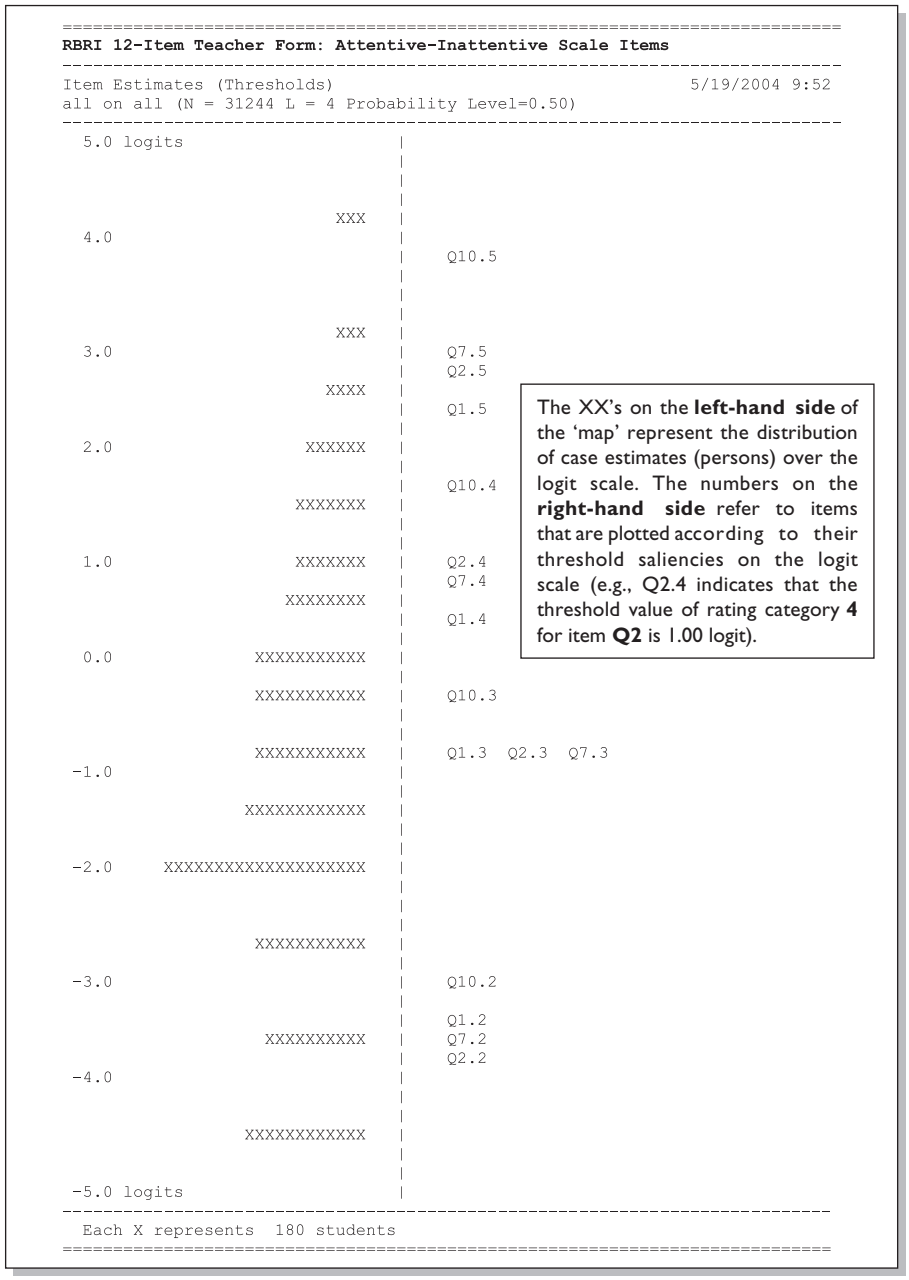

Figure 5. 'Person–item map' of *Attentive–Inattentive* scale items from the RBRI *12-item Teacher Form* for 30,018 children – aged 5–16 years

[12]The *logit* is a unit of measurement derived from the natural logarithm of the odds of an event, where the odds of that event is defined as the ratio of the probability that the event will occur to the probability that the event will not occur. A logit scale is used in both educational and psychological assessment because it has interval scale properties. That is, if the 'difficulty' or 'salience' of an assessment item (e.g., Item A) is 1.0 logit greater than the difficulty or salience of Item B, then the odds of an individual responding correctly (or more saliently) to Item B are 2.7 times the odds of the same individual responding correctly (or more saliently) to Item A, regardless of whether this person has high or low ability/attribute. Similarly, if the ability or attribute of Person A is 1.0 logit greater than the ability of Person B, then the odds of Person A responding correctly (or more saliently) to an item are 2.7 times the odds of Person B responding correctly (or more saliently) to the same item, regardless of item difficulty or its salience.

These properties of scales constructed via *Rasch measurement* are ***especially*** useful in the development of item banks from which items and/or indicators of known attribute salience can be drawn to develop further assessment instruments that are comparable. It is also extremely valuable (and **vital**) for applications in: (1) longitudinal, repeated-measures studies of the same cases, and (2) cross-sectional studies of different respondent cohorts at different times. Such procedures are not possible using traditional Classical Test Theory (CTT) methods, and have considerable advantages over traditional methods based in CTT – particularly those employing factor analytic approaches. For these reasons, *Rasch measurement* is used as the basis for constructing and describing scales for all cognitive, affective and behavioural assessment instruments developed by the Australian Council for Educational Research (ACER), as key elements in its national and international work in assessment and reporting.[13]

6.0 Concluding comments

In highlighting key methodological problems endemic to the design and use of psycho-behavioural rating inventories, and analyses of data derived from them, the purpose of the present paper is twofold.

First, it is argued that the design features of most psycho-behavioural rating inventories used routinely by epidemiologists, psychiatrists and psychologists to assess children and adolescents with psycho-behavioural 'problems' are less than adequate. In particular, the almost exclusive use of negative item nomenclature in such inventories increases the risk of prejudicial searches for 'pathology', regardless of its presence or absence. Given that serious decisions are frequently made on the basis of 'measures' obtained from such instruments, including: the labelling of a child as 'pathologic', subsequent referral to intervention therapy services, and prescription of medication by a physician, it is crucial that such instruments be of the highest quality in terms of both their design and measurement properties.

Second, on the basis of supporting evidence the paper argues that traditional *Classical Test Theory* and *factor-analytic* methodologies employed to – construct 'scales', 'measure' behaviour, report 'norms' and to specify 'cut-off' scores for the purposes of 'classification', 'diagnosis' and the provision of prevalence estimates – have long since passed their 'use-by-date'. Indeed, it is argued that claims of *validity* and *reliability* employing these traditional methodologies can no longer be justified. Rather, the need to adopt more rigorous approaches to *measurement* and analyses of the related data is urgent. It is hoped that both the traditional 'emperors' of psycho-behavioural inventory design, development and data-analytic methodology, and we, the product users, will heed such cries about our 'nakedness' before our sartorial delusions render our efforts ludicrous.

References

Note: due to space limitations here, the entire document - including the references cited in this paper - are available for download in pdf format from www.acer.edu.au/research/programs/learningprocess.html

[13] A full listing of psycho-behavioral and educational assessment instruments developed by ACER via *Rasch measurement*, are available at: www.acer.edu.au/tests/index.html. For examples of ACER's national and international work in assessment and reporting (among many others), see: Adams *et al.* (1988-1997); Lokan, Greenwood and Cresswell (2001); Masters (2002); Masters and Forster (1996, 1997); Rowe and Stephanou (2003). For examples of ACER's international publications and work programs with OECD, IEA and the World Bank, visit: www.acer.edu.au/about/international.html. Specific ACER projects and publications related to assessment and reporting are available at: www.acer.edu.au/research/reports.html, and www.acer.edu.au/research/programs/assessment.html.