The measurement of composite variables from multiple indicators: Applications in Quality Assurance and Accreditation Systems – Childcare

Ken Rowe, PhD

Australian Council for Educational Research

Background paper prepared for the National Childcare Accreditation Council
August 2006[1]

Introduction

Issues related to the ***measurement***, ***distributional*** and ***structural*** properties of data typically obtained in applied educational, epidemiological and psychosocial research are not well understood. Nonetheless, these issues are neither minor nor benign; indeed they are **vital**. All too frequently, assumptions about the measurement and distributional properties of 'scale scores' (or composite, latent variables) derived from multiple responses to attitude surveys, or to multiple indicators in quality assurance and accreditation rating inventories (e.g., NCAC, 2005a,b; NSW, 1999), are rarely questioned or examined. Indeed, when analyzing such data, investigators almost universally report findings from either univariate and/or multivariate applications of the *general linear model* (GLM) which, for claims of *valid* inference, assume that such 'scores' or 'ratings' (however computed) are continuous variables that are Normally distributed with homogeneous variances. The fact that these strong and crucial assumptions are mostly not justified needs to be emphasized, since the analysis of non-Normal and non-continuous data via omnibus univariate and multivariate GLM techniques, has well known implications for: (a) parameter estimation and interpretation, (b) the calculation of composite variables and their reliability/validity indices, (c) the construction of indices from performance indicator data, and (d) the interpretation of findings from explanatory modeling (Goldstein, 1995; Mardia, Kent & Bibby, 1979; McDonald, 1996; Rowe, 1989, 1998, 2001, 2002a,b, 2004, 2006a; Rowe & Rowe, 1992a,b, 1997, 1999, 2004).

This paper provides guidelines and procedures that are designed to ensure that the *measurement* and *distributional* characteristics of data are examined carefully **prior** to computing composite variables and fitting explanatory models. Guidelines and procedures designed to account for the *structural* properties of data are beyond the scope of this introductory paper, but for useful accounts, see: Bryk and Raudenbush (1992), Goldstein (1995), Kreft and de Leeuw (1998), Rowe (2002a,b, 2003, 2006a), Rowe and Hill (1998). However, before presenting these guidelines, several points that constitute vital philosophical bases for the procedures that follow are worthy of note:

- 'Good' research is ***entirely*** dependent on quality ***theoretical*** and ***conceptual*** foundations. Any research enterprise that is not grounded in substantive, theoretical questions is vacuous! Atheoretical research invariably leads to the conflation of 'theory' via the generation of inductively-derived 'findings' that are often spurious and invalid. Note that the axiom *"there is nothing more practical than a good theory"* is both a necessary and a sufficient condition of 'good' research, including traditional experimental and *ex post facto* designs, as well as to more recent empirical qualitative/quantitative approaches (see: Masters & Keeves, 1998; Keeves & Lakomski, 1998).

[1] This paper is an updated version of Rowe (2002a). Related correspondence should be addressed to Dr Ken Rowe, Research Director, *Learning Processes and Contexts* research program, Australian Council for Educational Research, Private Bag 55, Camberwell, Victoria 3124, Australia; *Tel*: +61 3 9835 7489; *Email*: rowek@acer.edu.au.

- The quality of any statistical analysis or modeling procedure is crucially dependent on: (1) the quality of the data to be analyzed or modeled, and (2) the extent to which the procedure is consistent with underlying substantive research questions.
- The foundation of all responsible data analysis and related statistical modeling is ***good measurement***!

Objective measurement: The Rasch approach[2]

In 1960, the Danish mathematician Georg Rasch laid the foundations of what has become known as *modern measurement theory* (or *item-response theory* - IRT). 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. *achievement*, *extroversion, alienation*, *benchmark performance*, etc.). Fitting Rasch's logistic model to indicator-response data yields an unbounded **logit scale** (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. This feature not only facilitates the setting of 'cut-scores', or 'pass marks' on tests and examinations, but also, for example, 'benchmarks' and/or performance standards in quality assurance and accreditation systems.

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 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; Stephanou, 2000). Moreover, 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.

This property of Rasch-calibrated scales is ***especially*** useful in the development of item banks from which items and/or indicators of known difficulty or attribute salience can be drawn to construct further assessment instruments that are comparable. It is also extremely valuable (and indeed **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. In fact, Rasch measurement has considerable advantages over traditional methods based in CTT, and those using exploratory factor analysis.

Despite these advantages, it is important to note that there are three key assumptions (or requirements) underling the Rasch measurement model, namely:

1. *Logistic model adequacy* (i.e., the logistic model is adequate for response-item/ indicator calibration to construct the measurement scale);
2. *Local independence* requires that the conditional distributions of the item scores for respondents at a given ability/salience level θ, are *independent* (i.e., the inter-item correlation for a selected 'ability' group is not significantly different from zero), and

[2] For comprehensive treatments and applications of Rasch measurement, see: Rasch (1960, 1977), Masters (1982), Masters and Keeves (1999), Masters and Wright (1997), Stephanou (2000), Wright and Mok (2000). For excellent introductory overviews, see Masters (2001a,b).

3. *Unidimensionality* (i.e., the indicators relate to a unidimensional construct (or quality standard) of interest.

Requirements 2 and 3 above are particularly stringent ones that only strictly apply in the case of single attribute scale construction. In the case of response indicators in multi-dimensional quality assurance and accreditation assessment instruments, for example, the covariance approach to constructing measurement scales described below is more appropriate. Nonetheless, this is a debatable issue, about which adherents to Rasch measurement would argue strongly is not the case. However, a demonstration of the construction of Rasch-calibrated measurement scales is beyond the scope and purpose of the present paper. For relevant examples, see Masters (2001b) and Stephanou (2000).

The measurement of composite variables: The covariance approach

Most theories and models in applied psychosocial research are formulated in terms of hypothetical constructs (or *latent* variables) that are not directly measurable or observable. As a means of data reduction, it is common-place to compute latent or composite variables such as *quality assurance domain, accreditation standard* and so on, from several observed indicators (or response items), each requiring responses in dichotomous (i.e., 0-1) or in Likert-type ordered categories (i.e., 1, 2, 3, 4, etc.). Traditionally, and mostly using negatively-biased Pearson product-moment inter-item correlation estimates, such composite scores have been computed as *factor scores*, or as simple, unit-weighted, additive indices of their indicators, regardless of either the measurement and distributional properties of the constituent indicators, or their relative contribution to the composite score. Typically, the composite scores are then treated as continuous variables in omnibus general linear model techniques, which assume that such variables are measured without error.

This approach leads to at least two major problems when trying to model relationships among composite scores, or to compare their magnitudes. First, the unit-weight addition of indicator variables in the formation of the scale scores ignores the possibility that some indicators may contribute more to the measurement of the composite than others. 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. However, there are further problems related to the estimation of composite scores, and are especially evident in prevailing methods used to compute them for use in psychosocial research, as well as in quality assurance and accreditation systems.

In psychosocial inquiry, for example, findings from several studies concerned with the development of psycho-behavioral rating inventories indicate that traditional data analytic methodologies employed to construct ‘scales’ and compute composite scores from ordinal-scaled response data can no longer be justified. Inventory developers and researchers who continue to use exploratory, ‘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 ordinal-scaled indicators), run the risk of generating biased and misleading estimates (see Hendrickson & Jones, 1987; Morris, Bergan & Fulginity, 1991; Rowe, 2002a,b, Rowe & Rowe, 1992a,b, 1997, 1999, 2006a; Scarr, 1985). Indeed, the widespread failure of researchers to take account of such properties in their data, amounts to what Hendrickson and Jones (1987) refer to as “…undisciplined romps through correlation matrices” (p. 105).

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 modeling. Consistent with the insights of Scarr (1985), it has been 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).[3] Whereas there is evidence for awareness of this problem among some researchers in child psychology and psychiatry, it is rare in other areas of applied psychosocial research, and warnings about such violations have remained patently unheeded. For example, Morris *et al*. (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.

Thus, when observed indicators are non-normal and non-continuous (e.g., dichotomous, ordinal/polytomous categories), the use of product-moment correlations is inappropriate (Jöreskog, 1990), yielding large negative biases in their estimates (Carroll, 1961). It should be noted that, in general, SEM techniques (including 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 scale types: categorical, ordinal (Likert-type rating scales) and continuous. Under such circumstances, the use of ordinary product-moment correlations is not appropriate (see Healy & Goldstein, 1976). Instead, ***tetrachoric*** (dichotomous with dichotomous) ***polychoric*** (ordinal with ordinal)[4] and ***polyserial*** correlations (continuous with ordinal) should be computed, and the correct asymptotic covariance matrix of such correlations should be analyzed by the method of Weighted Least Squares (WLS), using PRELIS (Jöreskog & Sörbom, 2005a). 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, 1979, 1988, 1989, 1993). 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 measurements. Means, variances, and covariances of ordinal variables have no meaning* (my 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 of ordinal variables in structural equation models requires other techniques than those which are used for continuous variables.

3 Similar comments can be made about such applications in quality assurance accreditation systems.

4 Unlike the product-moment correlation which is a measure of association (or standardized 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. For further technical details related to the estimation of polychoric correlations, see Jöreskog (1994), Olsson (1979), Poon and Lee (1987).

Cuttance (1987, pp. 245-250), Jöreskog (1990, 1994), and Muthén (1993) have provided detailed discussions on the treatment of dichotomous, ordinal and non-normal variables in structural equation models. In developing the WLS method available in LISREL 7 and LISREL 8 to assist in minimizing problems with non-normally distributed variables, Jöreskog and Sörbom provide a method for obtaining an appropriate weight matrix, correct parameter estimates, standard errors and a fit statistic. "The weight matrix required for such an analysis is the inverse of the estimated asymptotic covariance matrix **W** of the polychoric and polyserial correlations" (Jöreskog & Sörbom, 1993, p. 45).

During the past 30 years, these problems have been minimized by the use of confirmatory factor analysis (see Bentler, 1980; Bollen, 1989; Jöreskog, 1979; McDonald, 1978, 1985; Muthén, 1988). 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, 2002a,b, 2004, 2006a), Rowe and Rowe (1992a, 1997, 1999), Scott Long (1983), and Stevens (1995). In brief, however, 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, and the models will fit only when the indicator variables associated with any one latent variable are valid indicators of that trait. However, when the number of indicator variables becomes large, parameter estimation and model fit statistics are unstable unless the sample size is also large.

An Example: Measuring a *Quality Area* in a Childcare Accreditation System

For illustrative purposes, the following example focuses on one *Quality Area* of the *Quality Improvement Accreditation System* (QIAS) survey instruments for Long Daycare Centres, employed by the National Childcare Accreditation Council of Australia (see NCAC, 2005a,b). The 7 *Quality Areas* and the 33 constituent *Principles* are given in Appendix A.

For each of the six 'Principles' (indicators) relevant to ***Quality Area 4***: ***Children's Experiences and Learning***, raters are required to provide a response in one of four ordered categories, labeled: *Unsatisfactory*, *Satisfactory*, *Good Quality* and *High Quality*, and coded 1, 2, 3, 4, respectively – as shown in Table 1 below.

Table 1. Principles and rating categories for NCAC's QIAS *Quality Area 4*: *Children's Experiences and Learning*

Indicators (Principles)	Unsatis-factory	Satis-factory	Good Quality	High Quality
4.1 Staff encourage each child to make choices and participate in play	1	2	3	4
4.2 Staff promote each child's ability to develop and maintain relationships	1	2	3	4
4.3 Staff promote each child's language and literacy abilities	1	2	3	4
4.4 Staff promote each child's problem solving and mathematical abilities	1	2	3	4
4.5 Staff promote each child's enjoyment of and participation in the expressive arts	1	2	3	4
4.6 Staff promote each child's physical abilities	1	2	3	4

Using raw data[5] on these six 'Principles' (indicators), this example illustrates the specification and use of confirmatory, one-factor, congeneric measurement models to maximize the reliability of a composite score for ***Quality Area 4*** (in this case). The aim is to **maximize** the measurement **accuracy** of the constructed composite 'score' for ***Quality Area 4*** (as well as the six other *Quality Areas*) for subsequent application in the context of a quality assurance and accreditation system (i.e., to **minimize measurement error**).[6]

Construct measurement and reliability

For demonstration purposes, confirmatory factor analysis (CFA) is employed to compute the composite score for ***Quality Area 4*** using LISREL 8.72 (Jöreskog & Sörbom, 2005b) – preferably under a Robust Maximum Likelihood, weighted least squares method of estimation. This is obtained from fitting a one-factor, congeneric measurement model (Jöreskog, 1971) to the constituent ordinal ratings on the six indicators, based on a scaled covariance matrix of the polychoric correlations using PRELIS 2.72 (Jöreskog & Sörbom, 2005a). Composite scores computed by this method are single indices of their component items, each of which is weighted for its relative contribution to the composite. Unlike traditional unit-weighted methods for computing composites, the use of factor score regression indices obtained from CFA one-factor models minimizes measurement error in the indicators contributing to each composite scale, thus increasing the reliability (and validity) of the computed composite scores. For accreditation system applications, the use of maximally reliable composite scores is crucial (see: Holmes-Smith, 1994; Rowe, 2002c; Rowe & Darkin, 2001).

The one-factor, congeneric measurement model may be specified in matrix format – showing the regression of a vector of indicators ($\boldsymbol{x}_i$) on ξ_1 (the latent, composite variable) where the elements λ_{xi} are the partial regression coefficients of ξ_1 in the regression of $\boldsymbol{x}_i$ on ξ_1, and δ_i is a vector of measurement errors in $\boldsymbol{x}_i$, namely:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \\ \lambda_{41} \\ \lambda_{51} \\ \lambda_{61} \end{bmatrix} \begin{bmatrix} \xi_1 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix}$$

or, $$\boldsymbol{x}_i = \lambda_{xi}\xi_1 + \delta_i \qquad [1.1]$$

The assumed model implies that the covariance matrix of the observed indicators ($\boldsymbol{x}_i$) is of the form:

$$\Sigma = \Lambda_{xi}\Lambda_{xi}' + \Theta_\delta \qquad [1.2]$$

where Θ_δ is a diagonal matrix with elements $\theta_{\delta i}$ indicating the unique, measurement error variances of δ_i (i = 1, 2, 3, …, n).

The congeneric measurement model may also be illustrated diagrammatically for the ***Quality Area 4*** composite, as shown in Figure 1.1.

From the parameters of equation [1.2] the reliability (r_c) of a composite (ξ_c) is given as

---

5 These data derive from 12,036 raters for 33 *Principles* of NCAC's *Quality Improvement Accreditation System* (QIAS), obtained from 223 Long Daycare Centres, made up of ratings (with complete data) from: 223 *Self Study Reports*; 215 *Director Reports*, 2115 *Staff Reports*, and 214 *Validator Reports*. For specific details related to these data, see Rowe (2006b).

6 For an earlier overview of NCAC's QIAS, see: Rowe, Tainton and Taylor (2004).

$$r_C = \frac{w_C^{'}\left(\hat{\Sigma} - \hat{\Theta}_\delta\right)w_C}{w_C^{'}\,\hat{\Sigma}\,w_C}, \qquad [1.3]$$

where $\hat{\Sigma}$ is the estimated variance-covariance among the factor 'loadings' for the vector of congeneric indicators (***x***'s), Θ_δ is a vector of unique (measurement error) variances among those indicators, and w_C is a vector of factor score (FS) regression weights that maximize the reliability of the composite.[7] For specific details of these well-established but all too rarely used procedures, see: Alwin and Jackson (1980); Brown (1989); Fleishman and Benson (1987); Jöreskog (1971); Munck (1979); Werts *et al*. (1978). Further details including the rationale for this approach to computing composite variables and their reliabilities have more recently been outlined and demonstrated by: Ainley, Fleming & Rowe (2002); Hill *et al*. (1996); Holmes-Smith and Rowe (1994); Rowe (1998, 2002a,b); Rowe and Hill (1998); and by Rowe and Rowe (1997, 1999, 2004).

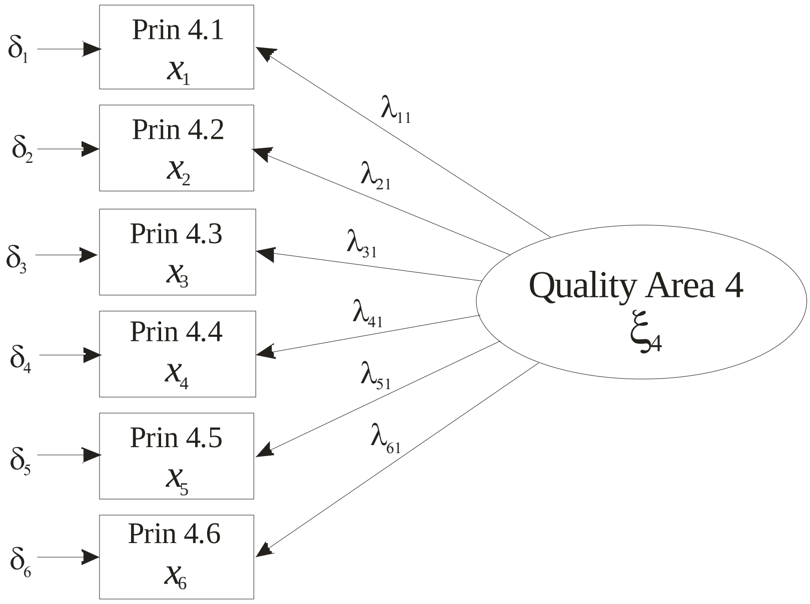

Figure 1.1 One factor, congeneric measurement model for the 6 *Principles* related to *Quality Area 4*

Hence, using the inter-item polychoric correlation matrix and the asymptotic variance-covariance matrix of these polychoric correlations [under a Robust Maximum Likelihood and Weighted Least Squares (WLS) method of parameter estimation], the results of fitting the above congeneric model to the data for the ***Quality Area 4*** composite scale are recorded in Figure 1.2 following.

[7] For a one-factor congeneric measurement model, the factor score regression coefficients (**FS**) represent the estimated bivariate regression of the factor (ξ) on all the observed indicator variables, given by: $\mathbf{FS} = \Lambda'\Sigma^{-1}$, where Λ is the estimated factor pattern matrix and Σ is the estimated covariance matrix of the observed indicators (see Jöreskog & Sörbom, 1989, p. 93; Lawley & Maxwell, 1971, p. 109). 'Factor' or composite score estimates (ξ_i) may be computed for any individual i with observed scores $\mathbf{x}_i$, using the simple product: $\xi_i = \mathbf{FS}\,\mathbf{x}_i$.

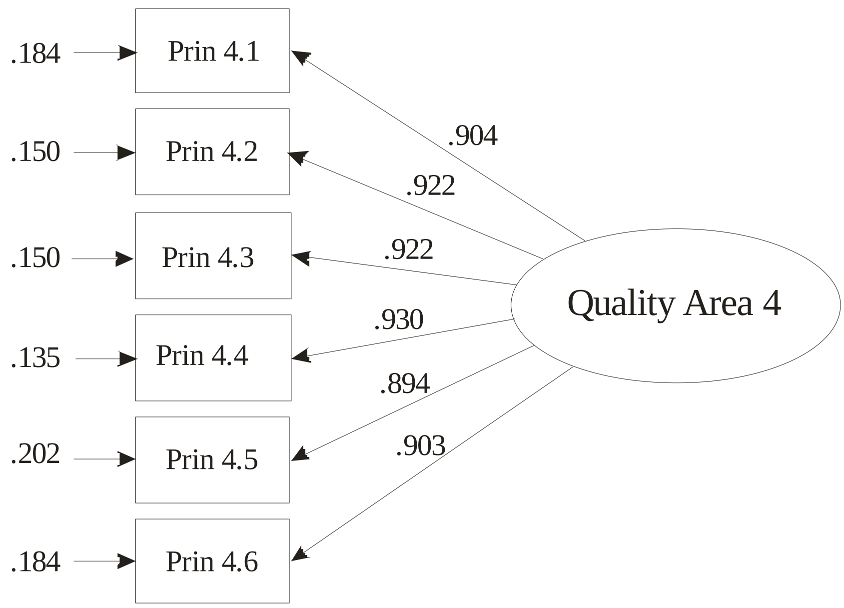

Figure 1.2 Standardized solution to one-factor model for *Quality Area 4*

Model Goodness-of-fit Indices: RMSEA = 0.018; Normed Fit Index (NFI) = 0.999; Non-Normed Fit Index (NNFI) = 0.999; Comparative Fit Index (CFI) = 0.999; Incremental Fit Index (IFI) = 0.999; Relative Fit Index (RFI) = 0.998.

By **every** criterion this solution is an **excellent** fit to the rating data obtained for the six *Principles*, with the model accounting for > 99% of the relative variances and co-variances in the data. These results (*inter alia*) confirm: (a) both the utility and stability of this *Quality Area* of the QIAS, and (b) the rating formats recommended by Rowe and Darkin (2001).

Total Coefficient of Determination (R^2) = 0.961 (compared with an *alpha* value of 0.890 - see last column of Table 2).[8] Note that for a one-factor model, R^2 is equivalent to the composite scale reliability coefficient (r_c) calculated from the maximally weighted factor score regression coefficients – give in equation 1.3 above, where Σ is the estimated variance-covariance matrix of the indicators; Θ_δ is estimated variance-covariance matrix among their measurement errors, and w_c is the vector of the related factor score regression coefficients (FS) for the ***Quality Area 4*** indicators ('Principles'), respectively: 0.156, 0.195, 0.194, 0.218, 0.140, 0.155.

A proportionally weighted scale score for the ***Quality Area 4*** composite that takes into account the individual and joint measurement error of the indicators can now be computed as a continuous variable for each case as follows:

Quality Area 4 = (*Prin 4.1**0.147) + (*Prin 4.2**0.184) + (*Prin 4.3**0.183) + (*Prin 4.4**0.206) + (*Prin 4.5**0.132) + (*Prin 4.6**0.148),

where *Prin 4.1* to *Prin 4.6* are the raw score ratings on the six 'Principles' (indicators) and the 'multiplier indices' are the proportional FS regression coefficients. This process ensures that the estimation of the scale/composite score (adjusted for measurement error) is proportionally weighted by the **actual** contribution made by each indicator. Note that these proportionally weighted FS regression coefficients given in the parentheses above add to 1; hence the composite score will range from a minimum of 1 to a maximum of 4. This means that the composite score for ***Quality Area 4*** (ξ_4) as given in Table 2 (together with the other

[8] With some reluctance, but for comparative purposes, the traditional estimate of so-called 'internal consistency', namely, Cronbach's (1951) *standardized item alpha* (α) is given in the final column of Table 2 (see footnote 6).

composites - ξ_n) all have the advantage of all being 'measured' in the **same metric**, regardless of the number of indicators or 'Principles' associated with each *Quality Area*.

Table 2. Indicator Factor-score Regression Weights and Related Composite Scale Parameters

Scale	Indicator ('Principle') Weights#						r_c	$\hat{\lambda}_c$	$\hat{\theta}_c$	α*
ξ_{c4} – ***Quality Area 4***	4.1	4.2	4.3	4.4	4.5	4.6				
Raw Factor Score Regression indices:	.156	.195	.194	.218	.140	.155				.890
Proportional Factor Score Regression indices:	.147	.184	.183	.206	.132	.148	**.961**	**.952**	**.082**	

* Cronbach's standardized item alpha.[9]

Note: $\boldsymbol{r_c}$ is the composite scale reliability coefficient calculated from the maximally weighted factor score regression coefficients obtained from fitting one-factor congeneric measurement models to the constituent scale items, given by:

$$r_C = \frac{w_C^{'}\left(\hat{\Sigma} - \hat{\Theta}_\delta\right)w_C}{w_C^{'}\,\hat{\Sigma}\,w_C}$$, where w_c is the vector of factor score regression weights.

$\hat{\lambda}_c = \sigma_c\sqrt{r_c}$ is the estimate of that part of the variance in the vector of indicator variables ($\underline{x}$) that is explained by the composite variable ξ_c.

$\hat{\theta}_c = \sigma_c^2\,(1 - r_c)$ is the estimate of the remaining variance in the indicator variables not explained by the composite variable - i.e., measurement error variance.

The computed composite score for ***Quality Area 4***, and similarly in the case of constructs for each of the other *Quality Areas*, may now be used to construct an ***Accreditation Performance Profile*** as illustrated in Figure 2 (p. 10).

An important advantage of an ***Accreditation Performance Profile*** as illustrated in Figure 2 (p. 10) is that regardless of the number of constituent indicators the composite score for each *Quality Area* is on the same metric as the individual indicators (i.e., 'Principles'), namely: *Unsatisfactory*, *Satisfactory*, *Good Quality*, and *High Quality* – in this case. Note that the factor-score regression weights for each indicator (relevant to its target *Quality Area*) may be:

(a) computed from fitting a one-factor, congeneric measurement model to the rating data on the indicators, as outlined above, or

(b) based on an *a priori* assignment of weights determined by policy makers.

---

9 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 that are measured (in this case) on 5-point ordinal scales. 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 generalizability, 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).

In either case, the presentation of the summary ratings shown in Figure 2 can be generated easily.

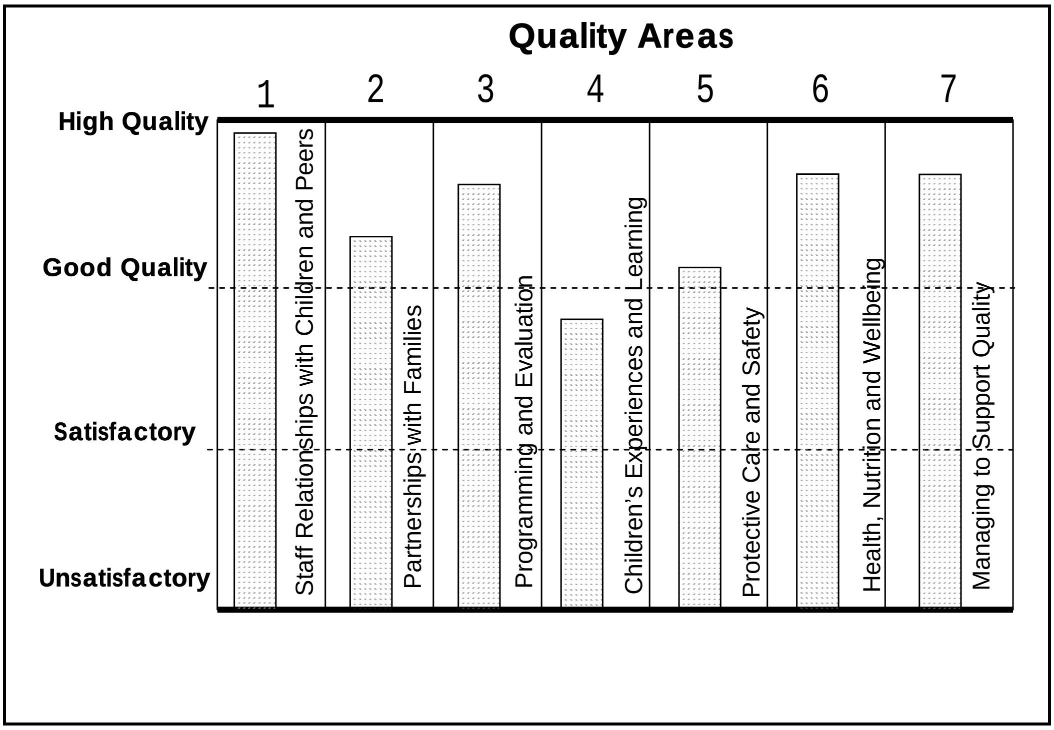

Figure 2. Hypothetical *Accreditation Performance Profile* for one long day care provider against accreditation criteria of NCAC's QIAS
(For the 7 *Quality Areas* and the 33 constituent *Principles*, see Appendix A)

Summary and concluding comments

This paper has briefly described and demonstrated the utility of confirmatory factor analysis (CFA) to compute maximally reliable composite scores from multiple indicators (‘Principles’) for NCAC’s *Quality Areas* in a *Quality Improvement Accreditation System* (QIAS). This was achieved by fitting a one-factor, congeneric measurement model to the constituent ordinal-scaled ratings on the six indicators (‘Principles’) relevant to **Quality Area 4: Children’s Experiences and Learning**. Composite scores computed by this method are single indices of their component items, each of which is weighted for its relative contribution to the composite. Unlike traditional unit-weighted methods for computing composites, the use of factor score regression weights obtained from CFA one-factor models minimizes measurement error in the indicators contributing to each composite scale, thus increasing the reliability (and validity) of the computed composite scores.

In brief, this method ensures that the estimation of the composite score assigned to a given service provider for a specific *Quality Area* is adjusted for measurement error and is proportionally weighted by the **actual** contribution made by each indicator or ‘Principle’. Note that these proportionally weighted FS regression coefficients add to 1; hence the composite scores for the seven *Quality Areas* have the advantage of being ‘measured’ in the **same metric**, regardless of the number of indicators associated with each *Quality Area*.

It is important to note, however, that regardless of any method used to compute composite scores, the ***measurement quality*** of each constituent indicator is dependent on the provision of **simple** nomenclature that is understood clearly and rated consistently by all raters. To this end, two considerations are important.

First, the wording of indicators to be rated should be stated using **positive** nomenclature. There is a considerable body of technical literature indicating that ratings on negatively-worded items or indicators are significantly less reliable than those that are worded positively, thus increasing the likelihood of measurement error.[10]

Second, it is **vital** that each indicator 'describes' a **single**, observable 'outcome' which is quantifiable in at least one category of a dichotomous ('absent' – 0, or 'present' – 1) or in ordered categories (i.e., 1, 2, 3, 4, etc.). Moreover, each indicator MUST have both ***face validity*** and ***construct validity*** with its target latent construct.[11]

References

Ainley, J., Fleming, M., & Rowe, K.J. (2002). *Evaluating systemic school reform: Literacy Advance in the early years*. Paper presented at the American Educational Research Association (AERA) conference, New Orleans (USA), April 1-5, 2002.

Alwin, D.F., & Jackson, D.J. (1980). Measurement models for response errors in surveys: Issues and applications. In, D. Schuessler (Ed.), *Sociological Methodology*. San Fransisco: Jossey Bass.

Australia (1998). *National Community Housing Standards* (1st ed.). Canberra, ACT: Commonwealth Department of Family and Community Services & NSW Office of Community Housing.

Bentler, P.M. (1980). Multivariate analysis with latent variables: Causal modelling. *Annual Review of Psychology, 31*, 419-56.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: John Wiley.

Brown, R.L. (1989). Using covariance modeling for estimating reliability on scales with ordered polytomous variables. *Educational and Psychological Measurement, 49*, 383-398.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Advances in Quantitative Techniques in the Social Sciences, Volume 1. Beverly Hills & London: Sage Publications.

Carroll, J.B. (1961). The nature of data, and how to choose a correlation coefficient. *Psychometrika, 26*, 347-372.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cuttance, P. (1987). Issues and problems in the application of structural equation models. In, P. Cuttance, & R. Ecob (Eds.), *Structural Modeling by Example: Applications in Educational, Sociological and Behavioral Research* (pp. 241-279). Cambridge: Cambridge University Press.

Fleishman, J., & Benson, J. (1987). Using LISREL to evaluate measurement models and scale reliability. *Educational and Psychological Measurement, 47*, 925-939.

Goldstein, H. (1995). *Multilevel statistical models*. London: Edward Arnold.

Gorsuch, R. (1983). *Factor analysis*. Hillsdale, NJ: Erlbaum.

Healy, M.J.R, & Goldstein, H. (1976). An approach to the scaling of categorized attributes. *Biometrika, 63*, 219-229.

---

[10] Nearly 30 years ago, Sandoval (1977) provided an evidence-based critique of the use of rating scales employing negatively-worded items by demonstrating that such items are highly susceptible to rater bias and response sets such as 'reverse halo effects' or 'reverse generosity errors' (Kerlinger, 1986; Sellitz, Wrightsman & Cook, 1976). In a subsequent comparative study of format effects in rating scales, Sandoval (1981) showed 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 measurements. In contrast, an inspection of the marginal distributions for negatively-worded items indicated that they tended to be highly skewed and leptokurtic. This is not a trivial issue (see Rowe & Rowe, 1997, 2004).

[11] ***Construct validity*** is established through a rational analysis of the content of an indicator or set of indicators – based on individual, subjective judgment. There are two major types of ***construct validity***: *face validity* and *logical validity*. ***Face validity*** is established when it is agreed (by consensus) that that an indicator (e.g., an indicator/Principle rating) is a valid measure of a relevant construct (i.e., **Quality Area** domain). ***Logical*** or ***sampling validity*** involves a careful definition of the **domain** of elements to be measured and the logical design of indicators to cover all the relevant areas of this domain.

Hendrickson, L., & Jones, B. (1987). A study of longitudinal causal models comparing gain score analysis with structural equation approaches. In P. Cuttance, & R. Ecob (Eds.), *Structural modeling by example: Applications in educational, sociological and behavioral research* (pp. 86-107). Cambridge: Cambridge University Press.

Hill, P.W., Rowe, K.J., Holmes-Smith, P., & Russell, V.J. (1996). *The Victorian Quality Schools Project: A study of school and teacher effectiveness*. Report to the Australian Research Council (Volume 1 – Report). Centre for Applied Educational Research, Faculty of Education, The University of Melbourne.

Holmes-Smith, P. (1998). *Analysis of the Quality Improvement and Accreditation System applying to long day care children's services: Report to the Commonwealth Child Care Advisory Council, Department of Health and Family Services*. Camberwell, VIC: Australian Council for Educational Research.

Holmes-Smith, P., & Rowe, K.J. (1994). *The development and use of congeneric measurement models in school effectiveness research: Improving the reliability and validity of composite and latent variables for fitting multilevel and structural equation models*. Paper presented at the International Congress for School Effectiveness and Improvement, World Congress Centre, Melbourne, January 3-6, 1994.

Huba, G.J., & Harlow, L.L. (1987). Robust structural equation models: Implications for developmental psychology. *Child Development, 58*, 147-166.

Jöreskog. K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109-133.

Jöreskog, K.G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity, 24*, 387-404.

Jöreskog, K.G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59,* 381-389.

Jöreskog, K.G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, Mass: Abt Books.

Jöreskog, K.G., & Sörbom, D. (1988). *PRELIS: A Program for Multivariate Data Screening and Data Summarization: A Preprocessor for LISREL*. Chicago: Scientific Software, Inc.

Jöreskog, K.G., & Sörbom, D. (1989). *LISREL 7: A Guide to the Program and Applications*. Chicago: SPSS, Inc.

Jöreskog, K.G., & Sörbom, D. (1993). *New Features in PRELIS 2*. Chicago: Scientific Software, Inc.

Jöreskog, K.G., & Sörbom, D. (2005a). *PRELIS 2, Version 2.72 for MS Windows: A Program for Multivariate Data Screening and Data Summarization; a Preprocessor for LISREL*. Lincolnwood, IL: Scientific Software International, Inc.

Jöreskog, K.G., & Sörbom, D. (2005b). *LISREL 8.72: Interactive LISREL for MS Windows*. Lincolnwood, IL: Scientific Software International, Inc.

Keeves, J.P., & Lakomski, G. (Eds.) (1998). *Issues in educational research*. Amsterdam: Pergamon.

Kerlinger, F.N. (1986) *Foundations of behavioral research,* 3rd edition. New York: Holt Rinehart & Winston.

Kreft, I.G., de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Lawley, D.N., & Maxwell, A.E. (1971). *Factor analysis as a statistical method* (2nd edition). London: Butterworths.

Mardia, K.V., Kent, J.T., & Bibby, J.M. (1979). *Multivariate Analysis*. London: Academic Press.

Marsh. H.W. (1987). The hierarchical structure of self-concept and the application of hierarchical confirmatory factor analysis. *Journal of Educational Measurement, 24*, 1-39.

Marsh, H.W. (1994). Confirmatory factor analysis models of factorial invariance. *Structural Equation Modeling, 1* (1), 5-34.

Marsh, H.W., & Grayson, D. (1994). Longitudinal confirmatory factor analysis: Common, time-specific, item specific, item specific and residual error components of variance. *Structural Equation Modeling, 1*(2), 116-145.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Masters, G.N. (1988). The analysis of partial credit scoring. *Applied Measurement in Education, 1*, 279-297.

Masters, G.N. (2001a). *The key to objective measurement in the psychosocial sciences*. Camberwell, Vic: Australian Council for Educational Research (MIMEO).

Masters, G.N. (2001b). *Educational measurement: ACER Assessment Resource Kit (ARK)*. Camberwell, Vic: Australian Council for Educational Research.

Masters, G.N., & Keeves, J.P. (Eds.) (1999). *Advances in Measurement in Educational Research and Assessment*. New York: Pergamon (Elsevier Science).

Masters, G.N., & Wright, B.D. (1997). The partial credit model. In W.J. van der Linden and R.K. Hambleton (Eds.), *Handbook of item response theory*. New York: Springer.

McDonald, R.P. (1978). A simple comprehensive model for the analysis of covariance structures, *British Journal of Mathematical and Statistical Psychology, 31*, 59-72.

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

McDonald, R.P. (1985). *Factor analysis and related methods*. Hillsdale, N.J: Erlbaum.

McDonald, R.P. (1996). Path analysis with composite variables. *Multivariate Behavioral Research, 31*, 239-270.

Miller, M.B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling, 2*, 255-273.

Morris, R.J., Bergan, J.R., & Fulginiti, J.V. (1991). Structural equation modeling in clinical assessment research with children. *Journal of Consulting and Clinical Psychology, 59*, 371-379.

Munck, M.E. (1979). *Model Building in Comparative Education: Applications of the LISREL Method to Cross-national Survey Data*. International Association for the Evaluation of Educational Achievement Monograph Series No. 10. Stockholm: Almqvist & Wiksell International.

Muthén, B.O. (1993). Goodness of fit with categorical and other non-normal variables. In K.A. Bollen & J. Scott Long (Eds.), *Testing structural equation models*. Newbury Park, CA: Sage Publications.

NCAC (2005a). *Quality Improvement and Accreditation System: Self-study Report* (3rd ed.). Surry Hills, NSW: National Childcare Accreditation Council Inc.

NCAC (2005b). *Quality Improvement and Accreditation System: Quality Practices Guide* (1st ed.). Surry Hills, NSW: National Childcare Accreditation Council Inc.

NSW (1999). *Community Housing Accreditation Kit*. Parramatta, NSW: Standards and Accreditation Unit, Office of Community Housing, NSW Department of Housing.

NSW (2002). *Outline of New South Wales Community Housing Standards and Accreditation System*. Parramatta, NSW: Standards and Accreditation Unit, Office of Community Housing, NSW Department of Housing.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*, 443-460.

Poon, W-Y., & Lee, S-Y. (1987). Maximum likelihood estimation of multivariate polychoric correlation coefficients. *Psychometrika, 52*, 409-430.

Rasch, G (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedogogiske Insitut.

Rasch, G (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14*, 58-94.

Raykov, T (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173-184.

Raykov, T. (1998). Cronbach's Coefficient Alpha and the reliability of composite scales with interrelated non-homogeneous items. *Applied Psychological Measurement*, *22*, 375-385.

Rowe, K.J. (1989). The commensurability of the general linear model in the context of educational and psychosocial research. *Australian Journal of Education, 33*, 41-52.

Rowe, K.J. (1998). Accounting for the hierarchical structure of data in psychosocial research: An annotated example using multilevel structural equation modeling. In M. Mok, & S. Burke (Eds.), *Collected papers on applications in multilevel modeling*. Sydney, NSW: Macquarie University Press.

Rowe, K.J. (2001). Educational performance indicators. In M. Forster, G. Masters and K.J. Rowe, Measuring learning outcomes: Options and challenges in evaluation and performance monitoring (pp. 2-20*). Strategic Choices for Educational Reform; Module IV – Evaluation and Performance Monitoring*. Washington, DC: The World Bank Institute.

Rowe, K.J. (2002a). *The measurement of composite variables from multiple indicators: Applications in Quality Assurance and Accreditation Systems (Child Care)*. Background paper prepared for the National Childcare Accreditation Council. Camberwell, VIC: Australian Council for Educational Research.

Rowe, K. J. (2002b). *Statistical analyses of Accreditation ratings on the National Community Housing Standards indicators: Report to the Community Housing Standards and Accreditation Units of the New South Wales and Queensland Departments of Housing*. Camberwell, VIC: Australian Council for Educational Research.

Rowe, K.J. (2003). Estimating interdependent effects among multilevel composite variables in psychosocial research: An example of the application of multilevel structural equation modeling. In S.P. Reise and N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 255-284). Mahwah, NJ: Lawrence Erlbaum Associates.

Rowe, K.J. (2004). *Analysing and reporting performance indicator data: 'Caress' the data and user beware!* Background paper to invited address presented at the 2004 Public Sector Performance & Reporting Conference (under the auspices of the International Institute for Research – IIR), Sydney, 19-22 April 2004. Available at: http://www.acer.edu.au/research/programs/learningprocess.html.

Rowe, K.J. (2006a). *Practical multilevel analysis with MLwiN & LISREL: An integrated course* (5th edition, revised). 22nd ACSPRI Summer Program in Social Research Methods and Research Technology, Australian National University, 23-27 January 2006. Camberwell, VIC: Australian Council for Educational Research.

Rowe, K.J. (2006b). *Analyses of Principle ratings on 7 Quality Areas of the Quality Improvement Assurance System for Long Day Care Centres*. Report to National Childcare Accreditation Council, June 2006. Camberwell, VIC: Australian Council for Educational Research.

Rowe, K.J., & Darkin, L. (2001). *An examination of options for developing quality assurance measurement instruments for accreditation of long day care centres: Report to National Childcare Accreditation Council*. Camberwell, Vic: Australian Council for Educational Research.

Rowe, K.J., & Hill, P.W. (1998). Modeling educational effectiveness in classrooms: The use of multilevel structural equations to model students' progress. *Educational Research and Evaluation, 4*(4), 307-347.

Rowe, K.J., & Rowe, K.S. (1992a). The relationship between inattentiveness in the classroom and reading achievement (Part A): Methodological Issues. *Journal of the American Academy of Child and Adolescent Psychiatry, 31*(2), 349-356.

Rowe, K.J., & Rowe, K.S. (1992b). The relationship between inattentiveness in the classroom and reading achievement (Part B): An Explanatory Study. *Journal of the American Academy of Child and Adolescent Psychiatry, 31*(2), 357-368.

Rowe, K.J., & Rowe, K.S. (1999). Investigating the relationship between students' *attentive-inattentive* behaviors in the classroom and their literacy progress. *International Journal of Educational Research, 31*(1/2), 1-138 (Whole Issue).

Rowe, K.J., & Rowe, K.S. (2004). *Developers, users and consumers beware: Warnings about the design and use of psycho-behavioral rating inventories and analyses of data derived from them*. Background paper to invited workshop presented at the International Test Users' Conference 2004, Hilton on the Park, Melbourne, 19-20 July 2004. Available for download in PDF format at: http://www.acer.edu.au/research/programs/learningprocess.html.

Rowe, K.J., Tainton, J., & Taylor, D. (2004). Key features of the Quality Improvement and Accreditation System. *Putting Children First: Newsletter of the National Childcare Accreditation Council*, Special Edition, June 2004, pp. 19-23.

Rowe, K.S., & Rowe, K.J. (1997). Norms for parental ratings on Conners' Abbreviated Parent-Teacher Questionnaire: Implications for the design of behavioral rating inventories and analyses of data derived from them (Leading article). *Journal of Abnormal Child Psychology, 25,* 425-451.

Sandoval, J. (1977). The measurement of the hyperactive syndrome in children. *Review of Educational Research, 47,* 293-318.

Sandoval, J. (1981). Format effects in two teacher rating scales of hyperactivity. *Journal of Abnormal Child Psychology, 9,* 203-218.

Scarr, S. (1985). Constructing psychology: Making facts and fables for our times. *American Psychologist, 40,* 499-512.

Scott Long, J. (1983). *Confirmatory factor analysis: A preface to LISREL*. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-033. London: Sage Publications.

Sellitz, C., Wrightsman, L.S., & Cook, S.W. (1976). *Research methods in social relations* (3rd ed.). New York: Holt Rinehart & Winston.

Stephanou, A. (2000). Using the Rasch model to study large scale Physics examinations in Australia. In M. Wilson, & G. Engelhard jr (Eds.), *Objective Measurement: Theory into Practice-V* (pp. 65-94). Stamford, CO: Ablex Publishing Corporation.

Stevens, J.J. (1995). Confirmatory factor analysis of the Iowa Tests of Basic Skills. *Structural Equation Modeling, 2*, 214-231.

Werts, C.E., Rock, D.R., Linn, R.L., & Jöreskog, K.G. (1978). A general method for estimating the reliability of a composite. *Educational and Psychological Measurement*, *38*, 933-938.

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

Appendix A: *Quality Areas* and *Principles* of NCAC's QIAS for Long Daycare Centres

Quality Area 1: Staff Relationships with Children and Peers

1.1 Staff interact with each child in a warm and friendly way
1.2 Staff guide each child's behaviour in a positive way
1.3 Staff initiate and maintain respectful communication with each child
1.4 Staff respect each child's background and abilities
1.5 Staff treat all children equitably
1.6 Staff communicate effectively to promote respect and professional teamwork

Quality Area 2: *Partnerships with Families*

2.1 Staff and families communicate effectively to exchange information about each child and the centre
2.2 Staff encourage family participation and involvement in the centre
2.3 The centre has orientation processes for children and families

Quality Area 3: *Programming and Evaluation*

3.1 The program reflects a clear statement of centre philosophy
3.2 Each child's learning is documented and is used in planning the program
3.3 The program assists each child to be a successful learner

Quality Area 4: *Children's Experiences and Learning*

4.1 Staff encourage each child to make choices and participate in play
4.2 Staff promote each child's ability to develop and maintain relationships
4.3 Staff promote each child's language and literacy abilities
4.4 Staff promote each child's problem solving and mathematical abilities
4.5 Staff promote each child's enjoyment of and participation in the expressive arts
4.6 Staff promote each child's physical abilities

Quality Area 5: *Protective Care and Safety*

5.1 Staff act to protect each child
5.2 Staff supervise children at all times
5.3 Staff ensure that potentially dangerous products, plants and objects are inaccessible to children
5.4 The centre ensures that buildings and equipment are safe
5.5 The centre promotes occupational health and safety

Quality Area 6: *Health, Nutrition and Wellbeing*

6.1 Staff promote healthy eating habits
6.2 Staff implement effective and current food safety and hygiene practices
6.3 Staff encourage children to follow simple rules of hygiene
6.4 Staff ensure toileting and nappy changing procedures are positive experiences
6.5 Staff support each child's needs for rest, sleep and comfort
6.6 The centre acts to control the spread of infectious diseases and maintains records of immunisations

Quality Area 7: *Managing to Support Quality*

7.1 Written information about the centre's management is readily available to families
7.2 Written information about the centre's management is readily available to staff
7.3 Staffing policies and practices facilitate continuity of care for each child
7.4 Management provides professional development opportunities for staff