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

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

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) underlining 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 & Fulginiti, 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). 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) 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
(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>
<thead>
<tr>
<th>Indicators (Principles)</th>
<th>Unsatisfactory</th>
<th>Satisfactory</th>
<th>Good Quality</th>
<th>High Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Staff encourage each child to make choices and participate in play</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4.2 Staff promote each child’s ability to develop and maintain relationships</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4.3 Staff promote each child’s language and literacy abilities</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4.4 Staff promote each child’s problem solving and mathematical abilities</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4.5 Staff promote each child’s enjoyment of and participation in the expressive arts</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4.6 Staff promote each child’s physical abilities</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Using raw data 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).

**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 ($x_i$) on $\xi_1$ (the latent, composite variable) where the elements $\lambda_{xi}$ are the partial regression coefficients of $\xi_1$ in the regression of $x_i$ on $\xi_1$, and $\delta_i$ is a vector of measurement errors in $x_i$, namely:

$$
\begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3 \\
  x_4 \\
  x_5 \\
  x_6 \\
\end{bmatrix} = 
\begin{bmatrix}
  \lambda_{x1} \\
  \lambda_{x2} \\
  \lambda_{x3} \\
  \lambda_{x4} \\
  \lambda_{x5} \\
  \lambda_{x6} \\
\end{bmatrix} \begin{bmatrix}
  \xi_1 \\
\end{bmatrix} + 
\begin{bmatrix}
  \delta_1 \\
  \delta_2 \\
  \delta_3 \\
  \delta_4 \\
  \delta_5 \\
  \delta_6 \\
\end{bmatrix}
$$

or,

$$x_i = \lambda_{xi} \xi_1 + \delta_i \tag{1.1}$$

The assumed model implies that the covariance matrix of the observed indicators ($x_i$) is of the form:

$$
\Sigma = \Lambda_{xl} \Lambda_{xl}' + \Theta_\delta \tag{1.2}
$$

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

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 ($\xi_c$) 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).
The Measurement of Composite Variables

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

where \( \hat{\Sigma} \) is the estimated variance-covariance among the factor ‘loadings’ for the vector of congeneric indicators (\( x \)'s), \( \hat{\Theta}_\delta \) 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. 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.

\footnote{For a one-factor congeneric measurement model, the factor score regression coefficients (FS) represent the estimated bivariate regression of the factor (\( \xi \)) on all the observed indicator variables, given by: \( FS = A'\Sigma^{-1} \), where \( A \) is the estimated factor pattern matrix and \( \Sigma \) 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 (\( \xi \)) may be computed for any individual \( i \) with observed scores \( x_i \), using the simple product: \( \xi_i = FS x_i \).}
The Measurement of Composite Variables

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).

Note that for a one-factor model, \( R^2 \) is equivalent to the composite scale reliability coefficient \( (\rho_c) \) calculated from the maximally weighted factor score regression coefficients – give in equation 1.3 above, where \( \Sigma \) is the estimated variance-covariance matrix of the indicators; \( \Theta_i \) 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:

\[
\text{Quality Area 4} = (\text{Prin 4.1} \times 0.147) + (\text{Prin 4.2} \times 0.184) + (\text{Prin 4.3} \times 0.183) + (\text{Prin 4.4} \times 0.206) + (\text{Prin 4.5} \times 0.132) + (\text{Prin 4.6} \times 0.148),
\]

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 \( (\xi_4) \) as given in Table 2 (together with the other

---

K.J. Rowe, August 2006

ACER
composites - $\xi_c$) 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

<table>
<thead>
<tr>
<th>Scale</th>
<th>Indicator ('Principle') Weights#</th>
<th>$r_c$</th>
<th>$\hat{\lambda}_c$</th>
<th>$\hat{\theta}_c$</th>
<th>$\alpha^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\xi_{4}$ - Quality Area 4</td>
<td>4.1 4.2 4.3 4.4 4.5 4.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Factor Score Regression indices:</td>
<td></td>
<td>.156 .195 .194 .218 .140 .155</td>
<td></td>
<td></td>
<td>.890</td>
</tr>
<tr>
<td>Proportional Factor Score Regression indices:</td>
<td></td>
<td>.147 .184 .183 .206 .132 .148</td>
<td></td>
<td></td>
<td>.961 .952 .082</td>
</tr>
</tbody>
</table>

* Cronbach’s standardized item alpha.\(^9\)

Note: $r_c$ 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:

$$ r_c = \frac{w'_c (\sum - \hat{\Theta}_s) w_c}{w'_c \sum w_c}, $$

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 ($\sum$) that is explained by the composite variable $\xi_c$.

$\hat{\theta}_c = \sigma_c^2 (1 - r_c)$ 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 $\alpha$: (1) the magnitude of $\alpha$ is a direct function of the number of items in a scale, regardless of their individual and shared error variance, and (2) $\alpha$ 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 $\alpha$ 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](image)

(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


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.


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.


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.


K.J. Rowe, August 2006


K.J. Rowe, August 2006


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

<table>
<thead>
<tr>
<th>Quality Area 1: Staff Relationships with Children and Peers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Staff interact with each child in a warm and friendly way</td>
</tr>
<tr>
<td>1.2 Staff guide each child’s behaviour in a positive way</td>
</tr>
<tr>
<td>1.3 Staff initiate and maintain respectful communication with each child</td>
</tr>
<tr>
<td>1.4 Staff respect each child’s background and abilities</td>
</tr>
<tr>
<td>1.5 Staff treat all children equitably</td>
</tr>
<tr>
<td>1.6 Staff communicate effectively to promote respect and professional teamwork</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality Area 2: Partnerships with Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Staff and families communicate effectively to exchange information about each child and the centre</td>
</tr>
<tr>
<td>2.2 Staff encourage family participation and involvement in the centre</td>
</tr>
<tr>
<td>2.3 The centre has orientation processes for children and families</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality Area 3: Programming and Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 The program reflects a clear statement of centre philosophy</td>
</tr>
<tr>
<td>3.2 Each child’s learning is documented and is used in planning the program</td>
</tr>
<tr>
<td>3.3 The program assists each child to be a successful learner</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality Area 4: Children’s Experiences and Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Staff encourage each child to make choices and participate in play</td>
</tr>
<tr>
<td>4.2 Staff promote each child’s ability to develop and maintain relationships</td>
</tr>
<tr>
<td>4.3 Staff promote each child’s language and literacy abilities</td>
</tr>
<tr>
<td>4.4 Staff promote each child’s problem solving and mathematical abilities</td>
</tr>
<tr>
<td>4.5 Staff promote each child’s enjoyment of and participation in the expressive arts</td>
</tr>
<tr>
<td>4.6 Staff promote each child’s physical abilities</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality Area 5: Protective Care and Safety</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Staff act to protect each child</td>
</tr>
<tr>
<td>5.2 Staff supervise children at all times</td>
</tr>
<tr>
<td>5.3 Staff ensure that potentially dangerous products, plants and objects are inaccessible to children</td>
</tr>
<tr>
<td>5.4 The centre ensures that buildings and equipment are safe</td>
</tr>
<tr>
<td>5.5 The centre promotes occupational health and safety</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality Area 6: Health, Nutrition and Wellbeing</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Staff promote healthy eating habits</td>
</tr>
<tr>
<td>6.2 Staff implement effective and current food safety and hygiene practices</td>
</tr>
<tr>
<td>6.3 Staff encourage children to follow simple rules of hygiene</td>
</tr>
<tr>
<td>6.4 Staff ensure toileting and nappy changing procedures are positive experiences</td>
</tr>
<tr>
<td>6.5 Staff support each child’s needs for rest, sleep and comfort</td>
</tr>
<tr>
<td>6.6 The centre acts to control the spread of infectious diseases and maintains records of immunisations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality Area 7: Managing to Support Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1 Written information about the centre’s management is readily available to families</td>
</tr>
<tr>
<td>7.2 Written information about the centre’s management is readily available to staff</td>
</tr>
<tr>
<td>7.3 Staffing policies and practices facilitate continuity of care for each child</td>
</tr>
<tr>
<td>7.4 Management provides professional development opportunities for staff</td>
</tr>
</tbody>
</table>