

Use of Data Collapsing Strategies to Identify Latent Variables in Questionnaire Data: Strategic Management of Junior and Middle School Data on the CHP Questionnaire

Peter Grimbeek, Fiona Bryer, Wendi Beamish, & Michelle D'Netto

Centre for Learning Research, Griffith University

A dataset of 399 junior and middle school students completed the Cognitive Holding Power questionnaire (CHPQ), which distinguishes between first-order procedural thinking to achieve specific goals and second-order solving of problems involving new situations. Factor analysis using the original 5-point scale indicated that these student responses were not completely consistent with the theorised two-factor structure. Some items contributed only marginally or became associated with the "wrong" factor. Analyses of these test data in the present study compared the outcomes of collapsing a 5-point Likert scale into 4- versus 2-category response options. By convention, four categorical points represent the minimum acceptable set for factor analysis (Byrne, 2001). However, collapsing from five to two response categories more fully corrects other methodological issues related to the occurrence of disordered difficulty in levels of response categories within items, indicative that participants' responses to items response patterns of the item set do not fit the expected sequence.

Introduction

The Cognitive Holding Power questionnaire (CHPQ) was designed to measure associations between differing settings for learning and differing levels of thinking. These differing levels of thinking have been conceptualised as representative of lower (first-order) versus higher (second-order) cognitive activity. Stevenson (1998) described first-order cognitive activities as use of procedural knowledge, and second-order cognitive activities as use of specific problem-solving procedures that deal with unfamiliar situations. The instrument was developed in the course of a doctoral study (Stevenson, 1984). Trials were conducted in secondary schools (Stevenson, 1998, 1992) and TAFE settings (Stevenson, 1984, 1990, 1991; Stevenson & McKavanagh, 1991; Stevenson & Evans, 1994). This group of studies has consistently reported outcomes compatible with CHPQ's premise that students will report differential use of general procedural and more specific cognitive procedures. That is, the test was found to measure two distinct dimensions of learning.

In part, the instrument was designed to identify the extent to which learning settings press high school students into different levels of thinking. Stevenson (1998) extended his previous study of Year 8 high school students in order to explore the performance of

the CHPQ in terms of its reliability across settings. He also examined the influence of teacher style and subject demands on the relative levels of first- and second-order cognitive activities. That is, a teacher's emphasis on one of another kind of cognitive activity was thought to affect the class culture for student thinking. Some teachers adopted a first-order approach, with students needing to copy and work as shown by the teacher. Other teachers placed an emphasis on second-order activities, with students needing to check results and find links between things.

D'Netto (2004) administered the CHPQ to 399 junior and middle school students to gauge the cognitive press of their environment. Their teachers were interviewed, and the perceptions of the students and their teachers were compared. Results indicated that, in these environments, both first- and second-order cognitive levels were used. A class environment that pressed for second-order thinking was one in which systematic enquiry-based tasks were established, high-order thinking was expected, and teacher action and student action were balanced in a process of "fading" teacher support and planned progression through different phases of skill development. However, younger students in the sample were less able to distinguish between the two kinds of press.

It seems likely that emerging cognitive changes though early adolescence would shape observed outcomes in the sense that these changing response tendencies may reduce the reliability of the instrument for younger students. It is worth noting that the two-factor model of first- and second-order abilities was based on the notion of student sensitivity to educational processes (e.g., subject demands and teacher influence) rather than on any consideration of developmental processes. That is, item difficulty and item response distributions might change as test-taker's ability changes.

Rationale

The present study focused primarily on the effect of a strategy of collapsing response categories (Beamish, 2004) on CHPQ data from junior and middle school settings. A fundamental issue with the use of Likert scale items is the problematic measurement properties of multi-choice response categories per item. Differing assumptions about the measurement properties of Likert items determine conflicting "rules of thumb" for analysing such Likert data.

If Likert data is assumed to be nonparametric, it can be viewed from a qualitative perspective. In this case, it might be expected that collapsing response categories would improve the intelligibility of the outcomes of analysis (i.e. "less is better"). This strategy for data analysis involves the strong assumption that Likert scale items are not interval data (Beamish, 2004). It follows that the interval between levels remains uncertain and unquantifiable. Beamish's (2004) use of nonparametric analysis of a nonlinear dataset to conduct her analysis of collapsed data makes this strong assumption. Under these conditions, decreasing the number of response categories by systematically collapsing across categories within items (data-slicing), as demonstrated by Beamish in relation to early intervention practices, facilitates dynamic inferences about decision-making in terms of distinct levels of agreement-disagreement, including certainty-doubt and misunderstanding-understanding.

One rationale in favour of collapsing across response categories, however, is that Likert scale response categories not only provide a positive opportunity for a smoother distribution of responses (i.e., a normal spread of choices across categories) but also allow "negative" opportunities for participants to misjudge the intensity of what is inherently a qualitative response. That is, the range of available response categories can obscure rather than clarify the intent of the respondent. A strategy for minimising respondent ambiguity is to collapse across response categories. The effect of this strategy on, for example, an acceptance scale (Beamish, 2004), might be to reduce the 5-point response categories of *Strongly disagree*, *Disagree*, *Undecided*, *Agree*, and *Strongly agree* to dichotomous categories representing the participant's choice between Disagreement (Collapsing across *Strongly disagree*, *Disagree*, *Undecided*) or Agreement (collapsing across *Agree*, *Strongly agree*). One constraint on this collapsing strategy is that all items should be collapsed in the same manner; that is, the method of collapsing should be constant across items. Another constraint is that a sound conceptualisation informs the decision to collapse categories.

When undertaking descriptive analyses employing tables or graphs, it is clear that collapsing responses into dichotomous categories has distinct advantages in terms of capturing trends in the data (Beamish, 2004). Likewise, contingency and other analyses that function more efficiently with larger numbers of participants per cell can benefit from Likert indicators—and other categorical or ordinal indicators—being collapsed into dichotomous rather than four or more categories.

This strategy of collapsing across response categories, however, runs counter to the contrary assumption that Likert scale items and the latent variables measured by them are equal interval data. It follows that collapsing across interval data points (i.e., response categories) reduces the sensitivity or power of the measurement both in terms of reliability of measurement and normalcy of response distribution (i.e., "more is better").

Collapsing data across response categories into fewer response categories (e.g., trichotomous or dichotomous) infringes methodological conventions about questionnaire data. There are several contributing reasons that account for the conventional rules of thumb for using multiresponse survey items in exploratory or confirmatory factor analyses of the soundness of test construction. Quantitative analysts prefer to assume that, even if Likert scale items are ordinal (Michell, 2003), the latent variables they express possess interval-scale measurement properties. The gap between measurement reality and measurement conventions is also bridged by assuming that scales with 4 or more points approximate interval measurement (Byrne, 2001), such that every point is equivalent in value and absolute distance from every other point in an ordered array. For such reasons, 4-point response scales are regarded as at the lower limits of acceptability for factor analysis.

A further rationale for the convention of the 4-response category cut-off point is that the distributional properties of items with fewer than four categories are held to be unreliable. Unreliability of alternative responses chosen by test respondents is of particular concern if the level of skew on an item exceeds 1 and if it is in different directions on different items. This concern is even more pressing if the level of kurtosis (peakedness or flatness of distribution) exceeds 1 and is in different directions

(i.e., peaked vs. flattened) on different items. Minimising the number of response categories can affect the likelihood of high levels of skew or kurtosis ($\Rightarrow 1$).

Although there are situations in which nonnormal distributions of data should be maintained (e.g., Beamish's study of practice consensus, mastery testing), linear scaling of tests is the general case that underlies test construction procedures. Therefore, in the present study, the simplifying assumption will be made that the data is parametric, but a series of factor analyses (exploratory & confirmatory) will investigate the effect of a data collapsing strategy (i.e., whether more or less is better). The strategy for combining response categories used in the present study is to use Rasch analytic procedures to identify out-of-order response categories and collapse so as to reduce their incidence (i.e., increase statistical intelligibility). Rasch analytic techniques were used as a guide to identify appropriate categories to collapse across. The question is whether the beneficial effect of reducing the incidence of disordered response categories outweighs the adverse effects of a reduced response distribution. Re-analysis of this CHPQ dataset provides an opportunity to test the contrary rules of thumb of less is better versus more is better by varying the number of response categories and reporting some estimates measures of skew and kurtosis and fit estimates normally associated with confirmatory factor analysis.

Data analyses

Because these further analyses constitute the major part of this paper, description of the original method of data collection is correspondingly brief.

Method

Of the total of 399 participants who completed the CHPQ in D'Netto's study, 43% ($N = 172$) were in junior year levels, and the remainder in the middle school years. These participants varied in age from 8 to 15 years, and 47% ($N = 187$) were male, with approximately 50% of these in junior and middle school years respectively. Of the total of 209 female students, approximately 38% ($N = 79$) were in the junior school year levels. These participants were drawn from four junior classrooms and five middle school classrooms.

The instrument used in the present analyses comprised 27 items, 13 of which expressed first-order and 14 of which expressed second-order cognitive activities (see items listed in Table 2). Participants responded to items on a 5-point scale (*Almost never, Seldom, Sometimes, Often, Very often*).

Diagnostic screening

The first step was to examine the dataset as a precursor to developing analysable versions with acceptable item qualities. Diagnostic screening of items was used to examine the distribution of responses across response categories per item (SPSS) and the ordering of response categories per item (WINSTEPS). WINSTEPS was developed for the purpose of Rasch item analysis. It treats these categorical responses (e.g., *Sometimes*) as separate and as categorical (or ordinal) rather than equal-interval. It also rescales item scores prior to further analyses and then reports item difficulty or test-taker ability in terms of these response categories. Thus, Rasch analysis provides an alternative approach to the more traditional computation of the average score for either an individual item or for the

whole test (or subsets of tests items). For this reason, WINSTEPS is a useful addition to SPSS frequencies when adopting a strategy of collapsing across the response categories of items.

Initial screening of the data using SPSS frequencies indicated that participants had not produced skewed responses to items. However, Rasch analytic examination of the 27 items, based on WINSTEPS, indicated that the average score per response category was out of sequence for six items. Table 1 illustrates the detection of an out of order sequence of responses for Q21 such that participants rated the two least positive responses (*Almost never, Seldom*) as having more or less equivalent difficulty. This out of order sequence can be contrasted with that for Q22, in which participants responded in a sequence across the five response categories in keeping with the putative difficulty for those categories. That is, Q22 reflects the normal sequence of responses, in which participants, on average, found it to be more difficult to make increasingly positive responses.

Table 1

WINSTEPS reporting of sequence of response category responses in terms of average measure and standard error for two of the 27 items

ITEM	RESPONSE CATEGORY CODE	AVERAGE MEASURE	STANDARD ERROR	LABEL
11	Almost never	-0.08	0.09	Q21
11	Seldom	-0.09	0.05	Q21
11	Sometimes	0.12	0.03	Q21
11	Often	0.35	0.04	Q21
11	Very often	0.61	0.06	Q21
12	Almost never	-0.29	0.08	Q22
12	Seldom	-0.07	0.05	Q22
12	Sometimes	0.24	0.02	Q22
12	Often	0.44	0.04	Q22
12	Very often	0.62	0.08	Q22

Note. Shaded scores indicate responses out of order.

Based on this WINSTEPS analysis, the average response per category was identified as out of sequence for six items that included four first-order items (Q16, Q23, Q24, Q28) and two second-order items (Q21, Q72). In each case, participant selections of response categories were such that they confused the difficulty level of the two least positive response categories (*Almost never, Seldom*). In the case of Q24, participants also confused the difficulty level of the two most positive response categories (*Often, Very often*). Accordingly, items were collapsed to form two comparable datasets. The first dataset was collapsed to form four response categories (*Seldom* [Almost never, Seldom], *Sometimes, Often, Very often*) that took into account most but not all of the items with undesirable qualities but did retain the requisite number of categories required by convention for factor analysis. The second dataset (after examining the skew of resulting variables) was collapsed to form two (i.e., dichotomous) response categories (Less often

[*Almost never, Seldom, More often*] and More often [*Often, Very often*]) that took into account all of the items with undesirable qualities but violated the number of categories conventions for factor analysis (with its interval-scale based assumptions about data).

Exploratory factor analyses

Table 2 presents factor analytic outcomes produced by the original 5-category array of response options. The purpose of the 5-category analysis was to provide a baseline for judging the efficacy of the collapsing categories strategy. A parallel sequence of factor analyses for the collapsed 4-category and 2-category datasets, presented in Table 3, were compared with both original outcomes and with each other.

The analytic strategy involved two steps. First, exploratory analysis (viz., SPSS Maximum likelihood factor analysis and Varimax rotation) was used in order to provide a purely empirical measure of the extent to which these two collapsed datasets supported the theorised data structure (see Tables 2 & 3). Second, confirmatory factor analysis (i.e., AMOS CFA) was used in order to provide a theoretically based measure of the extent to which the two datasets support the theorised model of the data structure. As AMOS also produces (a) univariate measures of skew and (b) univariate and multivariate measures of kurtosis, the outcomes can also be compared in terms of the relative levels of skew and kurtosis produced by these 5-, 4-, and 2-category Likert response scales.

Table 2 illustrates the analytic outcome of undertaking exploratory factor analysis with the 5-category dataset. Seven of the 27 variables displayed empirical problems: These items failed to load significantly (≥ 0.30), loaded on both factors, or did not load on the theorised factor. Although the dual-loading item (Q26) did not load at significant levels on the other factor and did load significantly on the theorised factor, another item (Q16) not only did not load significantly as theorised but also loaded more strongly (if nonsignificantly) on the nontheorised factor.

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy provided additional information about the factorability of the two datasets. As with Cronbach's Alpha, a value of 0.800 or above is considered to indicate an acceptable level of factorability. In the case of the 5-category dataset, these items would be regarded as fairly factorable (KMO = 0.770), and the two-factor structure accounted for a portion of the cumulative variance (20%).

Table 3 compares the results of undertaking exploratory factor analysis (Maximum likelihood, Varimax rotation) for the 4- and 2-category sets of items. In neither case did the items fit the theorised structure perfectly. In the case of the 4-category set, six items loaded nonsignificantly. In the case of the 2-category set, seven items loaded nonsignificantly. In terms of sampling adequacy, the KMO = 0.779 was better for the 4-category set than KMO = 0.723 for the 2-category set. In terms of cumulative variance explained, the 4-category set explained 20% and the 2-category 16%.

Table 2

Exploratory analyses (Maximum likelihood, Varimax rotation) using 5-category^a items
(loadings =>0.25 shown)

ITEMS/FACTORS	1	2
Q1 SO Ask Qs to chk rslts	.250	
Q2 SO Hve to try new ideas	.407	
Q3 SO Stdts encrgd to find links	.399	
Q4 SO Hve to find info myslf		
Q7 SO Chk rslts agnst known	.443	
Q11 SO Stdts encrgd to try new ideas	.597	
Q12 SO Feel that must chk rslts	.428	
Q13 SO Find links btn lrrnt things	.476	
Q15 SO Stdts encrgd to find out	.366	
Q19 SO Try out new ideas	.541	
Q21 SO Stdts encrgd to Q as chk	.408	
Q22 SO Feel I have to find links	.524	
Q27 SO Find info myself	.339	
Q29 SO Stdts encrgd to chck rslts	.454	.257
Q5 FO Let tchr tell me what to do		.287
Q6 FO Feel that must copy tchr		.554
Q8 FO Get all info from tchr		.448
Q9 FO Stdts encrgd to copy		.465
Q16 FO Stdts encrgd to do as told	.284	
Q17 FO Feel that must wrk as shown		.422
Q18 FO Rely on tchr to show links		.499
Q20 FO Copy what tchr does		.634
Q23 FO Accept rslts without Q		
Q24 FO Do things my way		
Q26 FO Stdts encrgd to wrk as shown	.285	.430
Q28 FO Rely on tchr for new ideas		.413
Q30 FO Work exactly as shown	.274	.430

^aWith respect to Likert scaling, 5-category responses is the same as a 5-point scale.

Table 3

Exploratory analyses (Maximum likelihood, Varimax rotation) for 4- (right) & 2-category (left) items (loadings =>0.25 shown)

ITEMS/FACTORS	1	2	ITEMS/FACTORS	1	2
Q1 SO Ask Qs to chk rslts	.274		<i>Q1 SO Ask Qs to chk rslts</i>	.269	
Q2 SO Hve to try new ideas	.423		Q2 SO Hve to try new ideas	.317	
Q3 SO Stdts encrgd to find links	.394		Q3 SO Stdts encrgd to find links	.302	
Q4 SO Hve to find info myslf			<i>Q4 SO Hve to find info myslf</i>		
Q7 SO Chk rslts agnst known	.464		Q7 SO Chk rslts agnst known	.492	
Q11 SO Stdts encrgd to try new ideas	.587		Q11 SO Stdts encrgd to try new ideas	.480	
Q12 SO Feel that must chk rslts	.460		Q12 SO Feel that must chk rslts	.402	
Q13 SO Find links btn lrrnt things	.493		Q13 SO Find links btn lrrnt things	.443	
Q15 SO Stdts encrgd to find out	.362		<i>Q15 SO Stdts encrgd to find out</i>	.289	
Q19 SO Try out new ideas	.527		Q19 SO Try out new ideas	.441	
Q21 SO Stdts encrgd to Q as chk	.420		Q21 SO Stdts encrgd to Q as chk	.359	
Q22 SO Feel I have to find links	.517		Q22 SO Feel I have to find links	.463	
Q27 SO Find info myself	.363		Q27 SO Find info myself	.342	
Q29 SO Stdts encrgd to chck rslts	.452	.263	Q29 SO Stdts encrgd to chck rslts	.408	
Q5 FO Let tchr tell me what to do		.290	<i>Q5 FO Let tchr tell me what to do</i>		.260
Q6 FO Feel that must copy tchr		.496	Q6 FO Feel that must copy tchr		.407
Q8 FO Get all info from tchr		.431	Q8 FO Get all info from tchr		.308
Q9 FO Stdts encrgd to copy		.424	Q9 FO Stdts encrgd to copy		.405
Q16 FO Stdts encrgd to do as told	.279	.289	<i>Q16 FO Stdts encrgd to do as told</i>		
Q17 FO Feel that must wrk as shown		.484	Q17 FO Feel that must wrk as shown		.448
Q18 FO Rely on tchr to show links		.518	Q18 FO Rely on tchr to show links		.439
Q20 FO Copy what tchr does		.553	Q20 FO Copy what tchr does		.474
Q23 FO Accept rslts without Q			<i>Q23 FO Accept rslts without Q</i>		
Q24 FO Do things my way			<i>Q24 FO Do things my way</i>		
Q26 FO Stdts encrgd to wrk as shown	.265	.494	Q26 FO Stdts encrgd to wrk as shown		.447
Q28 FO Rely on tchr for new ideas		.451	Q28 FO Rely on tchr for new ideas		.438
Q30 FO Work exactly as shown		.492	Q30 FO Work exactly as shown		.440

In short, the 4-category data set appeared to outperform both the 5- and 2-category datasets in terms of items loading on factors. It also appeared to be on a par with the 5-category set in terms of factorability and cumulative variance explained, with both these performing slightly better than the 2-category set. In summary, the collapsing categories strategy seems to perform optimally within the bounds of the 4-category minimum response scale convention for factor analysis outlined by Byrne (2001).

Confirmatory factor analyses

At this point in the life cycle of factor analysis, one option would be to undertake an iterative series of exploratory factor analyses, at each step removing nonsignificantly loading items (<0.30), until some final analysis yields a trimmed set of items all of which load significantly and in a conceptually sensible fashion. The alternative is to perform an iterative series of confirmatory factor analyses (CFA) with much the same aim. However, taking this approach has the advantage of explicitly presupposing that the theorised model illustrated in Figure 1 is the factor structure of choice. These CFA outcomes not only provide estimates of the strength of association between items and the factor of choice but also estimate the variance per item not explained by the model.

Another important feature of the confirmatory factor analytic process is that it also provides a list of four types of statistical measures relevant to CFAs and to structural equation models (SEM) more generally. These measures include correcting the chi-square test for model complexity, estimating the residual variance not accounted for by the model, comparing the tested model to a baseline model, and making more general estimates of goodness of fit:

- (a) Chi-square/degrees of freedom (*df*) computation (correcting chi-square for model complexity), which should approximate the 0-3 range;
- (b) RMR and RMSEA (estimating residual variance), which should approximate the 0-0.05 range;
- (c) NFI, TLI, CFI, and RFI (comparing tested model to baseline model), which should approximate the 0.9-1.0 range; and
- (d) GFI and AGFI (estimating goodness of fit), which should approximate the 0.9-1.0 range.

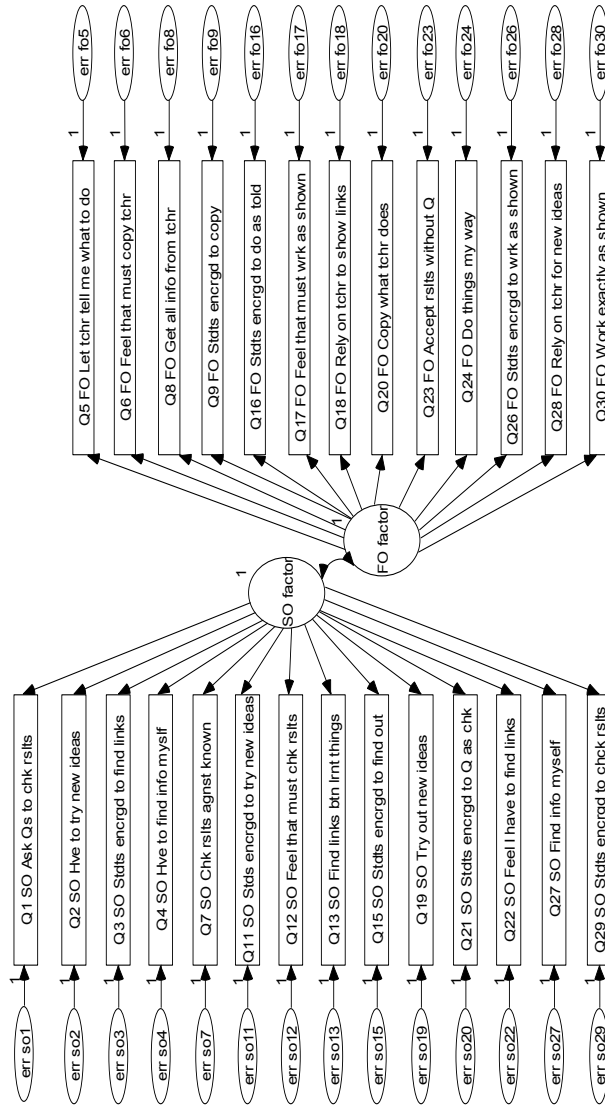


Figure 1. Initial confirmatory factor analysis (CFA) tested the specified 27-item two-factor model by measuring the strength of association between each item and its latent factor (regression weight or beta weight) and by taking into account that portion of the variance not explained by the model via error terms associated with each of the items (Factor variance set to 1).

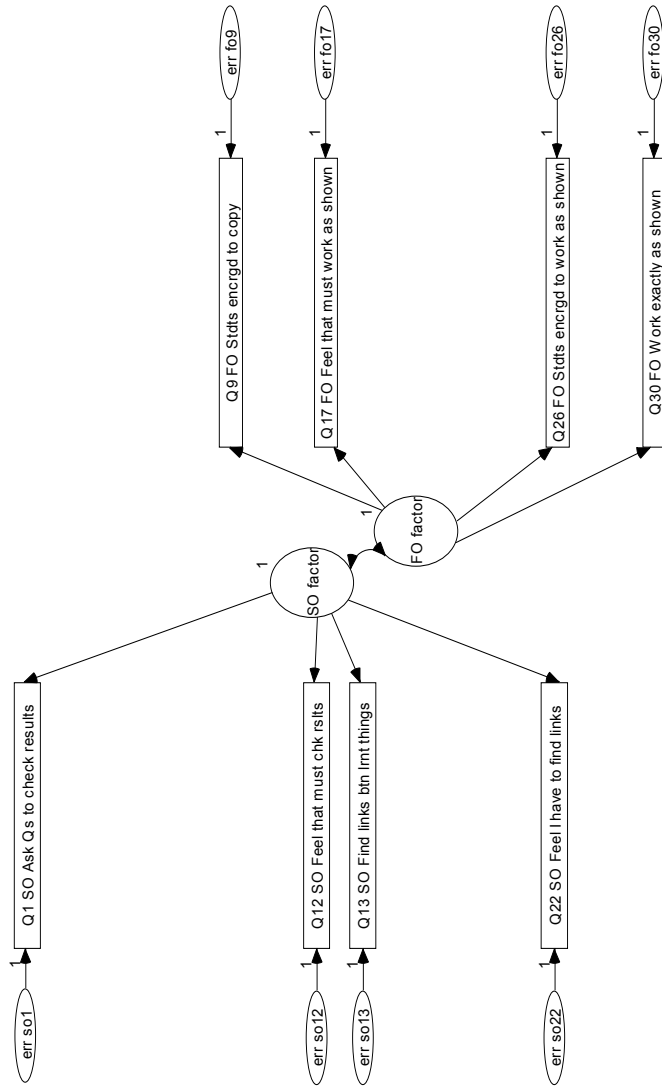


Figure 2. Follow-up confirmatory factor analysis (CFA) tested the specified 8-item two-factor model by measuring the strength of association between each item and its latent factor (regression weight or beta weight) and by taking into account that portion of the variance not explained by the model via error terms associated with each of the items (Factor variance set to 1).

Table 4 presents results from a range of tests for the 27-item CHPQ. What is clear from the results is that, regardless of the number of categories, the 27-item model did not fare well with the data collected from this sample. The number of acceptable measures increased as the number of response categories decreased. Table 5 provides various measures that complement the standard goodness of fit estimates in Table 4. In terms of Mardia's measure of multivariate kurtosis, the collective peakedness or flatness of items varied to the advantage of the 2-category set. In other respects, the 5-category model outperformed its alternatives. In summary, the 5-category model matched the 4-category models in terms of the number of items significantly associated with the theorised factors. It outperformed the collapsed category models in terms of minimising the number of excessively skewed or kurtotic items. However, the 5-category model trailed behind the collapsed category models in terms of Mardia's estimate of multivariate kurtosis and in terms of a range of goodness of fit measures. In these respects, the 2-category model outperformed all others.

Table 4

Estimates of goodness of fit for the 5-, 4-, and 2-category sets based on the 27-item two-factor model

MEASURE	5-CATEGORY SET	4-CATEGORY SET	2-CATEGORY SET
Chi-square	946.845	876.827	625.801
<i>df</i>	323	323	323
Probability	0.000	0.000	0.000
Chi/ <i>df</i>	2.931	2.715	1.937
RMR	0.078	0.058	0.013
RMSEA	0.070	0.066	0.049
NFI	0.652	0.590	0.564
RFI	0.524	0.554	0.527
TLI	0.625	0.663	0.697
CFI	0.655	0.690	0.721
GFI	0.840	0.854	0.894
AGFI	0.813	0.829	0.876

Table 5

Number of items with nonsignificant regression weights or excessive skew or kurtosis, plus estimates of multivariate kurtosis for the 5-category, 4-category, and 2-category 27 item CHP model

MEASURE	5-CAT. SET	4-CAT. SET	2-CAT. SET
Nonsignificant regression weight	1	1	2
Excessive skew (1+)	0	5	8
Excessive kurtosis (1+)	1	1	21
Mardia's measure of multivariate kurtosis	124.833	72.232	-2.126

^aCat. = Number of response categories available for test item.

The 27-item model was subjected to an iterative series of CFAs in which items with nonsignificant regression weights or excessively correlated errors were trimmed from the model. Figure 2 illustrates the eight-item two-factor model that emerged from this process.

Table 6 presents results from a range of tests for the 8-item CHPQ. What is clear from the results is that, regardless of the number of categories, the 8-item model fared very well with the data collected from this sample, with an optimal number of acceptable measures in the 4-response category dataset. It is also clear that the 4-category dataset outperformed both the 5- and the 2-category sets in terms of a range of goodness of fit estimates.

Table 6

Estimates of goodness of fit for the 5-, 4-, and 2-category sets based on the 8-item two-factor model

MEASURE	5-CATEGORY SET	4-CATEGORY SET	2-CATEGORY SET
Chi-square	29.260	20.455	22.184
<i>df</i>	19	19	19
Probability	0.062	0.368	0.275
<i>Chi/df</i>	1.540	1.077	1.168
RMR	0.037	0.024	0.008
RMSEA	0.037	0.034	0.021
NFI	0.924	0.947	0.915
RFI	0.888	0.922	0.875
TLI	0.957	0.994	0.980
CFI	0.971	0.996	0.986
GFI	0.981	0.987	0.986
AGFI	0.965	0.975	0.974

Note. Bolded values indicate fit within specified ranges.

Table 7 illustrates various measures that complement the standard goodness of fit estimates. Although all items were significantly associated with theorised factors, Mardia's measure of multivariate kurtosis was optimal for the 4-category dataset and less than optimal for either the 5- or 2-category set. In terms of excessive skew, only the 2-category set departed from zero. The 5-category set outperformed its alternatives in terms of the number of items with excessive kurtosis. In summary, the 4-category set mostly either matched or surpassed the 5- and 2-category sets.

Table 7

Number of items with nonsignificant regression weights or excessive skew or kurtosis, plus estimates of multivariate kurtosis for the 5-category, 4-category, and 2-category 8-item CHP model

MEASURE	5-CAT. SET	4-CAT. SET	2-CAT. SET
Nonsignificant regression weight	0	0	0
Excessive skew (1+)	0	0	2
Excessive kurtosis (1+)	0	3	6
Mardia's measure of multivariate kurtosis	9.007	3.402	-6.917

Discussion

This paper has described an analysis of previously collected data (D'Netto, 2004) that used exploratory and confirmatory factor analysis. The aim was to examine the effect of collapsing response categories on goodness of fit and associated distributional measures for the CHP model published by Stevenson and Ryan (1994). Limited collapsing improved reported estimates.

The strategy of collapsing across response categories appeared to confer an advantage in both exploratory and confirmatory factor analysis. That is, less is better. It is equally evident that the more restrained version of this strategy (the 4-category data set) generally outperformed not only the original 5-point scale but also the dichotomous scale. That is, more is better. This outcome is consistent with the convention reported by Byrne (2001) of using Likert scale data with four or more categories in preference to those with less. This outcome is also inconsistent with the implicit assumption sustaining the convention (i.e., that more categories are better), because, in this case, four was better than five. Furthermore, until a 3-category set is added to the mix of analyses, one cannot be assured that the 4-category set is, in fact, the ideal representation of this test data set.

It is likely that the superiority of the 4-category version of the dataset over the 3- or 2-category version in this instance is not necessarily an absolute. That is, the rules of thumb emerging from considerations of measurement properties require empirical testing. In every such analysis of educational test data, empirical testing is required to establish the optimal number of response categories. Failure to collapse across categories in some cases might render the data unintelligible and unanalysable.

Based on the present analyses, the responses of this mixed sample of junior and middle school students did not readily approximate the conventional CHP model. One speculation, based on these outcomes, is that the scope for application of the CHPQ does not extend smoothly across the entirety of these subsets of students. Even if middle-school students responded meaningfully to these items in a manner reflective of the postulated latent variables for first- and second-order reasoning, it is likely that junior school students might be less likely to do so for cognitive-developmental reasons. Future research might examine the extent to which these 27-item and 8-item models of the CHPQ generalise to junior school students. Moreover, the instrument was designed to identify the extent to which learning settings press students into different levels of

thinking (Stevenson, 1998). Another speculation, therefore, is that, in this case, the learning environment had somewhat modest effects.

Acknowledgements

This manuscript is based on access to a large dataset provided by Michelle D'Netto and her thesis supervisor Dr. Charlie McKavanagh (Centre for Learning Research). Michelle generated these data for her masters study. They were generous in making their data available to further analysis of data collapsing strategies, and this contribution is gratefully acknowledged.

References

- Beamish, W. (2004). *Consensus about program quality: An Australian study in early childhood special education*. Unpublished doctoral dissertation, Griffith University, Queensland, Australia.
- Byrne, B. M. (2001). *Structural equation modeling with AMOS*. London: Lawrence Erlbaum Associates.
- D'Netto, M. (2004). *The press for high-order thinking in new basics classrooms*. Unpublished masters dissertation, Griffith University, Queensland, Australia.
- Stevenson, J. (1998). Performance of the cognitive holding power questionnaire in schools. *Learning and Instruction, 8*(5), 393–410.
- Stevenson, J. C., & Evans, G. T. (1994). Conceptualisation and measurement of Cognitive Holding Power. *Journal of Educational Measurement, 31*(2), 1–20.
- Stevenson, J. C. (1991). Cognitive structures for the teaching of adaptability in vocational education. In G. T. Evans (Ed.), *Learning and teaching cognitive skills* (pp. 144–163). Hawthorn, VIC: Australian Council for Educational Research.
- Stevenson, J. C., & McKavanagh, C. (1991). *Cognitive structures developed in TAFE classes*. Paper presented at the annual conference of the Australian Association for Research in Education, Gold Coast, Queensland.
- Stevenson, J. C. (1990, December). *Conceptualisation and measurement of Cognitive Holding Power in technical and further education learning settings*. Paper presented at the annual conference of the Australian Association for Research in Education, Sydney.
- Stevenson, J. C. (1992, November). *Performance of the Cognitive Holding Power Questionnaire in Queensland schools*. Paper presented at the joint conference of the Australian Association for Research in Education and the New Zealand Association for Research in Education, Deakin University at Geelong.
- Stevenson, J., & Ryan, J. (1994). *Cognitive Holding Power Questionnaire manual*. Brisbane, Australia: Griffith University, Centre for Skill Formation Research and Development.