Author Name: Allen G. Harbaugh
Contact Email: allen.harbaugh@murdoch.edu.au

ALTERNATE DESTINIES FOR SURVEY ITEMS DESTINED FOR THE ISLAND OF MISFIT TOYS: AN ANALYSIS OF TEACHERS' PERCEPTIONS OF NAPLAN

Allen G. Harbaugh & Greg Thompson
Murdoch University

Abstract

This is a methodological paper describing when and how manifest items dropped from a latent construct measurement model (e.g., factor analysis) can be retained for additional analysis. Presented are protocols for assessment for retention in the measurement model, evaluation of dropped items as potential items separate from the latent construct, and post hoc analyses that can be conducted using all retained (manifest or latent) variables. The protocols are then applied to data relating to the impact of the NAPLAN test. The variables examined are teachers' achievement goal orientations and teachers' perceptions of the impact of the test on curriculum and pedagogy. It is suggested that five attributes be considered before retaining dropped manifest items for additional analyses. (1) Items can be retained when employed in service of an established or hypothesized theoretical model. (2) Items should only be retained if sufficient variance is present in the data set. (3) Items can be retained when they provide a rational segregation of the data set into subsamples (e.g., a consensus measure). (4) The value of retaining items can be assessed using latent class analysis or latent mean analysis. (5) Items should be retained only when post hoc analyses with these items produced significant and substantive results. These suggested exploratory strategies are presented so that other researchers using survey instruments might explore their data in similar and more innovative ways. Finally, suggestions for future use are provided.

Keywords: survey instruments, confirmatory factor analysis, latent class analysis

Any researcher that has frequently used confirmatory factor analysis (CFA) to assess the convergent validity of a congeneric latent factor model has encountered the unexpected loading result. Simply put, this is when you are working with a well-established survey instrument (or a theoretically sound instrument) with a common and reasonable set of items. However, one or more of the "really good" or "blatantly obvious" items fails to "fit" into the model properly. When enough items are present, dropping a few items is not generally seen as a threat to the construct validity under examination. Though, if you are like me, we are left scratching our heads trying to figure out why the one really strong item for everybody else did not work this time. While there are numerous hypothetical scenarios that might describe these findings, the focus of this paper is less on why the desired loading pattern was not observed and more on what might be done with the items otherwise destined for the *Island of Misfit Toys* (Muller & May, 1964).

Introduction: Latent Constructs and Confirmatory Factor Analysis

Many texts are available to explain the theory and practical application of Factor Analysis (e.g., Bartholomew, Steele, Moustaki, & Galbraith, 2002; Bollen, 1989; Byrne, 2010; Kline, 2011). As such, only a brief review of the basic decision rules for retaining or dropping survey items is discussed here. Furthermore, this paper assumes that the researcher will be working with a well-established or theoretically sound survey instrument. Thus, the focus is on confirmatory factor analysis (CFA) as opposed to exploratory factor analysis (EFA). That is, it is assumed that the researcher is confirming that a set of items measures the desired latent construct (model testing) as opposed to exploring which items among a large set might be influenced by the presence of one or more possible latent constructs

Author Name: Allen G. Harbaugh
Contact Email: allen.harbaugh@murdoch.edu.au

(model building)[1].

The distinction between EFA and CFA is well understood as a model-building vs. model-testing protocol. Even still, there are "exploratory" elements that seep into the CFA. Because reliability is an attribute of the data (the scores or subjects under study) as opposed to the model (the survey instrument or latent constructs to be studied), it is recommended that an examination of the internal reliability of the survey instrument be conducted each time it is used (Barnette, 1999; Thompson, 1994). It is this reason that the factor loading of the latent construct's influence on manifest items be examined each time a survey instrument is used. This generally involves an analysis of the model fit (e.g., CFI or RMSEA statistics), an analysis of the estimated parameters (e.g., factor loadings or covariances among latent constructs), and sometimes the choice to drop items from the measurement model. While some might argue that this is an attribute of the EFA, in this context it should be seen as a sample-specific analysis for purposes of internal reliability.

Items are commonly dropped from use for a latent construct measure for the following reasons: (1) poor or weak factor loadings, (2) substantive cross-loadings between factors, (3) improvement of model fit indices or internal reliability measures, and (4) when flagged by modification indices for error correlations or unjustifiable regression paths (Bollen, 1989; Byrne, 2010). Though the confirmation of a high internal reliability does not demonstrate construct validity on its own, it is understood that it is still a desired quality. As such, this process has become commonplace among researchers using factor analysis and structural equation modeling.

As would be expected with a strong model, most items hypothesized to measure a latent construct are often retained in the final analysis (thus providing one more piece of evidence that the model is indeed a good model). However, there are times when the model is adjusted by dropping a few select items from the subsequent analyses. Furthermore, there are times when the adjustment to the survey instrument is unexpected. For example, in a study examining achievement goal orientations, it was found that a common item for mastery orientation ("I feel most successful in this math class if I solve a problem by working hard") failed to load on the mastery orientation latent factor (Harbaugh, 2009)[2]. Though this is provided as an example, it is anecdotal in nature. However, based on personal communication with many other researchers, it is a common experience of needing to drop items in order to improve the fit of the latent construct.

When items are dropped from a survey instrument because of lack of fit with a latent construct, it still remains necessary for the researcher to demonstrate the construct validity of the proposed latent construct. Assuming this is achieved without any issues or concerns, the curious researcher may wonder, "What might be done with the dropped item or items?"

Retaining Dropped Items for Subsequent Analyses

Even if an item that was hypothesized to load on a latent construct fails to do so for a given sample, it may be possible to use the single manifest item as a separate measure. This section presents five key considerations we suggest should be examined when deciding to retain an item as a separate measure.

Theory driven.

As in all psychometric research, the data and analyses are always subservient to the theory. A correlation between two items is something, but its value is found in the theory that it confirms or supports; its value is not the mere existence of the correlation. This paradigm extends to the hypothesized manifest items intended to indirectly measure the latent construct of interest. At first glance, this

[1] It cannot be stressed strongly enough: The protocols of this paper would take on very different interpretations (statistically and theoretically) if applied to survey instrument design. Thus, it is assumed the researcher is working with a predefined factor model.

[2] Though the purpose of this paper is not to determine why such items fail to fit the model, in this case it was assumed to be an issue with the random order of the items. This particular item was randomly placed after a work avoidant item using the common phrase "work hard." It was hypothesized that this is what affected the results for this particular item.

Author Name: Allen G. Harbaugh
Contact Email: allen.harbaugh@murdoch.edu.au

would suggest that the item should not be retained. If it was hypothesized to load with the other items, and if the item failed to load, then theory would suggest it should be dropped. And, in general, this argument would support the conclusion of dropping the item and omitting it from future analyses. However, there are some surveys that have been adjusted to fit contexts different from those in which they were originally designed. In cases such as this, there may be an alternate theory that may appropriately be applied.

A theory may guide the creation of a revised survey instrument. For example, an instrument to measure one's self-efficacy to teach may be revised to measure one's self-efficacy to teach history. In such a case, an item that fails to load on the construct of interest (self-efficacy for teaching history) may in fact be a more general measure from the original survey (self-efficacy for teaching). In this case, it is the original theory that is potentially being measured by the poorly fitting item. Thus, it might be possible to retain this item as a measure of the broader theory while the convergent items serve as a measure of the latent construct related to the sub-theory.

Sufficient variance.

An item that does not demonstrate any variance across the subjects or participants provides little to no information in correlational analyses. Thus, if all responses are the same or all responses are near one extreme end of a scale, there is no additional information that can be garnished from this item (other than the fact that everybody agrees or disagrees with the measure). Though it may be possible to utilize such measures in alternate models (e.g., item response theory or logistic regression), these items prove to be problematic for normal-distributed maximum likelihood analyses (the common strategy employed in CFA analyses, even with Likert-type data)[3]. Thus, if an item is being considered to be retained for additional analyses, it should be confirmed that it demonstrates sufficient variance across the sample. In some cases, this may be an example of a strongly skewed item on a 5- or 7-point Likert-type scale, or the item may have a multimodal distribution of responses. In either case, a substantial (though possibly small) portion of the sample must be different from the rest of the sample.

Segregating items.

Some items that fail to align with the original latent construct model may demonstrate poor fit because of a violation of the test assumptions. For example, strongly skewed or bimodal items may not demonstrate good model fit with the CFA congeneric model when using normal-distribution maximum likelihood. While it may be possible to obtain different results using asymptotically distribution-free analyses or analyses that account for the ordinal nature of Likert-type data, these are generally not used frequently in practice when assessing congeneric latent factor models[4]. One possible scenario may be that the poorly fitting item is measuring the presence of a particular trait (e.g., writing in a diary everyday or never at all) or a willingness to disagree with a belief held by majority or consensus (e.g., most people support recycling but a few opponents disagree). As such, these might be mildly related to the latent construct of interest, but the weak correlation coupled with the violation of the model assumptions results in the item being dropped from the measurement model.

In these cases, it may be that the item should be used as a separate demographic-style variable or a variable that allows the researcher to separate the sample into 2 or more smaller subsamples. For example, it was shown that writing motivation may be different for individuals that write in diaries daily compared to those that never write in diaries (Troia, Harbaugh, Shankland, Wolbers, & Lawrence, 2013). Another example is when teachers disagree with the consensual belief that standardized tests result in a need to teach to the test (Harbaugh & Thompson, 2013, Thompson & Harbaugh, 2013). In each of these examples, the item of interest failed to load on the hypothesized latent construct, but the item indicated that there were possibly 2 different subsamples in the larger sample.

[3] And even with IRT or logistic regression, these analysis protocols still require even minimal variation. If all responses are identical, it becomes impossible to estimate associated population parameters.
[4] This is not to suggest that these strategies should not be explored because they are not frequently used or well understood. Instead, this is a suggestion for an alternate (if not perfect) strategy.

Author Name: Allen G. Harbaugh
Contact Email: allen.harbaugh@murdoch.edu.au

Latent class analysis.

Items that fail to load on a latent construct may still prove useful for other latent measurements models. One possibility is that a dropped item can serve as a key indicator item for a latent class. Thus, the item can be used to classify the data or respondents into categories or classes. As in the previous scenario, this latent construct may be mildly or indirectly related to the original latent construct for which the item was intended to measure. The advantage to this approach is that the use and presence of an item is explained by a theoretical model. Furthermore, this model can be tested for good model-data fit. Using latent class analysis, it is possible to assess the relationship of an ordinal item (the true nature of Likert-type scale data). Furthermore, it is possible to conduct additional analyses (e.g., chi-square tests of independence) with the resulting groups.

Post hoc analyses.

Finally, it is suggested that dropped items be retained for further analysis if the exploratory nature of the study would benefit from the confirmation (or disconfirmation) of a possible statistically significant relationship. To this end, dropped items could be retained as proxy-interval level measures or could be used as ordinal or dichotomized measures (as in the case for a variable separating the sample into smaller subgroups). These analyses could be traditional statistical analyses (e.g., t-tests, chi-square tests and ANOVA) or could be more advanced latent model analyses such as a moderation analysis using latent mean analysis (LMA).

Examples of Subsequent Analyses

As with all research in the social sciences, the process is driven by research questions and hypotheses. However, the subsequent analyses using manifest items that have been dropped from the measurement model are inherently exploratory. The logic is straightforward: the items were hypothesized to be part of a measurement model, and as such they were not intended to be used to answer a prescribed research question or to prove a given hypothesis. Thus, in the absence of an *a priori* question or goal, any additional analysis with the "new" measure has to be exploratory. While caution must be taken when interpreting results and presenting findings, there is still value in describing those empirical observations for future researchers.

This section presents a few possible analyses that can be conducted using dropped survey items. Some of these concepts were briefly addressed in the section addressing the assessment protocols to decide to retain items, but they will be elaborated here. As the focus has been on measurement models related to survey items, this discussion will focus on use of Likert-type data (though many of the ideas, suggestions and cautionary comments would extend to other type of manifest measurement items). Additional analyses will be grouped into 3 categories: ordinal level analysis, interval level analyses and analyses with dichotomized data. When a survey item is retained for further analysis separate from its original intended purpose, the researcher has the opportunity to reexamine the item carefully. The purpose of this paper is not to justify the use of analyses on data of different measurement levels from that which the analysis was originally intended[5]. Nor is it intended to justify the process of dichotomizing (or trichotomizing, etc.) data (for different opinions on the process, please refer to Dawson & Weiss, 2012; DeCoster, Iselin & Gallucci, 2009; MacCallum, Zhang, Preacher & Rucker, 2002). The purpose here is to simply suggest what strategies might be employed—with caution—when further exploring one's data.

Ordinal level analyses.

While one might argue that ordinal level analysis is most appropriate, there are limitations to what can be done with the data at this measurement level. As an ordered categorical variable, the main an-

[5] Though, it should be noted that in one aspect, this happens frequently and without question when Likert-type data are used as manifest items in factor analysis. Quite often, the estimation protocols used are intended for interval level data, and (with some justification from past research) the ordinal level measures are treated as interval level.

Author Name: Allen G. Harbaugh
Contact Email: allen.harbaugh@murdoch.edu.au

alyses available would be tests of independence (chi-square tests) with other categorical variables and ANOVAs with other interval level variables. For example, a chi-square test could be conducted to determine if there is a (unspecified) relationship between the different levels of agreement with a dropped survey item (e.g., "I enjoy marking exams") and dichotomous demographic grouping variable (e.g., "I have moved to a new school within the last 2 years"). As an example of the ANOVA, one could use the different levels of agreement as the factor levels when examining an interval level measure (e.g., an aggregate score for one of the other constructs being measured). Continuing with the current scenario, a set of items measuring job satisfaction could be averaged and used as the dependent variable in an ANOVA with the response levels for the enjoy-marking-exams item.

Interval level analyses.

Regression techniques are available when treating Likert-type data as interval level data. When relating the dropped item to categorical variables, logistic regression would be used. When relating the dropped item to interval or ratio level variables, standard regression would be used. Furthermore, the dropped item could be retained in a multiple regression analysis or used as a manifest variable in a structural equation modeling (SEM) analysis. In addition, the retained item could be used as a covariate in an ANCOVA.

Analyses with dichotomized data.

When the responses to a dropped item are dichotomized (or reduced to 3 or more smaller groupings), detailed information about the variable is lost. However, if this is done, there are a number of analyses available. First, all of the ordinal level analyses mentioned above can be conducted with dichotomized data. In addition, moderation analyses can be conducted using either multiple regression or SEM. For example, if the dropped item suggests a dichotomization of the sample into two subgroups, it is possible to determine if the relationships between a set of items are the same for each group, or if one group demonstrates different relationships.

Analyses for all types of data.

In addition to the analyses suggested for each grouping, it is possible to use the dropped survey item (at any of the measurement levels) along with other variables under study to conduct cluster analyses. This could be accomplished using classification and regression trees (CART), latent class analysis (LCA) or more traditional hierarchical clustering algorithms. In turn, these variable-level clusters can then be used to explore differences in other variables under examination in the study.

Practical Applications

This section provides a practical demonstration of some of these ideas applied to authentic data. The data used here was collected for a study of teachers' perceptions of the impact of NAPLAN on teaching, motivation and well-being. Data regarding achievement goal orientations were collected from $n_1 = 846$ teachers, and additional data regarding curriculum and pedagogy were collected from $n_2 = 809$ of the 846 teachers. Data was collected via a snowball-sampling protocol from April-June 2012. More information about the data collection process, sample demographics and additional variables measured can be found in previously reported studies (Harbaugh & Thompson, 2013; Thompson, 2012; Thompson & Harbaugh, in press).

Achievement Goal Orientations.

Twelve survey items were written to measure three goal orientations (mastery, performance-approach, and performance-avoidance) in relation to the NAPLAN test. Prior to the analysis of the items, it was noted that 2 of the 12 items did not make specific reference to NAPLAN. As might be expected, these two items demonstrated weak fit with the overall measurement model. While the analysis presented here does not completely represent the analysis protocol used to decide to drop the items from the measurement model, this does present a reasonable summary of the key concerns.

Author Name: Allen G. Harbaugh
Contact Email: allen.harbaugh@murdoch.edu.au

Table 1
Achievement Goal Orientations CFA Fit Statistics, Standardized Factor Loadings, Reliabilities and Distribution Frequencies.

	Final Model	Model + #2	Model + #5
Fit Indices			
$\chi^2(df)^*$	74.01 (31)	108.9 (38)	95.04 (38)
RMSEA	0.041	0.047	0.042
CFI	0.982	0.971	0.977
SRMR	0.033	0.037	0.034
Factor Loadings			
Mastery-#1	0.548	0.613	0.560
Mastery-#3	0.364	0.351	0.374
Mastery-#4	0.603	0.548	0.584
Perf-Appr-#6	0.593	0.592	0.594
Perf-Appr-#7	0.731	0.733	0.729
Perf-Appr-#8	0.607	0.606	0.608
Perf-Avoid-#9	0.849	0.849	0.842
Perf-Avoid-#10	0.816	0.816	0.819
Perf-Avoid-#11	0.729	0.728	0.735
Perf-Avoid-#12	0.554	0.554	0.556
Adjusted Model			
Item Added		Mastery-#2	Performance-#5
Mastery		0.292	0.201
Perf-Appr		0.031**	0.074**
Perf-Avoid		0.049**	0.308
Reliability (α)			
Mastery	0.474	0.505	
Perf-Appr	0.584		0.676
Frequency			
SD		7 (0.8%)	5 (0.6%)
D		0 (0.0%)	33 (3.9%)
N		14 (1.7%)	93 (11.0%)
A		304 (35.9%)	494 (58.4%)
SA		521 (61.6%)	221 (26.1%)

Note: $N = 846$; *all chi-square statistics significant at $\alpha < 0.0001$; Perf-Appr = Performance-Approach and Perf-Avoid = Perforance-Avoidance; **factor loadings not significant at $\alpha < 0.05$.

The goodness-of-fit indices (e.g., CFI and RMSEA) for the overall models were good to strong with and without the items considered for dropping (see Table 1). However, it was observed that the loadings for the two non-NAPLAN specific items were very weak ($\lambda < 0.31$); suggested cross-loadings were of comparable size. Additionally, when these non-NAPLAN specific items were included in the scale, the internal reliability measures (Cronbach α) were noticeably reduced. With these statistical findings coupled with the lack of a specific reference to NAPLAN, it was decided to drop these items from the measurement model.

From a theoretical perspective, the 2 non-NAPLAN specific items both could be conceived as general goal orientation measures for teaching. The non-NAPLAN specific item for the mastery orientation ("I actively seek out opportunities to improve my teaching skills and techniques") could be conceived as a general measure of a mastery orientation toward teaching in general. The non-NAPLAN specific item for the performance-approach orientation ("The opinions others have of my teaching ability are important to me") could be conceived as a general measure of a performance-approach orientation toward teaching in general. Thus, from a broader theoretical standpoint for achievement goal orientations (toward teaching in general as opposed to teaching with respect to NAPLAN), these two items could serve as proxy measures for the general constructs. This is also supported in that similar items have appeared as items in other achievement goal theory research (Midgley, et al., 1998).

When examining the two items more carefully, it is noted that Perf-Appr-#5 has a skewed response distribution with a modal response of "agree" (58.4%). Though the distribution is negatively

Author Name: Allen G. Harbaugh
Contact Email: allen.harbaugh@murdoch.edu.au

skewed, there is a substantial amount of variability with 15.5% of the respondents indicating a non-agreeing response (*i.e.*, SD, D or N). Furthermore, there is variability around the modal response in both directions (though more of the respondents favored one direction of the other). The item Mastery-#2 has a skewed response distribution. For this item, the modal response was the extreme end of the scale, "strongly agree" (61.6%). However, the skew is so extreme that less than 2.5% of the responses were non-agreeing (*i.e.*, SD, D or N). This lack of variability in the responses essentially results in a 2-point scale. As such, this greatly affects the validity of the item. For example, this item may be less a measure of mastery orientation and more a measure of an individual's likelihood of responding at the extreme end of a scale or not. Though both items were supported by theory, the lack of variability made the Mastery-#2 item unusable for further analysis.

As an example of additional analyses conducted with the Perf-Appr-#5 item, a multiple regression analysis was conducted in which general performance-approach orientation was included as one of the predictors (along with the other three NAPLAN-specific goal orientations). The result was that all 4 predictors were significant. In particular, the performance-approach orientation had a standardized regression coefficient of $\beta = 0.58$ ($p < 0.001$), while the non-NAPLAN performance approach had a standardized regression coefficient of $\beta = -0.10$ ($p = 0.002$). Furthermore, this result was examined further with an ANCOVA analysis in which the effect of the demographic variables on the dependent variable of pedagogy was explored after partialing out the influence of the (interval-level) covariates (of which the Perf-Appr-#5 item was one of the covariates). Even after accounting for the covariates, socioeconomic status was still a main effect on the perceived impacts on pedagogy. (More detailed analysis is available in previously reported studies [Harbaugh, Thompson & Sproul, 2013].)

Pedagogy and Curriculum.

The instrument for measuring teachers' perceptions regarding the impact of NAPLAN on curriculum and pedagogy was not a well-established instrument. It was derived from an observation instrument used to measure productive pedagogies (Lingard, Hayes, & Mills, 2003). However, the majority of items (13 items of 14 total) of the instrument were clearly supported by a substantive model regarding measurement of teachers' perceptions of pedagogy and curriculum. Because of this, a confirmatory approach was used to examine the measurement instrument. In this process, 2 items were flagged as potential problems. The first item was to load on perceptions of curriculum impact, and the second item was to load on perceptions of pedagogic impact.

It was decided to drop the problematic pedagogy item for the following reasons. Though measures of model fit and internal reliability were generally unaffected by the presence of this item (results presented in Table 2), it was noted that it had a substantively lower factor loading ($\lambda = 0.37$) than the other items on the scale ($\lambda > 0.51$)[6]. Additionally, an examination of the item's wording indicated a distinct flaw: "The criteria for judging student performance in NAPLAN are clear and explicit." All other items on the pedagogy scale measured a perception about how NAPLAN caused (promoted, facilitated, etc.) something in the classroom. This was the only item that assessed perceptions about an attribute specific to and generally confined to the test. Though previous research studies reported findings from additional analyses with this single manifest item, it is noted that none of the tests achieved statistical significance (Thompson & Harbaugh, 2013). As such, it did not seem worthwhile to retain this item for additional analyses in this context.

It was decided to drop the problematic curriculum item (#2) because of issues of cross-loading and decreased internal reliability. The overall measures of model fit (see Table 2) were generally unaffected by the inclusion of item #2. However, it was noted that over a quarter (17 or 61) of the univariate modification indices indicated a change relating to item #2. In addition, when allowed to cross-load on both the curriculum and pedagogy latent factors, significant factor loadings were observed. As might be expected with weak but significant cross-loadings, the internal reliability measures for both constructs decreased when item #2 was included in the scale. It was also noted that this

[6] Furthermore, it was noted that the next lowest factor loading was with an item that was reverse coded. Thus, it might be possible that the lower loading was not a result of poor-fit, but due to individuals responding too quickly and in the opposite pattern they might have if they read the question more carefully. This same argument could not be applied to the dropped item (#12).

Author Name: Allen G. Harbaugh
Contact Email: allen.harbaugh@murdoch.edu.au

item was the most strongly skewed of the items in this measurement instrument with 16.4% non-agreeing responses (e.g., SD, D or N). These issues strongly indicated dropping the item from the curriculum scale (even though this resulted in a scale with only 2 manifest measurement items).

Table 2
Perceptions About Pedagogy and Curriculum CFA Fit Statistics, Standardized Factor Loadings, Reliabilities and Distribution Frequencies.

	Final Model	Model + #2	Model + #12
Fit Indices			
$\chi^2(df)^*$	274.5 (50)	314.9 (60)	318.3 (60)
RMSEA	0.075	0.073	0.073
CFI	0.938	0.933	0.931
SRMR	0.051	0.055	0.050
Factor Loadings			
Pedag-#1	0.537	0.541	0.534
Pedag-#3	0.625	0.630	0.615
Pedag-#4	0.562	0.570	0.568
Pedag-#5	0.633	0.638	0.637
Pedag-#8	0.567	0.565	0.566
Pedag-#9r	0.514	0.509	0.514
Pedag-#10	0.548	0.546	0.540
Pedag-#11	0.707	0.705	0.705
Pedag-#13	0.627	0.627	0.643
Pedag-#14	0.648	0.645	0.652
Curric-#6r	0.833	0.918	0.841
Curric-#7r	0.871	0.789	0.862
Adjusted Model			
Item Added		Item-#2	Item-#12
Pedag		–0.211	0.373
Curric		0.480	0.027**
Reliability (α)			
Pedag	0.857	0.824	0.855
Curric	0.840	0.726	
Frequency			
SD		25 (3.1%)	97 (12.0%)
D		42 (5.2%)	230 (28.4%)
N		66 (8.2%)	199 (24.6%)
A		468 (57.8%)	257 (31.8%)
SA		208 (25.7%)	26 (3.2%)

Note: $N = 809$; *all chi-square statistics significant at $\alpha < 0.0001$; Pedag = Pedagogy and Curric = Curriculum; **factor loadings not significant at $\alpha < 0.05$.

Though the item was dropped from the measurement model, an examination of the item content suggested further analysis was warranted: "NAPLAN has encouraged me to give lessons that prepare students for the tests." In particular, this item was recognized as a consensus measure, a measure of whether an individual agrees with a belief held by the vast majority of a group of similar individuals. In the specific context of NAPLAN, teaching to the test as a concept was most likely heavily influenced by media reports, ongoing controversies and the advocacy of various professional associations and teacher unions. It was likely that for many teachers, NAPLAN and teaching to the test (TTTT) had become intertwined. In this case, those teachers who do not agree with the wider consensus may be a very interesting subset of the larger sample. This was essentially confirmed by the findings indicating that 16.4% of the sample did not agree with the suggestion that NAPLAN (a standardized, high-stakes tests) requires TTTT in the classroom. As this item still had sufficient variance among the responses (the modal response was "agree" with 57.8%), it was examined further for consideration as an additional grouping variable.

A consensus or concordance measure can be treated as an interval measure (indicating the degree to which an individual agrees or disagrees with a communal belief), or it can be treated as a dichoto-

Author Name: Allen G. Harbaugh
Contact Email: allen.harbaugh@murdoch.edu.au

mous measure (indicating agreement or disagreement with the communal belief). Using M*plus* 7.0, it is possible to analyze both of these approaches using latent class analysis (LCA). As the TTTT item was originally conceived as part of the survey instrument for teachers' perceptions of curriculum and pedagogy, latent clusters were examined using the latent factors and the TTTT item. As the purpose of this paper is not to explore research questions specific to these latent class analyses, only an abbreviated summary of the findings is presented here.

Table 3
Average Latent Class Probabilities for Most Likely Latent Class Membership by Latent Class

Latent Class Assignment	**Latent Class Probability**	
	Class #1	Class #2
Class #1	89.4%	10.6%
Class #2	2.2%	97.8%

A latent class model was imposed on the data in which 2 manifest items measured curriculum impact, 10 manifest items measured pedagogy impact, and 1 manifest item measured agreement with a need to TTTT. The first 2 constructs were latent measures (treating the Likert-scale items as interval level measurements) and the last measurement was treated as a categorical ordinal variable. For the two-class model, the sample-size adjusted BIC was 24858.0 and the entropy was 0.89. In addition, the average latent class probabilities for most likely latent class membership were reasonable (see Table 3). The first latent class generally disagreed with the need to TTTT; the second latent class generally agreed with the need to TTTT. Taken together, there is reasonable support for the presence of 2 possible latent classes. Next, the individual levels of the TTTT item were examined with respect to latent class membership (see Table 4). As would be expected with a consensus measure, underestimated frequencies were observed on the contradicting assignments (disagree latent class with A and SA and agree latent class with SD, D and N). Alternatively, overestimated values were observed on the concordant assignments (agree with agree and disagree with non-agree). This table demonstrates a strong relationship between response pattern on the TTTT item and latent class assignment with $\chi^2(4) = 139.3$, $p < 0.001$.

Table 4
Contingency Table for TTTT Item Responses and Latent Class Membership

Latent Class	SD	D	N	A	SA	TOTAL
disagree	17 (4.7)	21 (7.9)	33 (12.4)	71 (87.9)	10 (39.1)	152
	5.68	*4.67*	*5.85*	*–1.81*	*–4.65*	
agree	8 (20.3)	21 (34.1)	33 (53.6)	397 (380.1)	198 (168.9)	657
	–2.73	*–2.24*	*–2.81*	*0.87*	*2.24*	

Note: Expected frequencies indicated in parentheses; italicized numbers indicate signed Pearson residuals.

The same analysis was repeated with a dichotomized version of the TTTT item (SD, D and N were coded as non-agreeing and A and SA were coded as agreeing). Comparable results were found as in the previous analysis. For the two-class model, the sample-size adjusted BIC was 23746.5 and the entropy was 0.89. In addition, the average latent class probabilities for most likely latent class membership were reasonable (see Table 5). As above, the first latent class generally disagreed with the need to TTTT and the second class indicated agreement. Taken together, there is reasonable support for the presence of 2 possible latent classes. Next, the individual levels of the TTTT item were examined with respect to latent class membership (see Table 6). These findings also support the concept of a consensus measure with underestimated values on the off-diagonal and overestimated values on the main diagonal. As would be expected following the previous analysis, this table also demonstrates a strong relationship between response pattern on the TTTT item and latent class assignment with $\chi^2(1) = 137.7$, $p < 0.001$.

While concerns about dichotomizing a variable are still present, the results above provide support for the validity of this item as a measure of communal agreement. In addition, as the latent class anal-

Author Name: Allen G. Harbaugh
Contact Email: allen.harbaugh@murdoch.edu.au

ysis was combined with other measures regarding pedagogy and curriculum, this further suggests that agreement of disagreement with the consensual belief regarding TTTT (the general nature of the latent classes) is related to the other measures. As such, there is sufficient justification for the exploration of possible relationships with the TTTT or moderating effects associated with TTTT.

Table 5
Average Latent Class Probabilities for Most Likely Latent Class Membership by Latent Class (TTTT Dichotomized)

Latent Class Assignment	**Latent Class Probability**	
	Class #1	Class #2
Class #1	91.7%	8.3%
Class #2	2.7%	97.3%

Table 6
Contingency Table for TTTT Item Responses and Latent Class Membership (TTTT dichotomized)

Latent Class	D	A	TOTAL
disagree	71 (23.7) *9.73*	62 (109.3) *–4.53*	152
agree	73 (120.3) *–4.31*	603 (555.7) *2.01*	657

Note: Expected frequencies indicated in parentheses; italicized numbers indicate signed Pearson residuals.

The first post hoc analysis to be conducted with this item is a clarification of the evidence observed. Namely, the strength of the cross-loadings of the TTTT item when included in the model suggests that there is a relationship with the constructs. Furthermore, as there was not perfect discrimination with the LCA when examining the TTTT item with the curriculum and pedagogy constructs, there is reason to explore the possible relationship more carefully. This is because membership in one group did not completely determine the response on the TTTT item. Thus, there are additional differences between the groups. These are most likely attributable to the other constructs used in the classification analysis. Fortunately, M*plus* conducts a latent mean analysis (LMA) when reporting the LCA results. With the agreeing class used as the reference group (average latent means set to zero), significant differences were observed for the latent means for the disagreeing class. The average difference for the pedagogy construct was 0.255 ($p = 0.001$), and the average difference for the curriculum construct was 2.171 ($p < 0.001$). While these are sample-specific latent scales, there is evidence to suggest that in this sample, different response patterns were observed depending on whether the teacher agreed with the need to TTTT or not.

Suggestions: Practical Application and Future Research

The practical application of this approach is that researchers should explore the potential merit of using dropped measurement items for additional analyses. The principal caveat is that the goal of this exploration is for theory-building purposes. With this in mind, it might prove useful for other areas in education research and other disciplines in the social sciences where survey research is utilized to understand complex social phenomena and build-theory to suggest explanations for those phenomena. In particular, the exploration of covariation, influence of other factors after accounting for this covariation or moderation effects from categorizing variables would be useful in many areas of research relying on factor analytic techniques.

In addition to the strategies presented here, there may be other exploratory approaches that are available when considering more than one variable at a time. For example, it may be possible to use cluster analysis to distinguish groups from dropped measurement items. One possible arena for this would be if two survey items were dropped. If these two items demonstrates a mixed correlational

Author Name: Allen G. Harbaugh
Contact Email: allen.harbaugh@murdoch.edu.au

pattern,[7] it may be possible that the different clusters demonstrate different relationships with other variables measured. An analysis such as this could be conducted with CART, LCA or traditional hierarchical clustering protocols (Bartholomew et al., 2002; Ma, 2005; Pastor, Barron, Miller & Davis, 2007).

Additional research specific to the methodological approach as suggested here could address the power to detect differences in categorical or dichotomized variables. This could be accomplished using simulation studies in which latent classes are and are not related to specific variables used in the CFA. From this, it would be possible to explore how effective LCA is at discriminating the clusters in relation to the main variable under examination and in concert with other measured variables.

Finally, research could be conducted into the scenarios in which the lack of fit in the measurement model is a non-addressable issue. In such cases, the item would be seen as theoretically part of the model but not useable (for the measurement model or other analyses) because the data is faulty due to other issues. For example, if patterns of responses degraded over the time taken to complete a survey, the last few items of the survey may not actually measure anything (Barnette, 1999). On the other hand, there may be instances where additional explorations indicate that an alternate measurement model may better align to the data obtained. For example, point-inflation in 11-point Likert-type scales is when a larger portion of individuals respond to the middle or extreme responses (this is comparable to zero-point inflated Poisson models [Simonoff, 2003]). This response pattern may violate the normality assumptions for maximum likelihood CFA, but it may be possible to use LCA to classify the participants into different response patterns (scale usage). With these different classes, a comparable measurement model (in configuration) with better fit to the data may be achieved. (To our best assessment of the quantitative methodological literature, this has not been explored.)

Conclusion

The intent of this paper was to demonstrate how survey items dropped from a measurement model might be retained for additional analyses. By the very nature of this processing not driven by a hypothetical model, this type of exploration would be exploratory in nature. In addition, there are a number of key issues that should be examined when choosing to retain dropped measurement items. These are mainly statistical in nature, but the paramount concern should be that the retained items are related to theory and provide explorations that assist the research in the theory-building process. Furthermore, this paper also demonstrated additional analyses that can be used to check the viability of retained items for further analysis. This was demonstrated here using LCA, but comparable strategies following a similar theme could also serve a similar purpose. It is hoped that other researchers will be motivated to explore their data in similar and more innovative ways.

References

Barnette, J. J. (1999). Nonattending respondent effects on internal consistency of self-administered surveys: A Monte Carlo simulation study. *Educational and Psychological Measurement, 59*(1), 38–46.

Bartholomew, D.J., Steele, F., Moustaki, I., & Galbraith, J.I. (2002). *The Analysis and Interpretation of Multivariate Data for Social Scientists.* Boca Raton, FL: Chapman & Hall.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* New York, NY: John Wiley & Sons.

Byrne, B. M. (2010). *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Dawon, N. V., & Weiss, R. (2012). Dichotomizing continuous variables in statistical analysis: A

[7] In this context, a mixed correlational pattern would be where most of the sample follows a positive relationship (ranging from both extremes), but a portion follows one end of the negative relationship (say high on the first variable and low on the other) but very few follow demonstrate the other pattern (low on first variable and high on the other.

Author Name: Allen G. Harbaugh
Contact Email: allen.harbaugh@murdoch.edu.au

practice to avoid. *Medical Decision Making, 32*(2), 225-226. doi:10.1177/0272989X12437605

DeCoster, J., Iselin, A.-M. R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods, 14*(4), 349–366. doi:10.1037/a0016956

Harbaugh, A. G., & Thompson, G. (2013, April). The effects of NAPLAN: Teachers' perceptions of the impact on workplace stress and curriculum and pedagogy mediated by socioeconomic status. Paper presented at the *American Educational Research Association Annual Conference*, San Francisco, CA.

Harbaugh, A. G., Thompson, G., & Sproul, J. (2013, December). An Exploration of the Impact of NAPLAN: Relationships Among Teachers' Achievement Goal Orientations and Teachers' Perceptions of Impact on Curriculum and Pedagogy. Paper presented at the *Australian Association for Research in Education*, Adelaide, SA, Australia.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (3rd ed.). New York, NY: Guilford Press.

Lingard, B., Hayes, D., & Mills, M. (2003). Teachers and productive pedagogies: Contextualising, conceptualising, utilising. *Pedagogy, Culture & Society , 11*(3), 399–423.

Ma, X. (2005). Growth in mathematics achievement: Analysis with classification and regression trees. *The Journal of Educational Research, 99*(2), 78–86. doi:10.3200/JOER.99.2.78-86

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40. doi:10.1037/1082-989X.7.1.19

Midgley, C., Kaplan, A., Middleton, M., Maehr, M. L., Urdan, T., Andreman, L. H., Anderman, E., & Roeser, R. (1998). The development and validation of scales assessing students' achievement goal orientations. *Contemporary Educational Psychology, 23*(3), 113–131. doi:10.1006/ceps.1998.0965

Muller, R. & May, R. (Writers), & Roemer, L, & Nagashima, K. (Directors). (1964/2010). *Rudolph the Red-Nosed Reindeer* [Animated television special]. United States: DreamWorks Animation.

Muthén, L. K., & Muthén, B. O. (1998–2012). M*plus user's guide.* Los Angeles, CA: Muthén & Muthén.

Pastor, D. A., Barron, K. E., Miller, B. J., & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology, 32*(1) 8–47 doi:10.1016/j.cedpsych.2006.10.003

Simonoff, J. S. (2003). *Analyzing categorical data*. New York, NY: Springer.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837–847.

Thompson, G. (2012). *The Effects of NAPLAN: Executive Summary.* Perth, WA, Australia: Murdoch University.

Thompson, G., & Harbaugh, A. G. (2013). A preliminary analysis of teacher perceptions of the effects of NAPLAN on pedagogy and curriculum. *Australian Educational Researcher.* doi:10.1007/s13384-013-0093-0

Troia, G. A., Harbaugh, A. G., Shankland, R. K., Wolbers, K. A., & Lawrence, A. M. (2013). Relationships between writing motivation, writing activity, and writing performance: Effects of grade, gender, and ability. *Reading and Writing, 26*(1), 17–44. doi:10.1007/s11145-012-9379-2