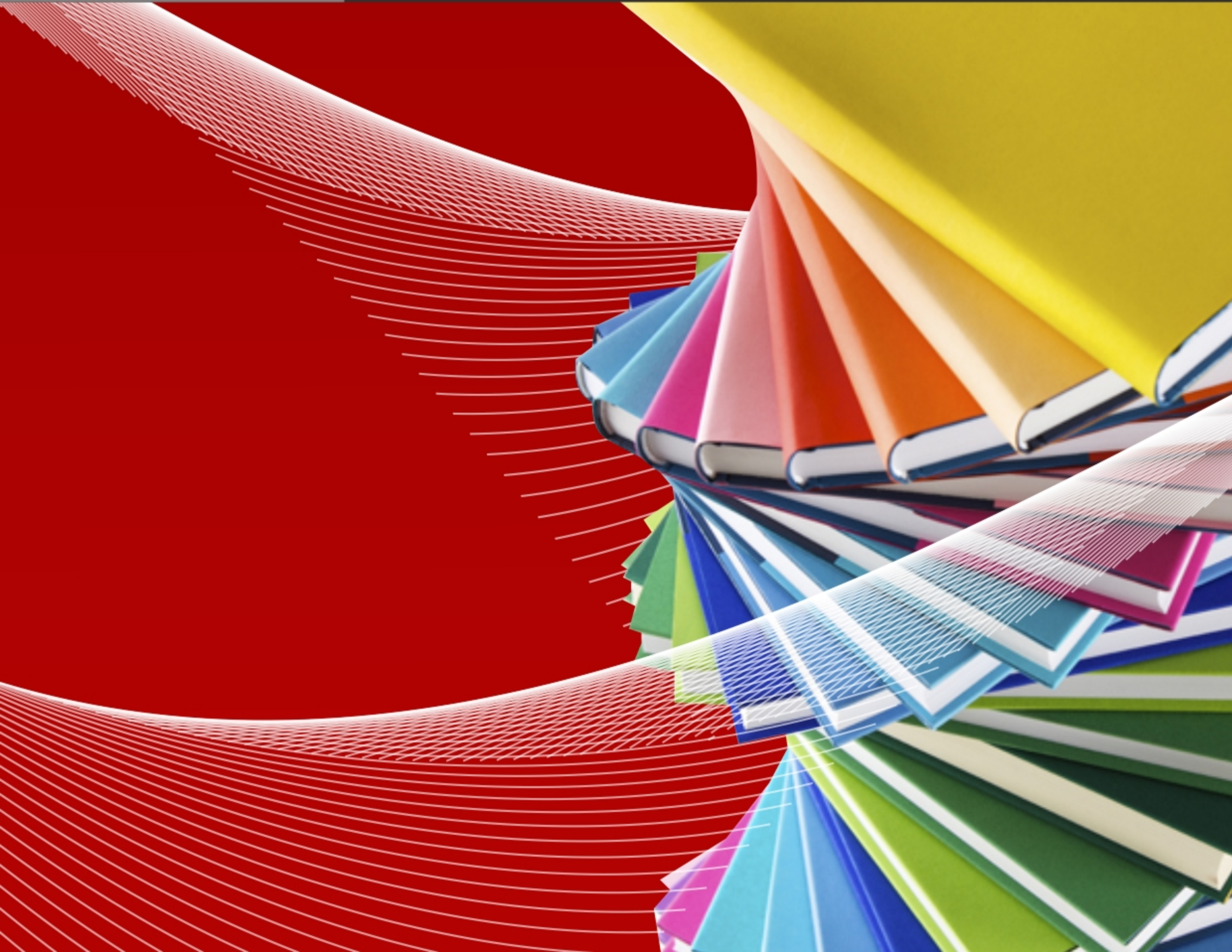
澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Outstanding Academic Papers by Students
學生優秀作品



Selection Errors in Scale Development: A Monte Study of Factor Analysis

Mark H. C. Lai

Supervisor: S. F. Cheung

University of Macau

UNIVERSIDADE DE MACAU
仁義禮知信
澳門大學

Abstract

Previous research had devoted a lot to the practice of exploratory and confirmatory factor analysis. Despite the large volume of reviews and simulations, little had been devoted to the decision of selecting and discarding items in scale development. The present research examined the impact of sample size, number of items, number of factors, ratio of strong indicators to weak indicators (i.e. items having strong / weak associations with the latent construct), and magnitude of weak loadings on the rates of occurrence of selection errors. A Monte Carlo simulation with a 6 × 3 × 2 × 3 × 3 design was conducted, and the real life practice of factor analysis in scale was examined. Results showed that when sample size was not large, the selection errors were not negligible, and selection errors also lead to reductions in population reliability. The utility of fit-indexes in identifying selection errors was also examined, with the model χ^2 being relatively more informative. It was also found that a large sample confirmatory factor analysis did not compensate for the instability of a small sample exploratory factor analysis. Some suggestions regarding sample size and procedures in scale development were discussed, with a general urge for more conservative procedure and careful sample size planning.

UNIVERSIDADE DE MACAU
澳門大學

Selection Errors in Scale Development: A Monte Carlo Study of Factor Analysis

From the time when Spearman (1904; 1920) developed it, factor analysis became an essential technique for scientific psychology. Though Spearman himself particularly used the technique to support his theory of intelligence, now factor analysis had been applied more broadly to the analysis of other constructs, and to the formulation of psychological scales (e.g. optimism by Scheier, Carver, & Bridges, 1994; depression by Radloff, 1977). As Nunnally and Bernstein (1994) and many others claimed, measurement scales are the fundamentals of psychology, since they allow psychological constructs like intelligence and personality to be studied in a scientific way. Most of the discussions on the process of scale development identified factor analysis as an important step (e.g. DeVellis, 2003; Fabrigar, Wegener, MacCallum, & Strahan, 1999), which confirmed its importance. With a hundred years of development, the technique became more refined and improved in accuracy (Cudeck & MacCallum, 2007, Chapter 2). However, in the literature there was relatively little discussion on the item selection decision. With the terminology of signal detection theory (Tanner and Swets, 1954), there are two kinds of errors in selecting items: *false alarm* (falsely including an item with weak relations[1] with the underlying factors) and *miss* (falsely excluding an item with strong relations with the underlying factors). The present study examined the stability of both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) regarding these selection errors within the framework of scale development.

Steps for Scale Development

I would first establish a general framework for the scale development process. Different sources (e.g. DeVellis, 2003; Worthington & Whittaker, 2006) presented slightly different conceptual models, but most of them would include the following steps:

1. **Item generation.** An item pool is generated, usually by means of some literature reviews, experts' opinions, or in-depth interviews.

2. **Item screening.** Items are subjected to a preliminary test to identify and eliminate items that are ambiguous or obscure in meaning, and those that overlap with other items.

3. **Item Selection.** The pool of items, after revision for wordings, is administered to a sample of participants. Usually factor analysis is used to analyze the scale and select items.

4. **Replication.** The retained items after the previous stage are usually administered to another sample. Their responses then are often analyzed to determine whether the scale still performs as intended in this second sample. Item selection or deletion occasionally takes place at this stage (e.g. Noone, Stephens, & Alpass, 2010).

5. **Testing for validity and other properties.** Sometimes the scale is further tested in other samples for determining its predictive or discriminative power using certain criterion constructs or for its normative data in certain populations (see DeVellis, 2003, Chapter 4; Nunnally & Bernstein, 1994, Chapter 3, for detailed discussions).

6. **Scale revision and short form.** In later studies, when evidence shows that the scale is suboptimal, some items in the scale are replaced. Occasionally some long scales are reduced in number of items (e.g. mini-IPIP by Donnellan, Oswald, & Baird, 2006) for research purposes. While the conceptual importance of the items will be of major concern in selecting items to the short form, factor analysis still acts as an important tool for the decision.

For the objectives of the present study, I exclusively deal with Step 3 and Step 4 (and the topics related to selection could also be generalized to Step 6 for selecting items for short form).

Impacts of False Alarms, Misses, and Misspecifications on the Selection Process

When weak indicators[2], items having only weak associations with all the factors of the

UNIVERSIDADE DE MACAU
仁義禮知信
澳門大學

construct, are retained after the selection, the error of *false alarm* is committed. When strong indicators, items having strong associations with at least one factor of the scale, are excluded, it is a *miss*. This study included a third type of error: When *strong* indicators are selected correctly but misattributed to be an indicator of another factor, the error of *misspecification* is committed[3]. For parsimonious concerns, often researchers would try to limit the length of their scales by selecting only a certain number of strong indicators on each factor (DeVellis, 2003). If a weak indicator replaced a strong indicator during the selection, the reliability and validity of the scale in the population may be attenuated, especially when the number of selected items is small and the difference between strong indicators and weak indicators is large (see Appendix B). In addition, as later section discussed, the problem of misspecification is probably more severe because it represented an incorrect structure of the construct captured by the scale.

Review on Factor Analyses

Factor analysis is a statistical technique to describe the variability of a set of variables with fewer latent variables, or factors (Kerlinger & Lee, 2000). To illustrate, consider the example of the well-known Center for Epidemiologic Studies Depression Scale (CES-D Scale; Radloff, 1977). The scores given by a sample of participants on 20 depression-related items were factor-analyzed and grouped into four factors, namely *depressed affect*, *positive affect*, *somatic and retarded activity*, and *interpersonal*. Each item had varying degrees of associations with the factors. For example in the original sample for developing the CES-D, the item “Happy” had about 44% (corresponded to the squared value of the standardized regression coefficient) of its variance explained by the factor *positive affect*. Thus, factor analysis allows researchers to represent a collection of items with a more parsimonious set of latent variables which are more easily handled (Fabrigar et al., 1999; for more understanding of factor analysis, refer to

Thompson, 2004). The latent variables also aid understanding of the construct being studied.

EFA

Researchers generally distinguished two types of factor analyses: exploratory and confirmatory. As the name suggested, EFA aims to explore the factor structure, including the number of factors to be extracted and the grouping of the items, of a set of items without the need for an a priori theory, as in the case of developing the CES-D (Radloff, 1977)[4]. Previous research focused a lot on factor extraction methods, ways to determine number of factors, and procedures of estimating the regression weights on an item when using the latent factors as the predictors. The latter one is denoted as *factor loading* (or simply *loading*) in the present study (also called pattern coefficients, see Henson & Roberts, 2006; Thompson, 2004).

The minimum sample size for EFA to be stable was also the focus of previous research. Whereas earlier scholars suggested a minimum of 150 for large number of indicators and 300 for small number of indicators (Guadagnoli & Velicer, 1988) or a participant-indicator ratio of 10 to 1 (Nunnally, 1978), recently researchers seemed to agree that sample size requirement could not be determined without concerns of other factors (Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005). For instance, MacCallum, Widaman, Zhang, and Hong (1999) suggested that strong loadings and large communality coefficients (i.e. the proportion of an item's variances accounted for by all the factors extracted) could make the solution more stable. In other words, under these conditions, a smaller sample could be adequate (de Winter, Dodou, & Wieringa, 2009).

Despite the huge volume of research on the best practice of EFA (e.g. Conway & Huffcutt, 2003; Ford, MacCallum, & Tait, 1986; Gorsuch, 1997; Hurley et al., 1997; Kahn, 2006; Worthington & Whittaker, 2006), the item selection process was relatively understudied. In the context of scale development, a researcher generally devises a large pool of items, collects responses in a sample, and uses EFA to determine the factor structure and the items to be included in the scale. Some items in the pool have strong estimated loadings, while others have weak ones. The latters are usually removed so that researchers can get a more parsimonious scale based on the magnitude and pattern of factor loadings. Typically research used what Hogarty, Kromrey, Ferron, and Hines (2004) called the “traditional approach”, in which a pre-determined cutoff value is set, and items with loadings on all factors lower than the cutoff are discarded (see also Osborne et al., 2008). However, there are some limitations in this approach.

Hogarty et al. (2004) pointed out that the cutoff value was usually arbitrarily set. A brief survey conducted by the author on 23 recent articles where a new scale was developed indicated that some researchers used a cutoff value of .3, some .5, and some used as high as .75 (see Appendix B). This result was consistent with the report from Osborne et al. (2008). Yet, seldom did researchers justify the chosen values or refer to established guidelines (as reported by Henson & Roberts, 2006, the median cutoff used was .4 for psychological journals). Even if they did provide some references, the reason for choosing that specific guideline was lacking. For example, Nunnally (1978) suggested .3 (which he assumed to be generally sufficient for a loading to be significantly different from zero), while Tabachnick and Fidell (2007) and Worthington and Whittaker (2006) suggested .32 (which means that 10% of the variances is explained by the factor). Due to this inconsistency, the same pool of items might give very different set of selected items using different cutoff values.

Also, as the estimation of the loadings is prone to sampling errors when sample size is not large enough (Sass, 2010), the item selection process may contain false alarm or miss. Hogarty et al. (2005)'s and Hogarty et al. (2004)'s research suggested that the latter was more common than the former. In their Monte Carlo studies, Hogarty et al. (2005) and Hogarty et al.

(2004) concluded that this approach performed well in selecting the strong indicators. Though I agreed with their conclusion, its generalizability might be limited. In Hogarty et al. (2004), the ratio of weak indicators to strong indicators ranged from 0.02 to 0.33; however, my review on recent articles suggested that the proportion of weak indicators in the pool of items may be larger. For instance, in Labbe and Maisto (2010), the number of items that did not reach the cutoff in EFA was half of the number of final selection (a ratio of 0.53, with 9 discarded and 17 selected); whereas in Wright, Creed, and Zimmer-Gembeck (2010), 32 items did not meet the cutoff and only 17 were selected (a ratio of 1.88). This suggested the plausibility of a higher proportion of weak indicators in real life than those tested in Hogarty et al.'s study.

Lastly, the result in Hogarty et al. (2004) also did not take into account the existence of *cross-loadings*, that is, an item’s loadings on factors other than the one with which the item has strongest associations. Usually researchers also set a cutoff for cross-loadings so that items that do not clearly belong to one factor are discarded (see Appendix B; Osborne et al., 2008).

The present study extended the study of Hogarty et al. (2004) to a context closer to the real practice in scale development. I examined the rates of occurrence of false alarms, misses, and misspecifications through simulations, under conditions of different sample sizes to evaluate the sample size requirement.

CFA

Recognizing the limitations of EFA, many researchers now used CFA to examine whether the factor structure suggested by EFA could be replicated in another sample. While CFA could refer to different techniques in the literature, in the present study it referred to the common practice of defining a *simple structure* of a scale (i.e., assuming every items loaded on only one factor; see McDonald, 1985) and used structural equation modeling (SEM) to estimate the

UNIVERSIDADE DE MACAU 澳門大學

loadings and assess whether the defined structure is acceptable in light of the data. Methodological issues associated with CFA shared those with SEM, like the performance of different families of fit-indexes and post hoc modification of the pre-defined structure (Sörbom, 1989). In addition, the effect of sample size on different indices had also been studied (Anderson & Gerbing, 1984; Gerbing & Anderson, 1985; Jackson, 2001; MacCallum, Widaman, Preacher, & Hong, 2001). For example, Fan, Thompson, and Wang (1999) found that the proportion of variances of different SEM fit-indexes by sample size ranged from less than 1% to 10%.

While there were a large volume of studies on the performance of SEM models in general, the individual items received less attention. Yet in the stage of scale development, both are essential, particularly when one recognizes the possibility of false alarms in CFA. Kahn (2006) shared the same opinion that one objective of CFA should be to test hypotheses that the item loadings were non-zero, yet in practice not very often did researcher report the significance level for individual loadings (see Labbe & Maisto, 2010; Wright et al., 2010; for examples where significance level were not reported). Because CFA estimation is also subject to sampling error, test for loadings may be able to detect items with zero loadings. Unfortunately, the goal of scale development usually is not restricted to finding items that have non-zero relations with the factors; rather, researchers want a parsimonious set of strong indicators. Thus, the logic of hypothesis testing may not be sufficient to fulfill this purpose, that is, to exclude those weak indicators with small non-zero loadings.

Whereas ways to detect false alarm needs further exploration, similarly in some situations miss also deserves attentions. Consider a case when the researcher, after carefully taking into account different concerns, decided that the final sets of items should contain two factors and in each factor exactly seven items. After the scale had been developed, the researcher identified a weak indicator in the scale which, for psychometric reasons, should be replaced by one with a stronger loading. In such occasions researchers usually would devise a new item to fill the gap (e.g. Scheier et al., 1994 in the revision of the Life Orientation Test). However, he or she could also choose to be more conservative before discarding items to save the cost in later replacement. In this sense the investigation on the occurrences of miss is relevant.

Compared to false alarms and misses, it may be easier to identify misspecifications. In the case of false alarm, if an item has a small but non-zero loadings with the factor, it still represents a true model structure; while for miss, as the item has been completely removed, researchers has no hint for this omission as it is not present in the covariance matrix (which is different from constraining it to zero). On the other hand, when an item loads on a wrong factor, the implied model puts constrain to the path between the item and the correct factor to zero, resulting in a model with poor fit. Thus, the use fit indexes should be able to identify item misspecification.

Besides studying the performances of different fit-indexes, this study also addresses the practical problem of sample size allocation between EFA and CFA. As it was usually more favorable to use two independent samples for EFA and CFA (to compensate for sampling errors), with the constraint of resources researchers would like to know whether they should have a larger sample for EFA, for CFA, or have the two samples equal in size. While Cheung (2009) discussed that the exploratory sample should not be too small, similarly there were certain sample size requirement for CFA to be stable. Thus I also compare the quality of item selections between different combinations of sample sizes.

In summary, the purpose of the present study is to evaluate the traditional approach of scale development in using EFA and CFA. Through simulations, I addressed particular questions including (a) the prevalence of selection errors in EFA; (b) whether the minimum sample size suggested by other scholars (e.g. MacCallum et al., 1999; Nunnally, 1978) is enough for an acceptable level of false alarms and misses; and (c) whether the CFA fit-indexes and testing of loadings could identify the selection errors in the EFA solution.

Method

A Monte Carlo study was conducted to test the performance of EFA and CFA across manipulated levels of five variables: (a) sample size, (b) number of factors, (c) number of strong indicators per factor, (d) proportion of weak indicators, and (e) magnitude of loadings for weak indicators. The statistical program R with version 2.12.2 (R Development Core Team, 2010) was used for all data analyses throughout the study.

Design

The simulation began by defining a population correlation matrix between all items using the *psych* package for R (Revelle, 2010). This correlation matrix was created by specifying the loadings of all items, both for strong indicators and weak indicators, on their primary factors (i.e., the factor to which they belong) and the interfactor correlations. Simulated sample correlation matrices were generated under the assumption of normality.

EFA Selection. The simulated sample correlation matrixes were then factor analyzed with principal axis factoring and promax rotation. Principal axis factoring and promax rotation were generally better than or at least as good as other alternatives (Fabrigar et al., 1999). The true number of factors to be extracted was specified, which could eliminate the bias of extracting an incorrect number of factors on the EFA solution (see Fava & Velicer, 1992; Fava & Velicer, 1996) so that the relation between sampling error and the selection error rates could be more clearly demonstrated. For each sample, then, following the suggestions by Worthington and

Whittaker (2006), items having estimated loadings lower than .32 on all factors were deleted. Besides, items were deleted if the difference between the absolute values of its strongest estimated loadings and its second strongest estimated loadings was smaller than .15. These selection criteria were chosen because they took into account cross-loadings, which was quite a common practice in the field (Appendix B). After the deletion, sample EFA solutions were compared with the population condition to determine the prevalence of selection errors.

CFA Model Fit. Next, same number of samples of raw data (rather than correlation matrices) as in the EFA stage were generated to simulate the replications in CFA, and each was matched with one EFA sample. For each pair, CFA was performed on the second sample to test the simple factor structure (with all cross-loadings and error covariances fixed to zero) implied by the EFA sample with the maximum iteration equals to 999. The maximum likelihood estimation in the *sem* package (Fox, 2010) of R uses the two-stage least squares procedure for fitting the model with an inputted covariance matrix, which provides similar results as in LISREL (Jöreskog, & Sörbom, 1996). The fit-indexes examined for the present study included the Goodness-of-Fit Index (GFI), Adjusted Goodness-of-Fit Index (Jöreskog & Sörbom, 1996), the root mean square error of approximation (RMSEA; Steiger, 1990), the Bentler-Bonett's (1980) Normed Fit Index (NFI), the Nonnormed Fit Index (NNFI; Bentler & Bonett, 1980; Tucker & Lewis, 1973), the Comparative Fit Index (Bentler, 1990), the standardized root mean squared residual (SRMR; Bentler, 1995), and the model χ^2.

CFA Selection. Another round of selection was performed based on the estimated loadings in the CFA solution. There are two common practices regarding CFA selection, which include assessing modification indexes (Sörbom, 1989) and examining the significance levels of the loadings. The first one had been challenged by MacCallum, Roznowski, and Necowitz

(1992) for directing to misleading decisions, while for the second its ineffectiveness in identifying weak indicators were dubious (see previous discussion of this report). As the goal was to include only the strong indicators, in this study the CFA selection retained items with the squared value of the unstandardized loading larger than 10% ($.32^2$) of the estimated item variance at a statistical significance level of .05, and discarded the otherwise[5].

Manipulated and Constant Parameters and the Symbols Used

Sample size for EFA (N_E). As a major concern of the present study, N_E was manipulated at four levels (100, 200, 400, 800). These levels were higher than those used by MacCallum et al. (1999) and were consistent with Jackson (2001). The level of 50 was not used here because many studies suggested that it was not enough for larger number of factors where average loadings were not high (de Winter et al., 2009).

Sample size for CFA (N_C). Levels of N_C were identical to those of N_E. However, for ease of communication, I only examine the four conditions with N_C equal N_E, plus the conditions with a large EFA sample (N_E = 800) with a small CFA sample (N_C = 100), and also a small EFA sample (N_E = 100) with a large CFA sample (N_C = 800). The conditions with equal sizes of EFA and CFA samples were typical in research practice (Appendix B). The two unbalanced conditions examined whether a large N_C could compensate for a small N_E, or vice versa, and the results could be useful for researchers when planning the sample sizes for scale development.

Number of factors (*f*). As indicated by different research (de Winter et al. 2009; MacCallum et al. 1999; Mundfrom, Shaw, & Ke, 2005), *f* could influence the stability of the EFA solution. Particularly, higher *f* corresponded to lower stability, which implies that the selection in EFA would be worse when the number of factors increases. Henson and Roberts (2006) reported that the number of factors extracted in psychological research ranged from 1 to 7, with a median of 3. Thus, in the present study two levels of *f* were employed (3, 5, and 7).

Number of items per factor (*p*/*f*), number of strong indicators per factor (p_S/*f*), and ratio of weak indicators to strong indicators (p_W/p_S). Past research had shown that *p*/*f* had impact on both EFA and CFA. For example, Mundfrom et al. (2005) showed that the stability of EFA increased when *p* increased, and its influence is stronger than absolute sample size or communality level. For CFA, research also suggested that increase in *p* could stabilize the sample estimation (Marsh, Hau, Balla, and Grayson 1998; Jackson, 2001; Velicer & Fava, 1998). In the present study, there were different combinations of p_S and p_W in the simulation. Specifically, p_S/f had two levels: 3 and 6, each with three corresponding levels of the ratio p_W/p_S: 0, or no weak indicators; 1, and 2. Thus there were in total six combinations of strong and weak indicators in the population. The choice for p_W/p_S was believed to be typical for applied research (Appendix B). Combining p_S and p_W, *p*/*f* ranged from 3 (3 strong indicators per factor with no weak indicators) to 18 (6 strong indicators per factor with 12 weak indicators), which was similar to previous research (Hogarty et al., 2005; MacCallum et al., 1999; both used 10/3 as the minimum), and also covered the maximum of 16 in the review by Henson and Roberts (2006) of psychological research using factor analyses.

Magnitude of loadings for strong indicators (λ_S) and weak indicators (λ_W). Unlike previous studies (e.g. Mundfrom et al., 2005), strong indicators in the present study did not assume a homogeneous value of loadings, but descended from .8 to .5 proportionally. For example, in conditions with six strong indicators, strong indicators had loadings of .80, .74, .68, .62, .56, and .50 respectively. This corresponded to the moderate to high range in other simulation studies (Hogarty et al., 2004; MacCallum et al., 1999). For λ_W, similar to Hogarty et al.'s (2004) criteria, three levels were used (.1, .2, and .3) to assess the effect of different degrees of weak indicators to the solution.

Table 1 shows a summary for the manipulated parameters. In brief, the Monte Carlo study used a 6 (N_E and N_C) × 3 (f) × 2 (p_S/f) × 3 (p_W/p_S) × 3 (λ_W) design. However, the extreme conditions with the number of items larger than N_E were not simulated. These included six conditions with $f = 7$, $p_S/f = 6$, $p_W/p_S = 2$, and $N_E = 100$. In addition, for conditions with no weak indicators, there were no variations in weak loadings. Thus, the actual number of conditions in the present study was 246. In all conditions, inter-correlations between factors were fixed to .3, which was a typical value shown by Sass (2010) to produce a relatively accurate EFA solution for promax rotation. The number of replications in each condition was 500.

Data Analyses After EFA

Descriptive analyses of EFA selection errors. Four indicators of the simulated data were computed: (a) *false alarm rate*-the percentage of replications in which at least one weak indicator is falsely included; (b) *miss rate*-the percentage of replications in which at least one strong indicator is falsely excluded; and (c) *misspecification rate*-the percentage of replications in which at least one strong indicator have its highest estimated loadings loaded on a factor different from the one that the indicator was supposed to load in the population. Central tendency of selection errors would be reported with regard to different manipulated variables.

Selection implied population reliability coefficients. To deduce the impact of selection errors, the population reliability *for each factor* was first computed both for the selected items in EFA and those after CFA, using the population interitem correlations by the formula $\alpha = \frac{k\bar{\rho}}{1+(k-1)\bar{\rho}}$, where k ($k \geq 2$) equals to the number of items selected and $\bar{\rho}$ the average value of the inter-correlations between the selected items. For cases where $k = 1$, the population reliability was calculated by the squared value of the factor loading[b]. This obtained reliability coefficients

UNIVERSIDADE DE MACAU 澳門大學

were then compared with the baseline value computed the correlation between only all strong indicators (α = .680 for $p_S/f = 3$; .813 for $p_S/f = 6$). This deviation from the baseline (denoted as$\Delta\rho_0$) were then analyzed by omnibus analysis of variances (ANOVA) to examine the effect of different manipulated variables and all possible two- or multi-way interaction between them, with the effect size measure of omega-squared (ω^2)[6].

Data Analyses After CFA

Utility of fit-indexes in identifying selection errors. The utility of CFA was investigated to see whether the model-fit could provide information for the prior EFA decision. Particularly, the eight fit-indexes reported in R were discussed. Following the traditional rule of thumb (Hu & Bentler, 1999), the cutoff for GFI, AGFI, NFI, NNFI, and CFI was .90, that is, values above .90 was regarded as good fit. For RMSEA, a value below .05 indicated a good fit; while for SRMR the value was .08 (see Hu & Bentler, 1998). For model χ^2 a significance level higher than .05 would be considered good. Comparing the results of EFA selection errors and CFA fit-indexes, two ratios were calculated.

Selection Error Sensitivity. The number of replications in which the EFA selection contained that error and the fit-indexes in CFA did *not* meet the conventional cutoff, divided by the total number of replications in which the EFA selection contained that error.

Overall Specificity. The number of replications in which the EFA selection did *not* contain any selection error and the fit-indexes in CFA *met* the conventional cutoff, divided by the total number of replications in which the EFA selection did *not* contain that error.

Comparisons of different sample size allocations. Lastly, I examined the change in false alarms, misses, and misspecifications after the CFA selection and tried to identify the effect of the manipulated variables. The deviation of the reliability coefficients from the baseline (only all strong indicators selected) after CFA, denoted as $\Delta\rho_1$, was compared with the one before CFA, denoted as $\Delta\rho_0$. Four planned contrasts were carried out with Wilcoxon paired sample test[7] to determine whether different distribution of EFA sample and CFA sample would make a difference in reliability. The four contrasts were:

1. $\Delta\rho_1$ ($N_E = 100$ & $N_C = 800$, *small-head*) versus $\Delta\rho_1$ ($N_E = 800$ & $N_C = 100$, *big-head*);
2. $\Delta\rho_1$ ($N_E = N_C = 100$, *small sample half-head*) versus $\Delta\rho_0$ ($N_E = 200$, *small sample all-head*);
3. $\Delta\rho_1$ ($N_E = N_C = 200$, *moderate sample half-head*) versus $\Delta\rho_0$ ($N_E = 400$, *moderate sample all-head*);
4. $\Delta\rho_1$ ($N_E = N_C = 400$, *large sample half-head*) versus $\Delta\rho_0$ ($N_E = 800$, *large sample all-head*).

Result

Selection Error in EFA

Three kinds of errors are identified in selecting and discarding items: false alarms, misses, and misspecifications. The error rates were reported in Table 2. As the results had a skewed distribution, the median were reported instead of the mean. Generally speaking, the most common error was the false alarms, followed by the misses, and misspecifications were found the least. Regardless of other factor, when λ_W is .3, which is very close to the cutoff applied in this study, most replications contains at least one false alarm in the selection (with a median value of 100). The median false alarm rate of conditions with a small sample ($N_E = 100$) regardless of other manipulated variables is higher than 95%. False alarm rate also increases with the increase of p_W/p_S ratio, f, and decreases with more strong indicators (p_S/f). Only when sample size is very large (N_E = 800) and when λ_W is not close to the cutoff would the false alarm rate generally be kept below 15%. The ratio of number of false alarms to the number of weak indicators follows a similar pattern as the false alarm rate.

The second selection error, miss, seems to occur less. Like false alarm, the miss rate rises up when f or p_W/p_S increases, and falls down when N_E, p_S/f or λ_W increases. When the sample size reaches 400, the miss rate seems no longer cause problems, with all median values lower than 2%. Lastly, replications with misspecification account for a significant proportion when the sample size is 100, particularly when the number of factors and the p_W/p_S ratio is high, and when λ_W and p_S/f are small. When N_E reaches 200, misspecification generally is small (< 2%).

Because in different conditions with the same sample size the total number of indicators in the item pool varies, it is important to take the participant-indicator ratio (before EFA took place) into account. The conditions are broken down into five groups similar to the one by Osborne et al. (2008): smaller than 3:1 (24.0%), between 3:1 to 5:1 (16.7%), between 5:1 to 10:1 (22.4%), between 10:1 to 20:1 (20.3%), and larger than 20:1 (16.67%). The percentage of conditions with the ratio higher than 10:1 in the present study is about the same as the one reported by Osborne et al. in 303 articles in PsycINFO (36.8%). As shown in Figure 1, even when the ratio is between 5:1 to 10:1, the median false alarm rate is still .6. For miss rate, when the ratio is at least 5:1 the median is less than 20%. For misspecification, most of the conditions would have a rate below 40% except for some conditions when the ratio reaches 3:1.

Reliability as Outcome in EFA

Deviation of reliability coefficient from the baseline ($\Delta\rho_0$) was used to further investigate the impact of these selection errors. As across conditions the number of strong indicators differed, $\Delta\rho_0$ was calculated by subtracting the baseline coefficient (.680 for p_S/f = 3, and .813

for $p_S/f = 6$) from the selection-implied one. As shown in Figure 2a, when both the sample size and the number of strong indicators are small, there is on average a .15 reduction (with a median $\Delta\rho_0$ of .11) in population reliability. With more strong indicators in the model, the reliability coefficient is more resistant to the negative effect of small sample size.

An omnibus ANOVA was performed to analyze the impact of the manipulated variables on the model reliability, with $\Delta\rho_0$ being the dependent variable. Because the sample size is unexceptionally high ($N = 126{,}000$), I concerned more with the ω^2 (see Table 3). N_E and the p_W/p_S ratio stood out to have the biggest main effects ($\omega^2 = .23$ and .09 respectively), indicating that reliability decreases less with a larger sample size or less weak indicators in the item pool relative to the strong indicators. Sample size also interacts with p_S/f, p_W/p_S ratio, and λ_W to influence $\Delta\rho_0$ (ω^2 for interaction effect = .77, .57, .57 respectively). As shown in Figure 2a, when p_S/f is high, the negative effect of small N_E decreases. In Figure 2b, again, with higher p_W/p_S ratio, it aggravates the negative effect of small N_E. Particularly, reliability coefficients are quite stable with no weak indicators in the sample pool, but not very stable in conditions with p_W/p_S ratio = 2, even when $N_E = 400$. Lastly, in Figure 2c, the negative effect of small N_E increases from λ_W equals .1 to λ_W equals .2. However, as λ_W becomes larger (to .3), the inter-item correlations increases and results in a smaller deviation of reliability from the ideal one.

Sensitivity and Specificity of Fit-Indexes to Selection Errors

Before studying the data from CFA, in some replications the CFA solution is not computable either because the solution did not converge, the covariance matrix was singular, or because the implied model after EFA selection was underidentified (e.g. when only one indicator was left for a factor). There are a total of 21 conditions with more than 10% of the replications not computable, and 20 are from conditions with $N_E = 100$. Particularly, in condition with $f = 7$,

$p_S/f = 3$, $pw/ps = 2$ and $\lambda_W = .1$, only 104 replications are complete. Analyses for the associations between different selection errors and CFA non-computability are shown in Appendix C. For subsequent analyses, only replications with the CFA solution available in R were used.

The major analyses for the selection error sensitivity and overall specificity of different fit-indexes are shown in Figure 3. In brief, the sensitivity of NFI was high (*M*s = 67%, 59%, and 62% respectively for false alarm, miss, and misspecification), but at the same time the overall specificity was among the lowest (*M* = 55%). Besides NFI, other fit-indexes produce a low sensitivity (for false alarm, 8% to 45%; for miss, 11% to 42%; for misspecification, 19% to 55%) and a relatively high mean specificity (from 65% for AGFI to 97% for SRMR) for all three errors. The significance level of the model χ^2 is the only one that with mean sensitivity higher than 30% on all errors (*M*s = 34%, 39%, and 55%) and overall specificity higher than 80%.

Selection Errors and Population Reliability of the Items after CFA

Table 4 shows the change in *number* of false alarms, misses, and misspecifications after running CFA and performing another item selection. As the mean value and the median value does not differ much (not more than .5), the mean values are reported. Generally, after the CFA selection, the number of false alarms decreases to a significant extent, while the number of misspecification also slightly decreases. Particularly, the large number of false alarms left by the EFA selection with large *f*, p_W/p_S, or λ_W is remedied. Nevertheless, the trade-off is the inflated number of misses, particularly when N_C is small (with an increase in number of misses by 3). To deduce the effect of CFA selection, I compared the population reliability coefficient of the selected items after CFA with the one for the items selected in EFA (which also equaled to the comparison between $\Delta\rho_0$ and $\Delta\rho_1$). After the CFA selection, the population reliability coefficient has a small increase with a mean value of .024 (across all conditions). In particular, when the original EFA sample is small ($N_E = 100$), the CFA selection improves the reliability of the selected items (by .026 to .070 for $N_C = 100$, and by .032 to .095 for $N_C = 800$, see Table 5) except for conditions with no weak indicators in the item pool, which shows nearly no change.

It seems that CFA does improve the scale development process. However, this improvement is based on the cost of an additional sample. Another interest of the present study was to determine, when the total sample size is constant, how different allocations to EFA and CFA would differ. The first comparison was between the one with a small N_E (100) and a big N_C (800), or the *small-head* approach, and the one with a big N_E (800) and a small N_C (100), or the *big-head* approach. The *big-head* approach is found to produce fewer cases with the population reliability coefficient greatly deviated from the baseline, particularly for conditions with high p_W/p_S, large f, small p_S, and small λ_W. In a paired-sample Wilcoxon test it is found that the *big-head* approach performed better with a statistical significance (median difference of 39 pairs = 0.02, 95%CI = [.006, .034], T = 130, $p < .001$).

Other comparisons were done between samples using the whole sample for EFA, or *all-head*, and those using half for EFA and half for CFA, or *half-head*. For a total sample size of 200, Wilcoxon test did not find significant differences between the two approaches, with the median difference between the reliability coefficients for *all-head* and for *half-head* being 0.007, 95%CI = [–.001, .016], T = 270, $p = .096$. Interestingly, with the increase in total sample size, the *half-head* approach outperforms the *all-head* approach, though the difference was small. At a total sample size of 400, the median difference for 42 pairs is –.005, 95%CI = [–.015, –.006], T = 630, $p = .002$; At 800, the median difference is –.002, 95%CI = [–.016, –.000], T = 785, $p < .001$.

Discussion

Results of EFA selection show that false alarm may be more than common, and that its prevalence rate is influenced by number of indicators as well as number of factors. Even when the sample size or the participant-indicator ratio is large, the false alarm rate is not ignorable (about 15%). Also, as shown in Table 2, the inclusion of more weak indicators increases the chance of getting a false alarm, which can be due to the decrease in participant-indicator ratio as well as the increase of the base rate of weak indicators. On the other hand, the inclusion of more strong indicators does not seem to reduce the false alarm rate a lot. The decrease of participant-indicator ratio may counter the effect of including more good quality items. In fact, when sample size is large ($N_E = 800$), p_S/f seems to make some differences (11.6 vs. 8.3 for median comparison, as shown in Table 2). Another factor that influences the false alarm rate is *f*, with higher *f* corresponds to higher false alarm rate.

While the falsely included strong indicators have the chance to be identified and deleted in follow-up replications, it is not the case for misses. Yet the reality is that researchers usually would need to have a parsimonious set of items with a good enough psychometric properties. Thus it is a dilemma whether to discard an item or not. Nevertheless, given the cost of re-creating an item and validating it, I would suggest one to be more conservative in making such a decision. As suggested by the present results, the miss rate is low (with a median value < 2%) across different conditions when the N_E reached 400. Finally, though maybe potentially dangerous, item misspecification is the least common among the three selection errors, and is the most identifiable by considering the model χ^2.

The population reliability was used as an indicator for the overall impact of the selection errors. From the mean plot and the ANOVA results, it is found that conditions with more strong indicators and less weak indicators are more stable and with a higher reliability. Certainly one reason for the low reliability is due to the lack of strong indicators to be selected and the chances of these items to be missed out, and the other reason would be the inflated false alarm rate with large p_W/p_S.

With respect to selection error, generally a participant-indicator ratio of 10 to 1 is favorable. However, the results should be interpreted with cautions, as the present study had not taken errors in determining number of factors into considerations. As the techniques for determining *f*, like parallel analyses, scree plot or bootstrapping each had their own limitations (cf. Thompson, 2004), in real practice the selection error rates could be higher. The required sample size also depends on other factors. As MacCallum et al.'s (1999) suggested, a smaller sample size could be enough if the item loadings were strong. Yet, this is usually not known prior to EFA, and as Figure 2c shows that when weak loadings are close to but lower than the cutoff value it may even result in lower reliability of the selected items. Thus, without evidence of how “good” the qualities of the items are (in terms of loadings), a larger sample is more adequate.

The introduction of CFA into the standard procedure of scale development in psychological research was probably a response to the selection errors (and also other biases like overextraction, underextraction) in EFA. In CFA or SEM in general, the model χ^2 and different fit-indexes were widely used in research. From the present result, NFI is the most successful among the eight in identifying selection errors (sensitivity) for all three decision errors but also leads to a high rejection of correct models (low in specificity). On the other hand, though the use of the model χ^2 is still under hot debate (Barrett, 2007), in the present context it nevertheless shows the highest utility, particularly for misspecification. A final remark is that as the present cutoff employed is just one example from the many different research practices, and if a stricter cutoff like .95 was used for NNFI instead of .90, its sensitivity might be increased while the specificity was still acceptable. Future research could consider using an approach similar to Hu and Bentler (1999) to evaluate the cutoffs of the fit-indexes.

Another function of CFA is to identify weak indicators and discard them. After selection, the reliability of the scale did improve, particularly if the previous EFA sample is small. Yet it also leads to more misses, and might result in a structure where a factor only includes three or fewer items. Though the problem was not handled here, some researchers may consider deleting the factor, which sometimes represents loss in conceptual integrity for the construct of interest. The CFA selection procedure, given its importance, needs further investigation in the future.

Finally, the comparison between the different allocations of sample sizes to EFA and CFA suggested that the utility of CFA is somehow limited. The comparison between the *big-head* approach and the *small-head* approach showed that a large sample CFA could not compensate for the errors in a small sample EFA in terms of reliability. On the other hand, despite its importance, the effect of the EFA sample size seemed to reach a ceiling after $N_E = 400$, as implied from the comparison of the *half-head* and *all-head* approach showing that the two approaches did not differ. Thus it seems that CFA played its role to improve the selection only on the basis that the previous EFA is large enough so that the selection errors are not causing large problems.

Limitations

Although the present research tried to capture and study the real practice in using factor analysis for scale development, as a Monte Carlo study it is not possible to model perfectly the things in real life. There were certain aspects that my simulation did not cover. For example, while Sass (2010) suggested that the magnitude of the interfactor correlation could influence the stability of the EFA solution and estimated loadings, the present study just chose a typical value of .3. Nevertheless, some construct may be composed of factors which are virtually uncorrelated, and some may have factors more highly correlated than .3. Under different levels the selection error rates, performance of CFA fit-indexes, and sample size requirements may change.

One should also be aware that the present simulations are based on simple factor structure. If the population factor structure were complex with cross-loadings besides the interfactor correlations (cf. Asparouhov & Muthén, 2009), all model specifications in CFA with a simple structure would lead to model misspecification. In the latter situation it would be hard to judge whether an item had been put to a wrong factor.

Suggestions for Using Factor Analysis

On the basis of the reported results in the present study and also the previous work by other researchers (cf. Brown, 2006; Conway & Huffcutt, 2003; Fabriegar et al., 1999; Henson & Roberts, 2006; Hahn, 2006; MacCallum et al., 2001; Osborne et al., 2008; Worthington & Whittaker, 2006), several suggestions for a better practice of EFA and CFA can be drawn, and these are shown in Table 6.

In summary, this study adopted a perspective of selection error to examine the performance of factor analyses, which I believed was practical to scale developers. Results showed that when sample size was small, when many factors were to be extracted, or when many weak indicators were present in the item-pool relative to the number of strong indicators, selection errors occurred at a non-negligible rates and lead to reduction in population reliability. Researchers should take different factors (quality of sample pool, complexity of the construct, etc) into account in doing sample size planning.

References

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*, 155–173. doi:10.1007/BF02294170

Asparouhov, T., & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling, 16*, 397–438. doi:10.1080/10705510903008204

Barrett, P. (2007). Structural equation modelling: adjudging model fit. *Personality and Individual Differences, 42*, 815–824. doi:10.1016/j.paid.2006.09.018

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246. doi:10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606. doi:10.1037//0033-2909.88.3.588

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guildford Press.

Cheung, M. W. L. (2009, Feb 17). *Common Myths (and Facts) in Data Analysis.* Powerpoint slides of workshop. Retrieved from http://courses.nus.edu.sg/course/psycwlm/internet/

Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods, 6*, 147–168. doi:10.1177/1094428103251541

Cudeck, R., & MacCallum, R. C. (Eds.). (2007). *Factor analysis at 100*. Mahwah, NJ: Lawrence

UNIVERSIDADE DE MACAU
仁義禮知信
澳門大學

Erlbaum Associates.

Cudeck, R., & O'Dell, L. L. (1994). Applications of standard error estimates in unrestricted factor analysis: significance tests for factor loadings and correlations. *Psychological Bulletin, 115*, 475–487. doi:10.1037//0033-2909.115.3.475

de Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research, 44*, 147–181. doi:10.1080/00273170902794206

DeVellis, R. F. (2003). *Scale development: theory and applications* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment, 18*, 192–203. doi:10.1037/1040-3590.18.2.192

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272–299. doi:10.1037//1082-989X.4.3.272

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation method, and model misspecification on structural equation modeling fit-indexes. *Structural Equation Modeling, 6*, 56–83. doi:10.1080/10705519909540119

Fava, J. L., & Velicer, W. F. (1992). The effects of overextraction on factor and component analyses. *Educational and Psychological Measurement, 56*, 907–929. doi:10.1207/s15327906mbr2703_5

Fava, J. L., & Velicer, W. F. (1996). The effects of underextraction on factor and component analysis. *Multivariate Behavioral Research, 27*, 387–415.

UNIVERSIDADE DE MACAU
澳門大學

doi:10.1177/0013164496056006001

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: a critical review and analysis. *Personnel Psychology, 39*, 291–314. doi:10.1111/j.1744-6570.1986.tb00583.x

Gerbing, D. W., & Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. *Multivariate Behavioral Research, 20*, 255–271. doi:10.1207/s15327906mbr2003_2

Gorsuch, R. L. (1997). Exploratory factor analysis: its role in item analysis. *Journal of Personality Assessment, 68*, 532–560. doi:10.1207/s15327752jpa6803_5

Guadagnoli, E., & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265–275. doi:10.1037//0033-2909.103.2.265

Haig, B. D. (2005). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research, 40*, 303–329. doi:10.1207/s15327906mbr4003_2

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*, 393–416. doi:10.1177/0013164405282485

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: the influence of sample size, communality, and overdetermination. *Educational and Psychologiocal Measurement, 65*, 202–226. doi:10.1177/0013164404267287

Hogarty, K. Y., Kromrey, J. D., Ferron, J. M., Hines, C. V. (2004). Selection of variables in exploratory factor analysis: an empirical comparison of a stepwise and traditional approach. *Psychometrika, 69*, 593–611. doi:10.1007/BF02289857

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424–453. doi:10.1037/1082-989X.3.4.424

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. doi:10.1080/10705519909540118

Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., & Williams, L. J. (1997). Exploratory and confirmatory factor analysis: guidelines, issues, and alternatives. *Journal of Organizational Behavior, 18*, 667–683. doi:10.1002/(SICI)1099-1379(199711)18:6<667::AID-JOB874>3.0.CO;2-T

Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: a monte carlo investigation. *Structural Equation Modeling, 8*, 205–223. doi:10.1207/S15328007SEM0802_3

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: user's reference guide*. Chicago: Scientific Software International.

Kahn, J. H. (2006). Factor analysis in counseling psychology research, training, and practice: principles, advances, and applications. *The Counseling Psychologist, 34*, 684–718. doi:10.1177/0011000006286347

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.). Fort Worth, TX: Harcourt College Publishers.

Labbe, A. K., & Maisto, S. A. (2010). Development of the Stimulant Medication Outcome Expectancies Questionnaire for college students. *Addictive Behaviors, 35*, 726–729.

doi:10.1016/j.addbeh.2010.03.010

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: when “good” indicators are bad and “bad” indicators are good. *Psychological Methods, 4*, 192–211. doi:10.1037/1082-989X.4.2.192

MacCallum, R. C. Roznowski, M., & Necowitz, L. B.(1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504. doi:10.1037//0033-2909.111.3.490

MacCallum, R. C., Widaman, K. F., Preacher, K. J., Hong, S. (2001). Sample size in factor analysis: the role of model error. *Multivariate Behavioral Research, 36*, 611–637. doi:10.1207/S15327906MBR3604_06

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor anlysis. *Psychological Methods, 4*, 84–99. doi:10.1037//1082-989X.4.1.84

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181–220. doi:10.1207/s15327906mbr3302_1

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

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5*, 159–168. doi:10.1207/s15327574ijt0502_4

Noone, J. H., Stephens, C., & Alpass, F. (2010). The process of retirement planning scale (PRePS): development and validation. *Psychological Assessment, 22*, 520–531. doi:10.1037/a0019512

UNIVERSIDADE DE MACAU 仁義禮知信 澳門大學

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Osborne, J. W., Costello, A. B., & Kellow, J. T. (2008). Best practices in exploratory factor analysis. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 86–99). Thousand Oaks, California: Sage Publications.

R Development Core Team (2010). R: a language and environment for statistical computing (reference index version 2.12.0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Raykov, T. (2007). Reliability if deleted, not 'alpha if deleted': evaluation of scale reliability following component deletion. *Structural Equation Modeling, 60*, 201–216. doi:10.1348/000711006X115954

Radloff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401. doi:10.1177/014662167700100306

Revelle, W. (2010). psych: procedures for personality and psychological research (version 1.0-90) [R package]. Evanston, Illinois: Northwestern University. Retrieved from http://personality-project.org/r

Sass, D. A. (2010). Factor loading estimation error and stability using exploratory factor analysis. *Educational and Psychological Measurement, 70*, 557–577. doi:10.1177/0013164409355695

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of Life Orientation Test. *Journal of Personality and Social Psychology, 67*, 1063–1078.

UNIVERSIDADE DE MACAU 澳門大學

doi:10.1037//0022-3514.67.6.1063

Sörbom, D. (1989). Model modification. *Psychometrika, 54*, 371–384. doi:10.1007/BF02294623

Spearman, C. (1904). “General intelligence,” objectively determined and measured. *American Journal of Psychology, 15*, 201–293. Retrieved from http://library.isb.edu/digital_collection/general_intelligence.pdf

Spearman, C. (1920). Manifold sub-theories of “the Two Factors”. *Psychological Review, 27*, 159–172. doi:10.1037/h0068562

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173–180. doi:10.1207/s15327906mbr2502_4

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (5th ed.). New York: Allyn and Bacon.

Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*, 401–409. doi:10.1037/h0058700

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: understanding concepts and applications*. Washington, DC: American Psychological Association.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1–10. doi:10.1007/BF02291170

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3*, 213–251. doi:10.1037/1082-989X.3.2.231

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: a content analysis and recommendations for best practices. *The Counseling Psychologist, 34*, 806–838. doi:10.1177/0011000006288127

Wright, M., Creed, P., & Zimmer-Gembeck, J. (2010). The development and initial validation of a brief daily hassles scale suitable for use with adolescents. *European Journal of Psychological Assessment, 26*, 220–226. doi:10.1027/1015-5759/a000029

UNIVERSIDADE DE MACAU
仁義禮知信
澳門大學

Footnotes

[1] The present study took a position different from the theoretically ideal. While the latter implied that all indicators with non-zero relations should be retained, as pointed out by Fabriger et al. (1999) the inclusion of items with low communalities affected the EFA estimation. Little, Lindenberger, and Nesselroade (1999) also suggested that if one has a strong theory for the understanding of the construct (which is a favorable situation for scale development), selecting fewer indicators with high communalities resembles more to the goal of parsimony. The present study followed Hogarty et al. (2004) and Hogarty et al. (2005) to exclude items with weak loadings in scale development.

[2] Here the term *indicators* and *items* are used interchangeably.

[3] When selected weak indicators are specified to a factor that they have no associations, the error is counted as *false alarm*.

[4] Sometimes researchers would have their theory in advance and skip the process of EFA and go directly to CFA. However, as argued by Haig (2005) and Hurley et al. (1997), the two techniques were usually complementary, with EFA generating specific hypotheses and CFA testing them.

[5] More often researchers preferred to use significant test to determine whether the loadings were zero. In fact, I had tried using that approach for CFA selection, and the mean deviation in reliability using that approach was very similar to the one used in this study (with a mean difference smaller than .00001). However, one conceptual problem lead us to discard that method: A non-significant loading would mean a lack of evidence that the loading is higher than zero, yet in the present study all loadings are above zero in the population. Thus, being unable to support that the item is significantly larger than zero is always a Type II decision error.

[6] The formula for omega-squared:

$$\omega^2 = \frac{Sum\ of\ squares\ of\ the\ effect - degree\ of\ freedom \times Mean\ squares\ of\ error}{Total\ sum\ of\ squares + Mean\ squares\ of\ error}$$

[7] a non-parametric test for comparison of two means, used because of the non-normality of the distribution of Δρ.

UNIVERSIDADE DE MACAU
仁義禮知信
澳門大學

Table 1

List of Variables to be Manipulated in the Simulation

Variable	Symbol	Value	Levels of manipulation
Sample size	N	(N_E, N_C) = (100,100), (200,	
For EFA	N_E	200), (400, 400), (800, 800),	6
For CFA	N_C	(100, 800), (800, 100)	
Number of factors	f	$f = 3, 5, 7$	3
Number of items per factors	p		
Strong indicators	p_S/f	$p_S/f = 3, 6$	2
Weak indicators	p_W/f	$p_W/f = 0, p_S/f, 2p_S/f$	3
Magnitude of items	λ	$\lambda_{Si} = .80 – (i-1)*.30/p_S$,	
Strong indicators[a]	λ_S	where $i = 1, 2, \ldots p_S-1, p_S$	3
Weak indicators	λ_W	$\lambda_W = .1, .2, .3$	

[a]The set of magnitude for strong indicators are fixed across all conditions.

UNIVERSIDADE DE MACAU 澳門大學

Table 2

Median Values of Replications Committing Selection Errors After Exploratory Factor Analyses

	False alarm				Miss				Misspecification			
	N_E=100	N_E=200	N_E=400	N_E=800	N_E=100	N_E=200	N_E=400	N_E=800	N_E=100	N_E=200	N_E=400	N_E=800
f												
3	96.6	77.6	34.7	4.4	33.1	5.1	0.2	0.0	1.3	0.0	0.0	0.0
5	100.0	93.3	59.3	9.9	78.0	15.3	0.6	0.0	11.8	0.0	0.0	0.0
7	100.0	98.4	73.0	14.7	96.7	33.8	1.6	0.0	39.7	1.3	0.0	0.0
p_S/f												
3	100.0	94.2	54.2	11.6	84.6	26.0	1.4	0.0	19.9	1.0	0.0	0.0
6	99.4	94.7	55.4	8.3	64.3	9.2	0.2	0.0	1.6	0.0	0.0	0.0
p_W/p_S												
0	---	---	---	---	56.5	12.8	0.4	0.0	1.1	0.0	0.0	0.0
1	99.3	86.5	46.6	7.2	74.6	13.5	0.2	0.0	8.0	0.1	0.0	0.0
2	100.0	98.0	70.3	13.8	78.0	16.8	0.9	0.0	18.6	0.2	0.0	0.0
λ_W												
0.1	95.0	29.7	1.0	0.0	88.4	20.2	0.7	0.0	29.5	0.7	0.0	0.0
0.2	99.4	92.1	55.4	9.4	78.5	14.3	0.8	0.0	10.0	0.2	0.0	0.0
0.3	100.0	100.0	100.0	100.0	70.2	13.5	0.3	0.0	2.3	0.0	0.0	0.0

Note. Based on percentages of 246 conditions. For each condition, the percentage equaled to the number of replications with the selection error divided by the total number of replications, or 500.

UNIVERSIDADE DE MACAU 澳門大學

Table 3

ANOVA Results for the Deviation of Population Reliability from Baseline ($\Delta\rho_0$)

Source	*df*	*SS*	*F*	ω^2
f	2	34.63	11900.99	.04
p_S	1	37.73	25933.35	.05
p_W/p_S	2	72.09	24775.64	.09
λ_W	2	1.60	549.38	.00
N_E	3	180.90	41448.39	.23
$f \times p_S$	2	2.18	748.20	.00
$f \times p_W/p_S$	4	19.12	3284.76	.02
$p_S \times p_W/p_S$	2	3.94	1354.39	.01
$f \times \lambda_W$	4	5.04	865.81	.01
$p_S \times \lambda_W$	2	4.18	1435.84	.01
$p_W/p_S \times \lambda_W$	2	3.24	1114.02	.00
$f \times N_E$	6	29.67	3398.85	.04
$p_S \times N_E$	3	59.79	13699.14	.08
$p_W/p_S \times N_E$	6	44.93	5147.79	.06
$\lambda_W \times N_E$	6	44.17	5059.94	.06
$f \times p_S \times p_W/p_S$	4	10.26	1763.47	.01
$f \times p_S \times \lambda_W$	4	0.34	58.30	.00
$f \times p_W/p_S \times \lambda_W$	4	1.21	207.20	.00
$p_S \times p_W/p_S \times \lambda_W$	2	0.22	74.46	.00
$f \times p_S \times N_E$	6	5.38	616.68	.01
$f \times p_W/p_S \times N_E$	12	7.50	429.38	.01
$p_S \times p_W/p_S \times N_E$	6	8.82	1009.85	.01
$f \times \lambda_W \times N_E$	12	4.54	259.92	.01
$p_S \times \lambda_W \times N_E$	6	10.56	1210.07	.01
$p_W/p_S \times \lambda_W \times N_E$	6	5.76	659.63	.01
$f \times p_S \times p_W/p_S \times \lambda_W$	4	0.65	110.99	.00
$f \times p_S \times p_W/p_S \times N_E$	11	1.13	70.33	.00
$f \times p_S \times \lambda_W \times N_E$	12	1.28	73.15	.00
$f \times p_W/p_S \times \lambda_W \times N_E$	12	0.44	25.25	.00
$p_S \times p_W/p_S \times \lambda_W \times N_E$	6	1.10	125.51	.00
$f \times p_S \times p_W/p_S \times \lambda_W \times N_E$	10	0.38	26.13	.00
Error	122835	178.70		

Note. All effects have a *p*-value <.0001.

UNIVERSIDADE DE MACAU 仁義禮知信 澳門大學

Table 4

Mean Change in Selection Errors from Before to After Confirmatory Factor Analyses

		False alarm						Miss						Misspecification					
	N_E	100	200	400	800	100	800	100	200	400	800	100	800	100	200	400	800	100	800
	N_C	100	200	400	800	800	100	100	200	400	800	800	100	100	200	400	800	800	100
f																			
3		−4.2	−3.2	−2.5	−2.1	−4.3	−2.1	1.8	0.6	0.1	0.0	0.1	1.8	0.0	0.0	0.0	0.0	−0.1	0.0
5		−7.1	−5.2	−4.3	−3.4	−7.2	−3.4	3.0	0.9	0.1	0.0	0.4	3.0	−0.3	0.0	0.0	0.0	−0.3	0.0
7		−7.8	−7.3	−5.9	−4.8	−8.4	−4.7	4.0	1.3	0.1	0.0	0.8	4.2	−0.5	−0.1	0.0	0.0	−0.6	0.0
p_S/f																			
3		−5.1	−3.9	−3.0	−2.4	−5.8	−2.4	2.8	1.1	0.1	0.0	0.6	2.9	−0.4	−0.1	0.0	0.0	−0.4	0.0
6		−6.9	−6.4	−5.5	−4.5	−7.0	−4.4	2.7	0.8	0.1	0.0	0.1	3.1	−0.1	0.0	0.0	0.0	−0.1	0.0
p_W/p_S																			
0		---	---	---	---	---	---	2.8	0.9	0.1	0.0	0.0	3.1	0.0	0.0	0.0	0.0	0.0	0.0
1		−4.6	−3.5	−2.9	−2.3	−4.8	−2.3	2.9	0.9	0.1	0.0	0.4	3.0	−0.2	0.0	0.0	0.0	−0.3	0.0
2		−8.0	−6.9	−5.6	−4.6	−8.4	−4.5	2.5	0.9	0.1	0.0	0.5	3.0	−0.3	−0.1	0.0	0.0	−0.4	0.0
λ_W																			
0.1		−2.4	−0.5	0.0	0.0	−3.0	0.0	2.8	1.0	0.1	0.0	0.4	3.0	−0.3	−0.1	0.0	0.0	−0.3	0.0
0.2		−4.8	−2.8	−0.9	−0.1	−5.2	−0.1	2.8	0.9	0.1	0.0	0.5	3.1	−0.3	0.0	0.0	0.0	−0.4	0.0
0.3		−10.2	−12.1	−11.8	−10.2	−10.5	−10.1	2.6	0.9	0.1	0.0	0.2	2.9	−0.1	0.0	0.0	0.0	−0.2	0.0
Overall		−6.0	−5.2	−4.2	−3.5	−6.4	−3.4	2.7	0.9	0.1	0.0	0.4	3.0	−0.2	−0.0	−0.0	0.0	−0.3	0.0

Note. Based on 1,260,000 samples. Each cell represented the mean values for all valid replications corresponded to the manipulated level of the studied variables. Replications with the confirmatory factor analyses solution not computable were excluded (n = 3,909).

UNIVERSIDADE DE MACAU 仁義禮知信 澳門大學

Table 5

Mean Change in Population Reliability from Before to After Confirmatory Factor Analyses

	N_E	100	200	400	800	100	800
	N_C	100	200	400	800	800	100
f							
3		.035	.019	.012	.010	.044	.004
5		.048	.022	.013	.010	.065	.005
7		.052	.025	.012	.009	.079	.004
p_S/f							
3		.063	.031	.016	.011	.087	.006
6		.026	.013	.009	.008	.032	.003
p_W/p_S							
0		−.005	.000	.000	.000	.001	-.005
1		.042	.020	.013	.010	.054	.004
2		.070	.032	.017	.013	.095	.007
λ_W							
0.1		.031	.008	.000	.000	.050	-.005
0.2		.062	.029	.009	.001	.082	-.004
0.3		.045	.036	.034	.032	.055	.026
Overall		.044	.022	.013	.010	.061	.004

Note. Based on 1,260,000 samples. Each cell represented the mean change in population reliability for all replications corresponded to the manipulated level of the studied variables. Replications with the confirmatory factor analyses solution not computable were excluded (*n* = 3,909).

UNIVERSIDADE DE MACAU 仁義禮知信 澳門大學

Table 6

Suggestions for Using Exploratory and Confirmatory Factor Analyses in Scale Development

1. Justify the use of a particular cutoff value, and do an a priori analysis of the consequences of the cutoff. For example, if one uses a cutoff of .32 and selects 10 items, then the expected minimum reliability coefficient is .53 using Raykov (2007)'s formula. If one uses a cutoff of .5, then coefficient is .77. Compare the results of the coefficient based on the EFA estimated loadings with the expected minimum as a piece of information for selection.
2. Do not simply aim for an infinitely large item pool, because as the proportion of weak indicators increase, the odds for decision errors also increase. Make sure that the items are generated with a sound theoretical background.
3. Beware of the possibility of committing false alarm, miss, and misspecification. Particularly, as miss is an irreversible selection error, be cautious before discarding an item, especially when the sample size for EFA is not large (smaller than 200, or participant-indicator ratio lower than 5:1, from Figure 1).
4. Be more conservative if the item pool is not large. Discarding items can result in models with factors having less than three indicators, which poses problems for conceptual integrity of the factor (see Little et al., 1999). Besides working on the quality of the item pool, researchers could consider using significant test instead of an absolute cutoff (as suggested by Cudeck & O'Dell, 1994).
5. When the total sample size is moderate or large (more than 400), divide the sample into one for EFA and one for CFA. When the toal sample size is small (less than 400), in the absence of firm theoretical reasons, use the whole sample for EFA.
6. In looking at CFA model fit, take into account whether the model χ^2 indicates a significant poor-fit. It is particularly informative for model misspecifications, compared to other fit-indexes. See Barrett (2007) for the pros and cons of deciding model rejection based on the model χ^2
7. Continue to replicate the scale even after the second CFA replication, because selection errors can still be present after CFA.

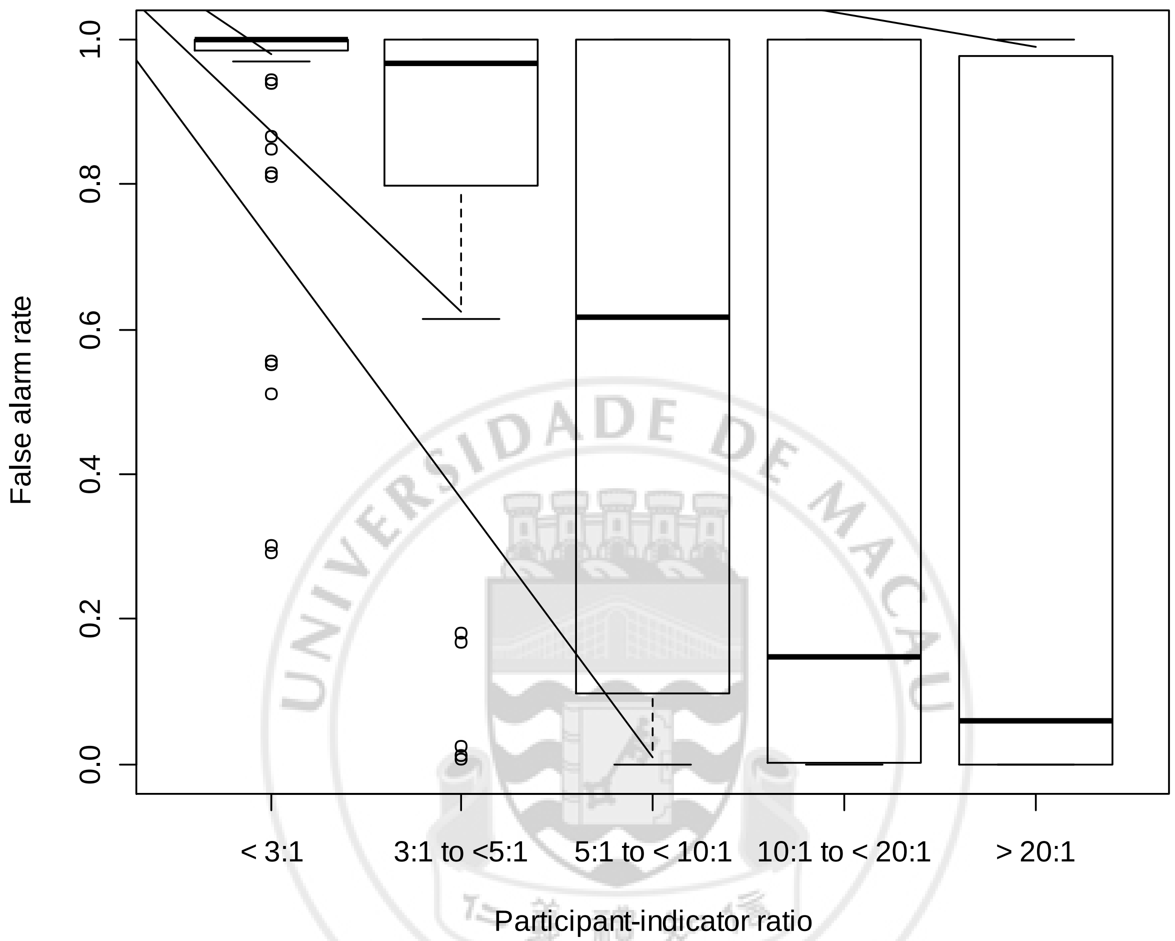


Figure 1a. Boxplot of false alarm rate by participant-indicator ratio after exploratory factor analysis. The box denoted the first quartile, the median, and the third quartile of the data. Dots showed outliers which were outside 1.5×interquartile. For all the above levels of participant-indicator ratio, the proportion of cases with at least one item with weak loadings “falsely” included was not negligible, but with the increase of the ratio the false alarm rate decreased.

UNIVERSIDADE DE MACAU 澳門大學

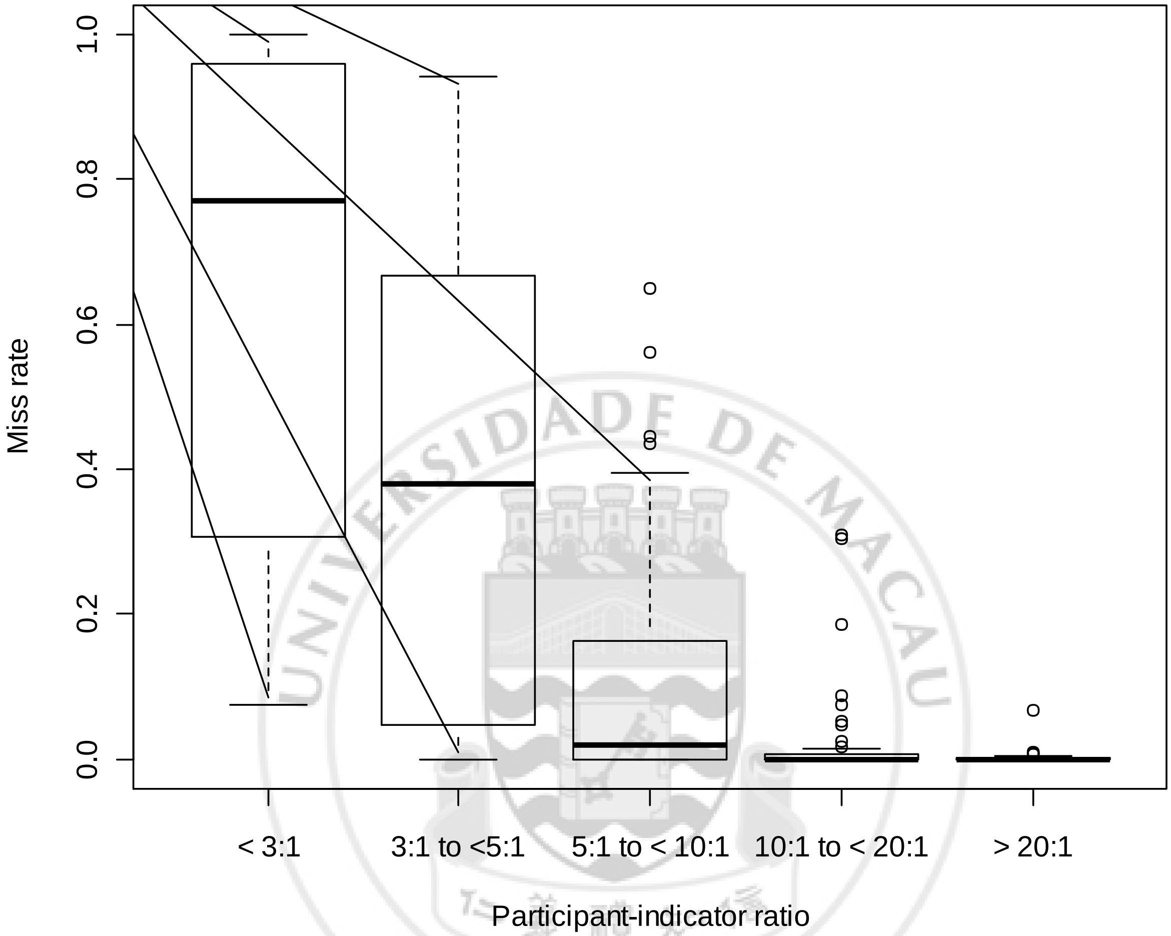


Figure 1b. Boxplot of miss rate by participant-indicator ratio after exploratory factor analysis. The box denoted the first quartile, the median, and the third quartile of the data. Dots showed outliers which were outside 1.5×interquartile. Only when the ratio reached 7:1 would there be generally less than 10% of the cases with at least one item with a strong loading be "falsely" discarded.

UNIVERSIDADE DE MACAU 澳門大學

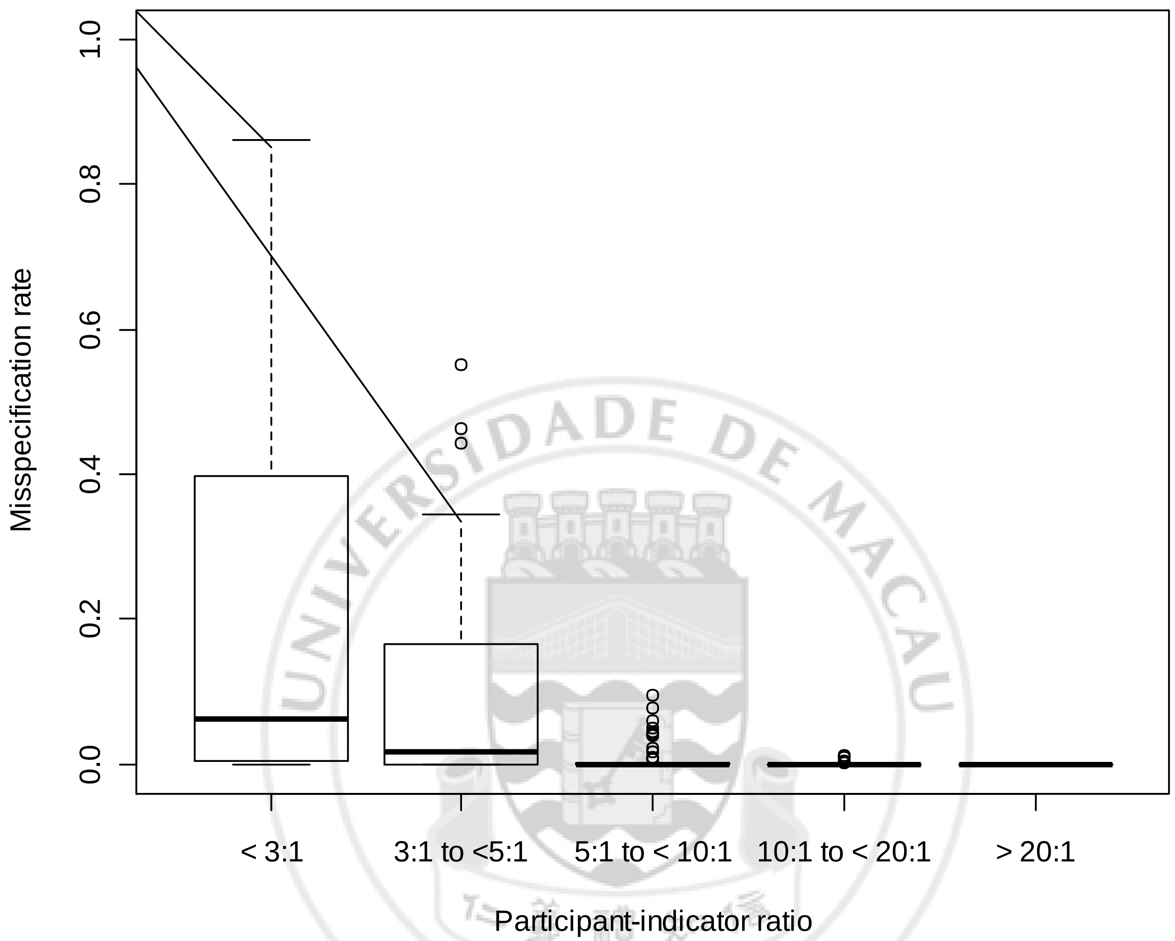


Figure 1c. Boxplot of misspecification rate by participant-indicator ratio after exploratory factor analysis. The box denoted the first quartile, the median, and the third quartile of the data. Dots showed outliers which were outside 1.5×interquartile. When the ratio reached 5:1 there were generally less than 10% of the cases with at least one item mis-specified to a factor other than the one it was supposed to load in the population.

UNIVERSIDADE DE MACAU 澳門大學

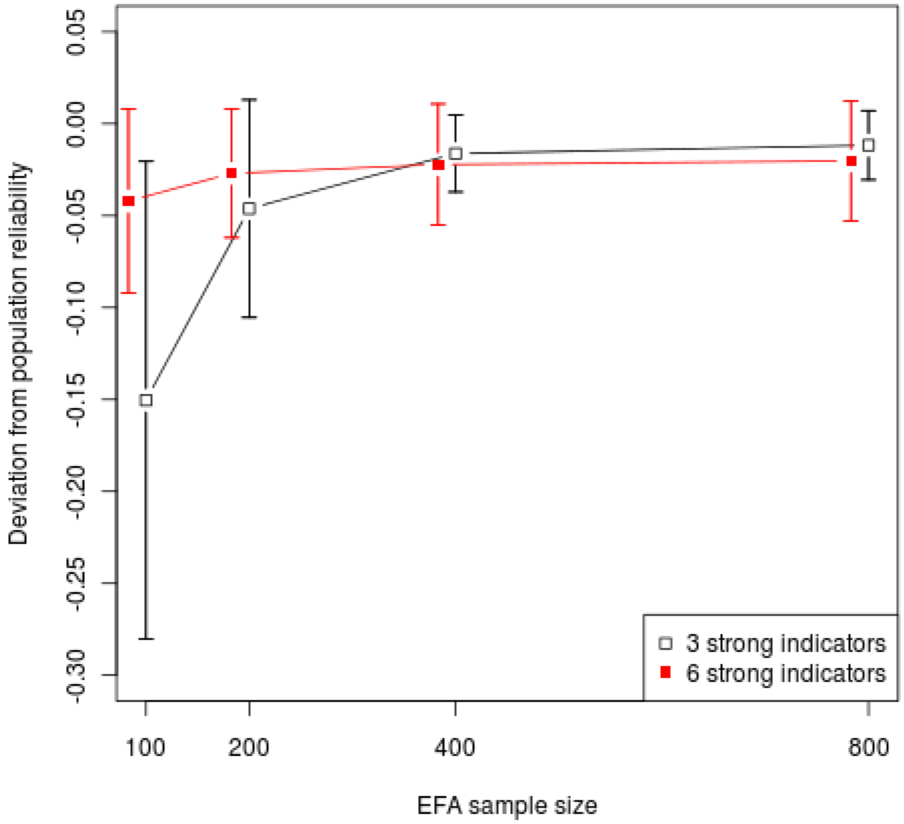


Figure 2a. Comparison of population reliability deviation from the baseline (with all and only strong indicators selected). The line graph showed the mean deviations of all replications with the error bar indicating the standard deviation. For conditions with more strong indicators were smaller, the variability was smaller, and there was less deviation from the baseline when sample size was small or moderate. When sample sizes increased, the two groups converged.

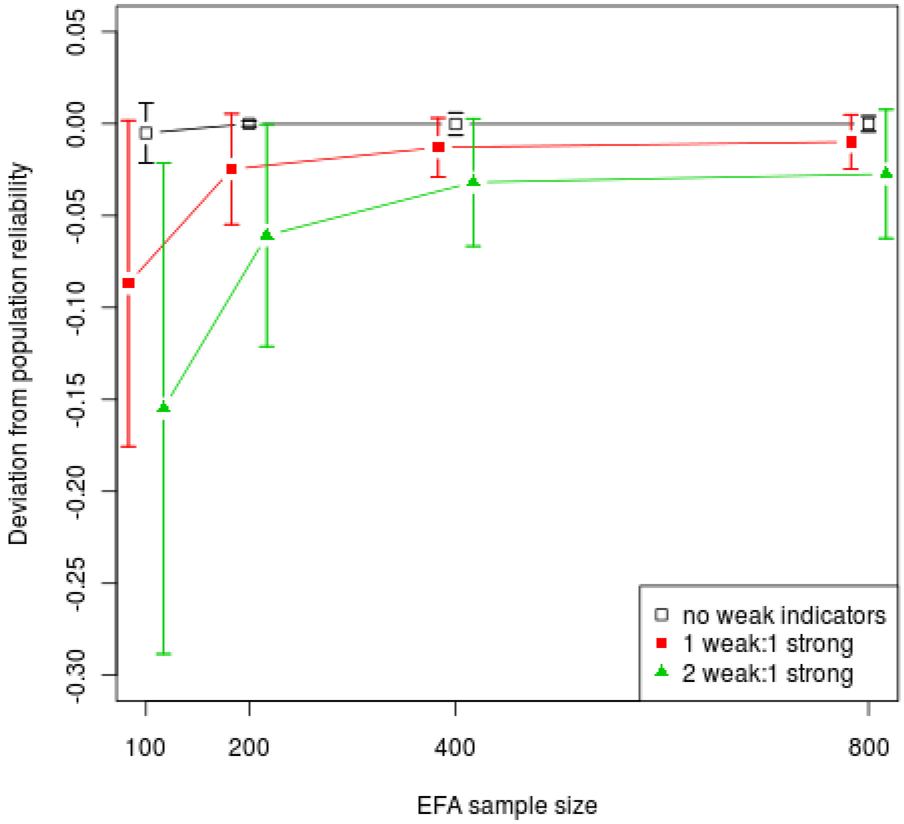


Figure 2b. Comparison of population reliability deviation from the baseline (with all and only strong indicators selected). The line graph showed the mean deviations of all replications with the error bar indicating the standard deviation. For conditions with fewer or none weak indicators in the original item pool, the variability was smaller, and there was less deviation from the baseline. When sample sizes increased, the mean deviation for the three groups started to converge, yet the variability was still higher for the high-weak-indicator-proportion group.

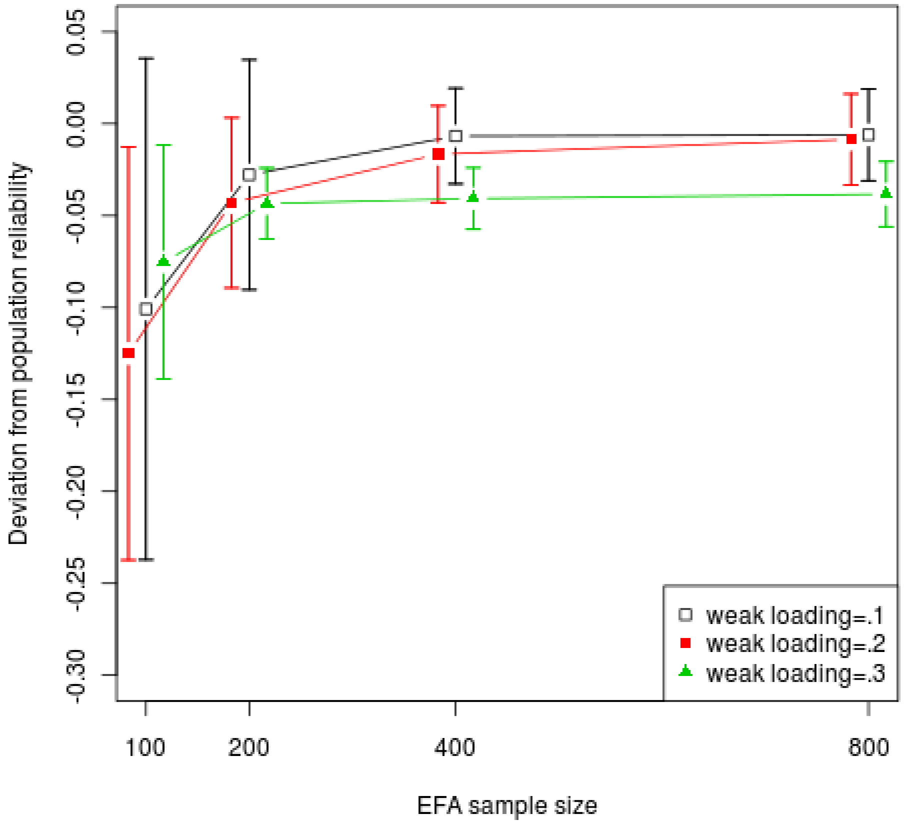


Figure 2c. Comparison of population reliability deviation from the baseline (with all and only strong indicators selected). The line graph showed the mean deviations of all replications with the error bar indicating the standard deviation. For conditions with weak loadings in population equaled .3, the variability was smaller, yet the mean deviation from the baseline was larger.

仁義禮知信 澳門大學

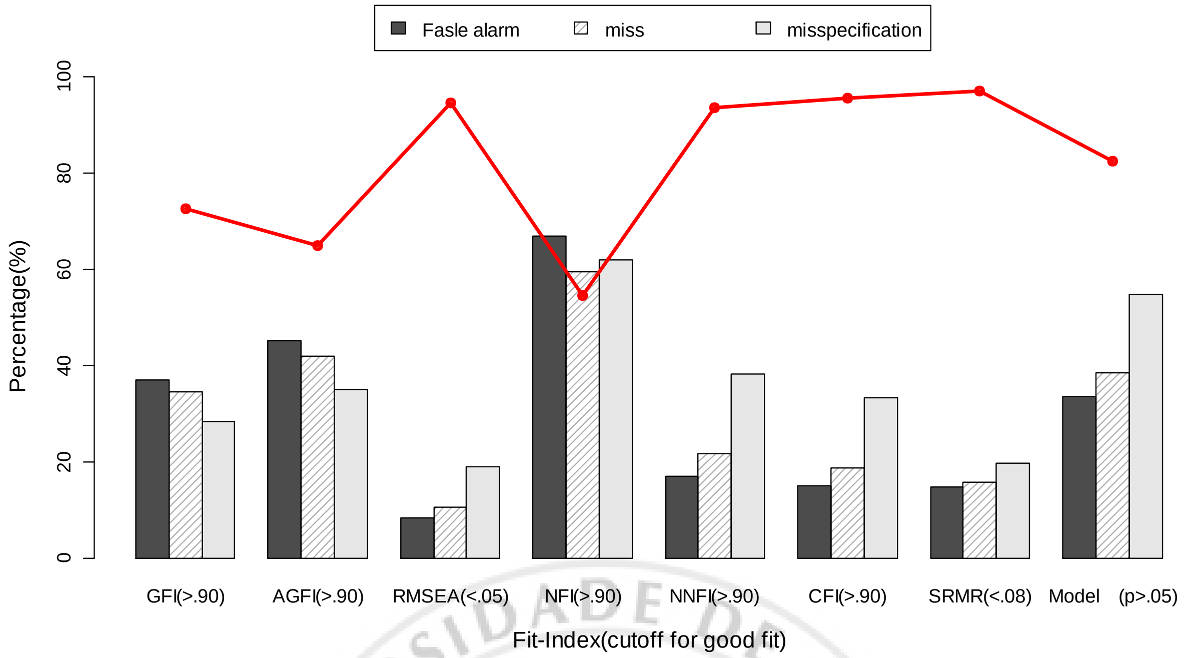


Figure 3. Selection error sensitivity and overall specificity of fit-indexes in confirmatory factor analysis to selection errors in prior exploratory factor analysis. The bars showed the selection error sensitivity, which is the ratio between the number of replications with a poor fit and with a kind of selection error to the total number of replications without that error. The red line is the overall specificity, which is the ratio between the number of replications with a good fit and without any selection errors to the total number of replications without any selection errors.

Appendix A

Consider a scale with p items. Let λ_i be the symbol for factor loadings, where i equals to 1, 2, …, $p-1$, p. The variances of an individual item, σ_{ii}^2, could be divided as

$$\sigma_{ii}^2 = \lambda_i^2\sigma_T^2 + \sigma_U^2 \text{ (1)},$$

where σ_T^2 equals to the variance of the latent factor score (or true score), and σ_U^2 equals to the variance accounted by the uniqueness (i.e. the variance of an item score not explained by the common factor) of the item. As suggested by the classical test theory and the formula used by Raykov (2007, p.203), the reliability of a measure, ρ_Y, is defined as

$$\rho_Y = \frac{\left(\Sigma_{i=1}^{p}\lambda_i\right)^2}{\left(\Sigma_{i=1}^{p}\lambda_i\right)^2 + \Sigma_{i=1}^{p}\Psi_{ii}} \text{ (2)},$$

where Ψ_{ii} refers to variance-covariance matrix for the uniquenesses of items, which has σ_U^2 as its diagonal elements. Assume that the uniquenesses does not correlate with each other, which means that the only source of covariance between items is the common factor variance (and no common method variance). Thus, Ψ_{ii} is a diagonal matrix.

Consider a set of five items each having a loading of .6 ($\lambda_1 = \lambda_2 = \ldots = \lambda_i = .7$) on a factor in the population. This means that for each item, 36% of the variances could be explained by the common method variance. Assume that σ_T^2 equals one, which implies that $\sigma_{Ui}^2 = 1 - .36 = .64$. Using formula (2), the reliability for this scale in the population would be

$$\rho_Y = \frac{(.6\times 5)^2}{(.6\times 5)^2 + .64\times 5} = .738,$$

which meet the convention of .7 level. However, if one of them was replaced by an item with a loading of .1, reliability would then drop to

$$\rho_Y = \frac{(.6\times 4 + .1)^2}{(.6\times 4 + .1)^2 + (.64\times 4 + .99)} = .638$$

which beneath the conventional level.

UNIVERSIDADE DE MACAU
澳門大學

This lowered reliability of the scale could then attenuate its validity using other criterion variable W (Thompson, 2003). If the latent factor of the scale have a population correlation $\rho_{YW} = .5$ with W, its estimated value r_{YW} using the scale with five .6 items would be, by the formula

$$r_{YW} = \sqrt{\rho_{Y}\rho_{W}}\rho_{YW}$$

and assume that the measurement for W contains no error, then $r_{YW} = .429$. However, if one item with loading equaled to .1 is selected, r_{YW} would drop to .399. With the strong indicator loadings higher or the item selected fewer, or the number of weak indicators selected increases, this attenuation effect would be more significant (see Figure A1).

UNIVERSIDADE DE MACAU
仁義禮知信
澳門大學

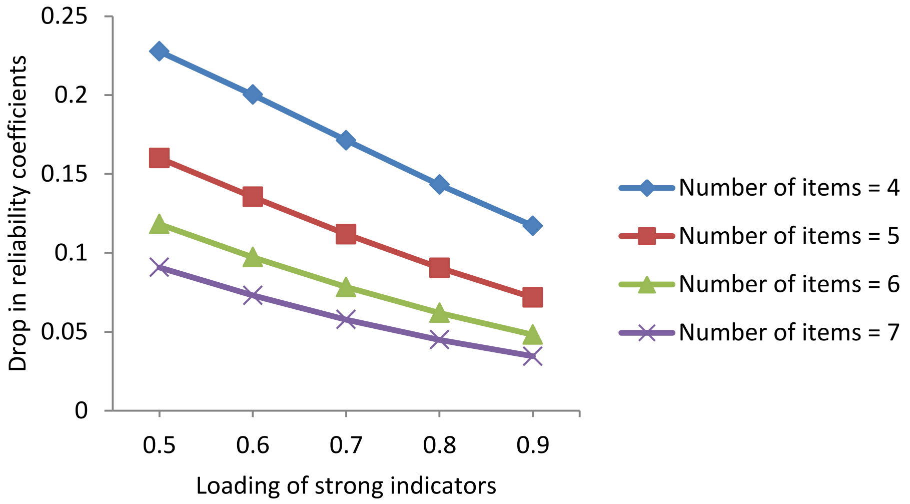


Figure A1. Relations between loading of strong indicators and drop in reliability coefficients under varying conditions of total number of items. In each condition, number of weak indicators equaled 1, and its loading equaled 0.1. Drop in reliability coefficients referred to the decrease in reliability coefficients when one strong indicator was replaced by a weak indicator.

UNIVERSIDADE DE MACAU 仁義禮知信 澳門大學

Appendix B

Characteristics of 23 recent articles about scale development in PsycINFO:

Study	Cutoff for low factor loadings	Cutoff for cross-loadings[a]	Any guidelines cited?	N_E	N_C	Initial number of items	Final number of items
1. Kim, Han, & Yoon	.50	.30	Yes	488	451	103	15
2. Wood, Worthington, Exline, Yali, Aten, & McMinn	.75	.20	No	394	93	11	9
3. Şimşek	.35	.10	No	352 / 178	352 / 227 / 178[c]	21	17 / 12
4. Wright, Creed, Zimmer-Gembeck	.40	Not mentioned[b]	No	212	236	49	14
5. Osberg et al.	.40	NA	No	228	343	29	15
6. Wei, Alvarez, Ku, Russell, Bonett	.45	.30	No	328	328[c]	41	25
7. Wang & Chang	.60	Not mentioned[b]	No	164	433	56	27
8. Armfield	Based on highest loadings		No	1083	NA	16	8
9. Vanderlinde & Braak	NA	Not mentioned[b]	No	471	471[c]	35	18
10. Stankov, Higgins, Saucier, & Knežević	Based on low communalities		No	452	NA	132	24
11. Livanis & Tryon	NA	NA	No	307 / 277	NA	34	31
12. Storch, Rasmussen, Price, Larson, Murphy, & Goodman[e]	.40	NA	No	130	130[c]	10	10
13. Ku & Minas	Not mentioned[b]	Not mentioned[b]	No	208	NA	34	26
14. Abramowitz et al.	NA	NA	No	478	477 / 423	20	20
15. Waters & Cross	Not mentioned[b]	NA	No	2809	2809	14	13
16. Labbe & Maisto	.40	NA	Yes	91	294	26	16
17. Davis et al.	.50	NA	No	300	150	6	4
18. Neilands, Chakravarty,	NA	NA	No	380	1001	13	13

UNIVERSIDADE DE MACAU 仁義禮知信 澳門大學

Darbes, Beougher, & Hoff							
19. Timothy et al.	.50	NA	Yes	558	545	7	6
20. Yost	.30	Not mentioned[b]	No	213	258	37	23
21. Vissers, Keijsers, van der Veld, de Jong, & Hutschemaekers	Based on highest loadings		No	750	199	69	12
22. Cua, Junttila, & Schroeder	NA	NA	No	NA	435 / 164	5	5
23. Noone, Stephens, & Alpass	Reliability		No	1532	1532[c]	52/ 52 / 52 / 52	50/ 49 / 49 / 50

Note. EFA = Exploratory factor analysis; CFA = Confirmatory factor analysis. All studies are published in 2010, and found using the keywords “scale” and “develop*”. Cells with NA denoted a lack of information in the original articles.

[a]The cutoff set so that items with the second largest regression weight on the factors larger than this value was discarded.

[b]The article stated that a cutoff was used as a criterion for determining deletion of items, but the value was not mentioned.

[c]Confirmatory factor analysis shared the sample with exploratory factor analysis.

[d]CFA was done before EFA.

UNIVERSIDADE DE MACAU 仁義禮知信 澳門大學

Appendix C

Besides sensitivity and specificity for different ratios, *non-computable sensitivity* was calculated to shown the association between CFA non-computability and each selection error. This ratio corresponded to the number of replications in which the EFA selection contained that error and the CFA solution was not computable with the EFA implied model divided by the total number of replications in which the EFA selection contained that error. Results were reported in Table C1. Small sample size, more factors to be extracted, fewer strong indicators, more weak indicators, and lower magnitude of weak loadings associate with increased non-computable sensitivity, but the effects may be due to the increase in number of non-computable cases. Generally, misspecification leads to more occurrences of non-computable CFA.

UNIVERSIDADE DE MACAU 仁義禮知信 澳門大學

Table C1

Group means of non-computable sensitivity

	Mean number of non-computable cases	False alarm	Miss	Misspecification
N_E-N_C				
100-100	64.6	.15	.14	.22
200-200	1.4	.02	.02	.10
400-400	7.2	.00	.01	.00
800-800	0.1	.00	.00	---
100-800	26.4	.06	.06	.10
800-100	0.0	.01	.32	---
f				
3	3.5	.01	.08	.09
5	14.4	.04	.05	.14
7	30.8	.08	.09	.18
p_S/f				
3	28.7	.07	.12	.17
6	2.4	.01	.02	.09
p_W/p_S				
0	1.8	---	.02	.03
1	12.7	.03	.07	.13
2	24.3	.01	.09	.19
λ_W				
0.1	31.1	.10	.13	.20
0.2	17.6	.04	.08	.17
0.3	6.2	.01	.04	.08
Overall	15.9	.04	.07	.14

Note. Non-computable sensitivity$=\frac{N(\text{CFA solution was not computable } \mid \text{EFA selection contained that error})}{N(\text{EFA selection contained that error})}$.

Conditions in which no selection errors were found in EFA were excluded.

UNIVERSIDADE DE MACAU 仁義禮知信 澳門大學