Using Factor Mixture Modeling to Counter Faking

Raul Corrêa Ferraz

Follow this and additional works at: https://scholarcommons.sc.edu/etd

Part of the Experimental Analysis of Behavior Commons

Recommended Citation

This Open Access Dissertation is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact digres@mailbox.sc.edu.
USING FACTOR MIXTURE MODELING TO COUNTER FAKING

by

Raul Corrêa Ferraz

Bachelor of Science
Federal University of Santa Maria, 2018

Master of Arts
University of South Carolina, 2020

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy in
Experimental Psychology
College of Arts and Sciences
University of South Carolina
2023

Accepted by:
Alberto Maydeu-Olivares, Major Professor
Dexin Shi, Committee Member
Sarfaraz Serang, Committee Member
Lynn A. McFarland, Committee Member
Cheryl L. Addy, Interim Vice Provost and Dean of the Graduate School
ACKNOWLEDGEMENTS

I would like to thank my advisor, Alberto Maydeu-Olivares, for his advice and help throughout my graduate school career and in completing this project. Despite your busy schedule, you always made time for a good walk around campus, answering my questions and sharing your wisdom. This dissertation would not have been possible without your guidance. I would also like to thank my committee members Dexin Shi, Sarfaraz Serang, and Lynn Ann McFarland for their helpful feedback and insights, as well as their patience and flexibility throughout the dissertation process.

Thanks also to my parents, Vânia and Orides, and my sister, Laura, who have always encouraged me in everything I have ever done. It was very challenging to come to the distant Columbia, SC, far away from everything I had ever known, and navigate the challenges of a different country and of a graduate program. Without your unwavering support I would never have succeeded.

I would also like to thank my loving and caring partner, Mariajosé, for her understanding and support. You were always there to help me and cheer me on. I will forever be grateful for the stroke of luck that brought us together.
ABSTRACT

Self-reports (SRs) of typical behavior are often the only existing feasible method to gather data on important drivers of human performance. In applications such as personnel selection, SRs are vulnerable to intentional distortions, often referred to as faking. A review of the literature suggests that so far, the methods proposed to address faking are unsatisfactory. In a recent breakthrough, Pavlov et al. (2019) showed that high-stakes scale scores are best modeled as a function of a) propensity to fake, b) honest scores, and c) the interaction of these two terms. Pavlov et al. did not, however, propose any method to extract honest scores in assessment settings, when only high stakes data is available. In this dissertation I investigate using factor mixture modeling (FMM) with class specific intercepts and factor loadings to this aim. I assume that responses to high stakes items are a function of the respondents’ “honest” factor scores on the attribute being measured, an unobserved categorical “tendency to fake” latent class, and their interaction. I perform simulations using parameter estimates based on Pavlov et al.’s data to determine the extent to which factor scores estimated using a two class factor analysis model outperform the current standard, a single class model. Results suggest that only under specific conditions FMM scores provide higher correlations with true factor scores compared to single class models. Moreover, empirical findings indicate that the theoretical potential of FMM to detect faking is not realized in practice, where class separations are not as defined.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS ................................................................................................................ iii

ABSTRACT ........................................................................................................................................ iv

LIST OF TABLES ............................................................................................................................... viii

LIST OF FIGURES ............................................................................................................................. ix

CHAPTER 1: INTRODUCTION ........................................................................................................... 1

1.1 Where is the study of faking of interest? Personality and I/O psychology ......................... 6

1.2 The Big Five model of personality ......................................................................................... 6

1.3 Faking ......................................................................................................................................... 9

1.4 Statistical background ............................................................................................................... 22

CHAPTER 2: METHODS .................................................................................................................. 32

2.1 Preliminary data analysis ......................................................................................................... 32

2.2 Simulation (Monte Carlo) study ............................................................................................ 35

2.3 Outcomes .................................................................................................................................. 38

2.4 Empirical application ............................................................................................................. 39

CHAPTER 3: SIMULATION RESULTS ............................................................................................ 43

3.1 Mean scores ............................................................................................................................. 43
APPENDIX D: EXAMPLE R SCRIPT FOR SCORING AND CLASSIFICATION ... 133

APPENDIX E: R SCRIPT FOR CROSS TABULATION......................................................... 140
LIST OF TABLES

Table 3.1 Simulation results for small models ($p = 6$): correlations between true and estimated factor scores .................................................................................................................................................. 60

Table 3.2 Simulation results for intermediate models ($p = 12$): correlations between true and estimated factor scores .................................................................................................................................................. 61

Table 3.3 Simulation results for large models ($p = 18$): correlations between true and estimated factor scores .................................................................................................................................................. 62

Table 3.4 Simulation results for small models ($p = 6$): classification .................................................................................................................................................. 63

Table 3.5 Simulation results for intermediate models ($p = 12$): classification .................................................................................................................................................. 64

Table 3.6 Simulation results for larger models ($p = 18$): classification .................................................................................................................................................. 65

Table 4.1 CFA models and goodness-of-fit indexes for the rating scale version of the BFT ($n = 622$) .................................................................................................................................................. 84

Table 4.2 Information criteria for two-class factor mixture models of the BFT ($n = 609$) .................................................................................................................................................. 85
LIST OF FIGURES

Figure 1.1 Selected contributions in the history of the Big Five model of personality .....31
Figure 2.1 Regression model for Extroversion scores with interaction by faking.............40
Figure 2.2 Simulated data example using a two-group factor analysis model and analyzed using an MGFA .................................................41
Figure 2.3 Simulated data example using a two-group factor analysis model and analyzed using an FMM.........................................................42
Figure 3.1 Simulation results for the conditions with highest class separation .............66
Figure 3.2 Simulation results for the conditions with lowest class separation.............67
Figure 3.3 Simulation results: lowest class separation with increased factor loadings .....68
Figure 3.4 Simulation results: lowest class separation with decreased item intercepts.....69
Figure 3.5 Simulation results: lowest class separation with decreased uniquenesses ......70
Figure 4.1 Visual guide to determine plausible starting values .....................................86
Figure 4.2 Neuroticism scale scores in Wetzel et al.’s data, group-coded .....................87
Figure 4.3 Extroversion scale scores in Wetzel et al.’s data, group-coded .....................88
Figure 4.4 Openness scale scores in Wetzel et al.’s data, group-coded........................89
Figure 4.5 Agreeableness scale scores in Wetzel et al.’s data, group-coded.................90
Figure 4.6 Conscientiousness scale scores in Wetzel et al.’s data, group-coded..........91
Figure 4.7 FMM classification of Neuroticism scale scores in Wetzel et al.’s data.........92
Figure 4.8 FMM classification of Extroversion scale scores in Wetzel et al.’s data.......93
Figure 4.9 FMM classification of Openness scale scores in Wetzel et al.’s data.........94
Figure 4.10 FMM classification of Agreeableness scale scores in Wetzel et al.’s data ....95

Figure 4.11 FMM classification of Conscientiousness
scale scores in Wetzel et al.’s data.................................................................96
CHAPTER 1: INTRODUCTION

Self-reports (SRs) of typical behavior are often the only existing feasible method to gather data on important drivers of human performance. In industrial and organizational (I/O) psychology, personality is one of those drivers and is used in applications such as personnel selection, training, licensing, or promotion, or to predict individual performance in a work role, productivity as part of a team, and voluntary turnover (e.g., Barrick & Mount, 1991). In these applications, SRs are vulnerable to intentional, goal-oriented, distortions, often referred to as faking or impression management.

Faking is a documented problem in several real-life situations. For example, there is evidence that job applicants exaggerate characteristics that they perceive to be desirable in order to be hired (Donovan et al., 2003). Faking is also a concern for promotion purposes, with internal candidates distorting their responses based on coaching rumors spread within companies and departments (Landers et al., 2011). More worryingly, a study on the US Army entrance exam found that recruiters themselves often indicate to prospective candidates what answers may result in a failed test (White et al., 2008).

Different approaches have been proposed to deal with the problem of faking. These approaches can be categorized into four main groups.

The first approach is to deter or discourage faking. Typical methods include introducing warnings that faking can be detected and describing what the consequences
of that detection will be (Doll, 1971; Vasilopoulos et al., 2005), requesting written elaboration on provided answers (Schmitt & Kunce, 2002), and manipulating the grouping of items so that desirable answers are not obvious to the respondents (Morgeson et al., 2007).

Another approach to deal with faking is to try to detect distorted responses. The most widely used method within the detection approach is the use of the so called “lie scales” (e.g., Ellingson et al., 2001), which include items deemed unlikely to be endorsed by honest respondents such as “I never lie”. Within this approach we can also include the use of over-claiming or “bogus” items, where for example inexisten products are used in a test of brand knowledge (Phillips & Clancy, 1972). Finally, the use of idiosyncratic item responses has seen some use in directed faking experiments, but has received less attention in applied contexts (Kuncel & Borneman, 2007). The combination of deter and detect approaches is popular but faces challenges in applications where coaching is a concern (Hough & Connelly, 2013), and correlations with cognitive abilities can be an issue (Kuncel, Borneman, et al., 2011).

The third approach to deal with faking is to use measurement methods that prevent or difficult faking. Research in this area has for example explored item characteristics of tests that are known to predict criteria of interest (Schmidt & Hunter, 1998) as well as the use of subtle item content (Schmit et al., 1995). Within this approach we also include the use of forced-choice measures (e.g, Christiansen et al., 2005), and conditional reasoning measurement (James, 1998). These approaches are insightful and fruitful, but test development can prove challenging (Converse et al., 2010).

The fourth approach consists of using a statistical model that may mimic the
process used by individuals to respond to high stakes items. The goal is not only identifying individuals that fake in SRs, but also obtaining scores free of faking effects (Brown & Böckenholt, 2022; Zickar et al., 2004; Ziegler et al., 2015). This dissertation is positioned within this approach.

The model investigated in this dissertation is motivated by a recent study by Pavlov et al. (2019). They used a within-subjects experimental design gathering baseline (honest) scale scores, and high-stakes scale scores to show that job applicant’s scale scores are best modeled as a function of a) their propensity to fake, b) their true (i.e., honest) scores, and c) the interaction of these two terms. As a result, scores free of faking can be accurately estimated from high-stakes scores, propensity to fake, and their interaction.

Pavlov et al. did not, however, propose any method to extract honest scores in assessment settings, when only high stakes data is available. It is simply not possible to regress assessment scores on their “honest” counterparts because such honest scores and tendency to fake are not observed. However, it is in principle possible to estimate the unobserved “honest” scores and the tendency to fake if multiple measurements are available. The item responses in the high stakes setting provide multiple measures of the unobserved “honest” scores and the unobserved tendency to fake, modeled as common factors, leading to a two factor model with interactions. The common factor model is unrestricted (i.e., exploratory), as every item is assumed to depend on (1) a “tendency to fake” common factor, (2) a common factor representing the attribute of interest (free of faking effects), and (3) their interaction.

Unfortunately, at the time of this writing there is no statistical theory for
estimating such a model. Several approaches have been proposed for modeling restricted (i.e., confirmatory) factor analysis models with interactions (Arminger & Muthén, 1998; Jöreskog & Yang, 1996; Kenny & Judd, 1984; Klein & Moosbrugger, 2000; Ping, 1995). However, they cannot be applied to this setting as they require assuming that the responses to some items depend only on tendency to fake, or on the attribute being measured, an unrealistic assumption.

As a workaround, I propose to treat the tendency to fake as a discrete unobserved (latent) variable instead of as a continuous latent variable. In other words, as a latent class instead of a common factor. More specifically, when a single attribute is being measured, a model in which item responses are assumed to depend on a common factor, (for example) three classes and their interactions, amounts to reducing the infinite range of \textit{trait by tendency to fake} interactions to just three possible interactions. Measurement models involving both discrete and continuous latent variables are referred to as factor mixture models (FMM). In particular, models with interactions between the common factor and the latent classes are analogous to multiple group factor models with class specific intercepts and slopes and unknown group membership. As a result, the number of parameters grows very rapidly as the number of latent classes increases and easily exceeds the number of identified parameters in an unrestricted factor analysis model (McDonald, 1967). Hence, this investigation will be limited to FMMs with two classes.

In theory, FMM is the prime candidate to model faking responses, as it can account for both measurement error in the substantive factor(s) of interest (for example, the Big Five) and heterogeneity that is due to an unobserved group membership through the latent class variable. In applications where faking is a concern and again assuming
only a single attribute is being measured, a two-class FMM involves one class of faking respondents and one class of honest respondents. For each individual, we then estimate a) a factor score that assumes membership to each class, b) the posterior probability of class membership, c) a factor score weighted by the probabilities of class membership, and d) a predicted class membership.

Mixture models in general are notoriously difficult to reliably estimate (Hipp & Bauer, 2006; McLachlan & Peel, 2000). Their likelihood is multimodal, and as a result, local optima are frequently encountered. Therefore, a key research question is the extent to which factor mixture modeling results can be trusted when making selection decisions without risking adverse impact. This research focuses on one factor mixture models because failure to reliably estimate the simplest model will cast doubts on the feasibility of the approach to model faking processes in more complex cases.

This research is organized as follows. In the Introduction, first I briefly review the Big Five model of personality and its significance to the industrial/organizational (I/O) psychology literature. Second, I review previous studies on faking and strategies to handle it. Third, I provide the statistical background relevant to the development of FMM. Fourth, in the Methods section I provide preliminary findings from a few analyses of Pavlov et al.’s (2019) data which will inform the parameters used in the simulation. I describe the simulation conditions and introduce the dataset used in the empirical application. Fifth, results for all studies are described. Finally, I discuss important findings, limitations, and suggest future research.

1 Although FMM applications are not restricted to personality, the predominance of the topic in the faking literature is enough to warrant a brief description of the most common personality taxonomy. A personality measure is also used in the empirical study in Chapter 4.
1.1 Where is the study of faking of interest? Personality and I/O psychology

What is often cited as the first model of work performance is that of Campbell and Pritchard (1976), who proposed that Performance = Ability × Motivation. In the 1980s, models of work performance grew to include cognitive abilities and job knowledge (Hunter, 1983), and eventually job experience (Schmidt et al., 1986), but dispositional tendencies such as personality were not incorporated, possibly due to the mixed findings in early validity studies of personality in the context of personnel selection (Ghiselli, 1966; Guion & Gottier, 1965). In the early 1990s, construct-oriented studies of personality–criterion relationships (Barrick & Mount, 1991; Hough et al., 1990) motivated the inclusion of personality variables as predictors of work performance more widely.

The Borman et al. (1991) model of work performance incorporated two facets of Conscientiousness — dependability and achievement orientation — and accounted for more than twice the variance in job performance ratings than Hunter’s (1983) model. Schmidt and Hunter (1992) summarized evidence for a similar model that included general mental ability, job experience, and conscientiousness as determinants of job performance. At the time of these publications, many I/O psychologists conceded that Conscientiousness might be a useful predictor of job performance across work settings, and cautiously started to investigate other personality variables. Soon, Openness to Experience gained importance as a predictor of innovation and creativity (Bartram, 2005; Hough, 1992; Hough & Dilchert, 2007).

1.2 The Big Five model of personality

The development of the Big Five model traces back to the lexical hypothesis. The
lexical hypothesis can be defined by two postulates: (1) characteristics that are important to a group of people become a part of that group's language (Cattell, 1943), and (2) the more important characteristics are more likely to be described with a single word (John et al., 1988). Galton (1884) was one the first scientists to investigate the lexical hypothesis within the context of personality, simply by studying a dictionary and identifying words “expressive of character”. He found roughly a thousand of these words. Throughout the next few decades, other researchers (Klages, 1929; Partridge, 1910; Perkins, 1926) investigated personality by using dictionaries and characterology publications, with Baumgarten (1933) being credited with the first publication of a psycholexical classification of psychological terms.

In 1934, Thurstone pioneered the use of factor analysis methods in the study of 60 adjectives descriptive of personality and proposed the first (to the best of my knowledge) five-factor model of personality. Curiously, Thurstone never followed up on this early analysis. In parallel, Allport and Odbert (1936) reported that they identified nearly 18,000 terms used to describe psychological characteristics. They separated these words into four categories (“columns”), with 4,504 words specifically used to described personality traits (“Column I”). It was terms from this Column I that Cattell, inspired by Thurstone’s findings, investigated using factor analysis throughout the 1940s (1943, 1944, 1945, 1947, 1948), identifying 35 personality factors. This list was later reduced to five factors by Fiske (1949). In this same line of research, Tupes and Christal (1958, 1961) used the same 35 trait variables developed by Cattell and also arrived at a five-factors solution.

Norman (1963, 1967) did an independent analysis of Allport and Odbert’s Column I and identified 2,797 trait-descriptive terms. Norman’s study was the basis for
Peabody and Goldberg’s Big Five personality traits (Goldberg, 1990, 1992, 1993; Peabody & Goldberg, 1989), which would lead to the influential International Personality Item Pool (IPIP; Goldberg, 1999). The IPIP is a public domain resource with over 3,300 items that have been derived based on joint-administrations of several measures (Goldberg et al., 2006).

In general, all previously described studies had a “bottom-up” approach: terms indicative of personality were identified, and factor analysis was used to cluster them into themes. Costa and McCrae have argued that this lexical-hypothesis approach is not suitable to the development of a hierarchical model of personality (Costa & McCrae, 1995). They point out the under-representation of some important traits in the lexicon and that broader terms can account for a disproportional amount of the covariance. These limitations motivated their “top-down” approach to the development of the NEO-PI-R (Costa & McCrae, 1992, 1995), a proprietary, hierarchical measure of personality. In the top-down approach, first the domain is defined, and then the items that are meant to measure those domains are written and/or selected and analyzed. In the literature, this is also termed the “questionnaire tradition”, as historically many scales were developed with specific applications in mind (McCrae & John, 1992). Costa and McCrae can also be credited with popularizing the name of the Big Five domains as Neuroticism (N), Extroversion (E), Openness to Experience (O), Conscientiousness (C), and Agreeableness (A). A simplified summary of the history of the Big Five is provided in Figure 1.

In I/O Psychology as in other areas, personality is typically measured with self-report inventories (SRIs), a standard method of assessment that has been developed over the last century (Heymans & Wiersma, 1906; Woodworth, 1920). SRIs typically consist
of self-descriptive statements (e.g., “I keep a cool head in emergencies”) and response formats such as "true" or "false" or Likert (1932) scales indicating the degree of agreement. These measures of noncognitive characteristics differ from their cognitive counterparts (such as college admission tests) in that they measure typical performance, as opposed to optimal or maximum performance (Bandalos, 2018). A challenge in the use of responses to self-reports of typical performance is that they can be intentionally or unintentionally distorted (i.e., impression management or self-deception; Paulhus, 1984). In this study, I focus primarily on Likert instruments and on impression management: the intentional, goal-oriented distortion of responses (“faking”).

1.3 Faking

Due to the intrinsically misleading nature of faking, it is of interest to study intentional distortion in a controlled environment. Much of the research on faking has adopted experimental designs where some individuals are instructed to fake high or low scores and their responses are compared to those of respondents that are instructed to answer as truthfully as possible (Griffith & Robie, 2013; Viswesvaran & Ones, 1999). While this obviously deviates from the typical use cases for personality assessment in real-life applied settings, it provides valuable information on the effects of intentional distortion.

In terms of test scores, some studies have investigated the change in mean scores across different conditions. In fake-good conditions, mean scores generally increase from half to a full standard deviation. Similar magnitudes are observed in fake-bad conditions, only in the opposite direction. The differences between honest and faking conditions are larger in within-subject study designs compared to between-subjects study designs.
The effect size also tends to be larger when the honest condition precedes the (enhanced) self-presentation condition (Hooper, 2007). Finally, correlations between honest and faked scores are generally low, indicating that faking introduces construct-irrelevant variability that do not reflect traits of interest (Ellingson et al., 1999; Zickar & Robie, 1999).

Some studies have also investigated how construct validity is affected by faking. For example, Zickar and Robie (1999) compared the factorial structure of answers to a multidimensional personality test across honest, directed faking (e.g., “choose the item that will make you look the best”), and coached faking (directed faking with the addition of example items and an explanation as to why certain answers are desirable) conditions. In the faking conditions, higher factor intercorrelations were observed compared to the honest condition. The authors suggest that these differences could be modeled through a new, common factor to all items, that is independent of the variance attributed to the substantive constructs (e.g., a bifactor model). Unfortunately, the authors did not test such model, but their findings are consistent with other studies in this literature that have found what has been interpreted as a “ideal-employee factor” (Cellar et al., 1996; Paulhus et al., 1995; Schmit & Ryan, 1993). For example, Schmit and Ryan (1993) used confirmatory factor analysis to compare the factor structure of a Big Five measure in a sample of college students and a sample of job applicants. In the job applicant sample, participants responded differently to items that they view as work-related and desirable for employee selection purposes. Although contradictory findings exist in the literature (Ellingson et al., 1999, 2001), these discrepant findings seem to be due to a failure of some studies to consider individual differences in faking (McFarland & Ryan, 2000).
To summarize, directed-faking studies indicate that the effects of intentional distortion on psychometric properties are meaningful and negative. However, these findings do not necessarily generalize to the behavior of job applicants in real-life selection situations. Therefore, it is important to determine whether real-life job applicants fake and if so, to what extent.

1.3.1 Do respondents fake in real life situations?

Overall, previous research seems to agree that some job applicants do distort their responses to portray themselves as better candidates than they are in reality. The extent of faking is, however, a point of contention. Donovan, Dwight, and Hurtz (2003) surveyed college students who had recently applied for a job. Approximately 32% of participants responded that they had exaggerated their personal characteristics or traits to make themselves look better, and 62% de-emphasized negative attributes. Further, over 15% of respondents indicated that they had given completely made-up responses. Although limited by its self-reported design, this study provided insight into the amount of distortion that may happen in actual applicant settings.

Hogan, Barrett, and Hogan (2007) published an influential article on the Journal of Applied Psychology about faking on personality measures (among others) in real-world settings. They examined test scores from job applicants who were denied employment and who were allowed to complete the tests again at a later opportunity. This research design intended to induce in the participants the motivation to improve their original scores. Hogan et al.’s (2007) data showed however that mean scores did not change meaningfully across the two administrations. While this finding is insightful, it was not determined whether individuals who failed the first administration were unable or
unwilling to fake their responses. In addition, the retested individuals reflected the entire distribution of personality test scores, because some components of their assessment (such as cognitive tests) had larger weight on the individuals’ pass-fail status.

Landers, Sackett, and Tuzinski (2011) also published research on retesting after initial failure. Within an organization, applicants for a managerial position who failed a first assessment had a 1.40 SD mean gain on the retest score. Over time, these participants also engaged in what the authors called “blatant extreme responding” (maximizing the answer to every item) more often than applicants who did not choose to retest.

Evidence from meta-analytic studies can provide a good panorama of the prevalence of faking in real-life contexts. Birkeland et al. (2006) examined faking on personality inventories that were used as part of different employment processes. They compared mean test scores of job applicants to mean test scores of non-applicants, such as job incumbents, on the same tests. Applicants’ mean scores were on average 0.13 to 0.52 standard deviation higher than those of incumbents. Ones et al. (1993) also conducted a meta-analysis, investigating integrity test validities. At face value, integrity tests are great candidates from research on faking because they directly relate to “irresponsible or counterproductive behaviors (e.g., disciplinary problems, disruptiveness on the job, tardiness, or excessive absenteeism)” (Ones et al., 1993). The authors did not find evidence for meaningful response distortion.

Two observations should be kept in mind when interpreting the findings of the two meta-analyses above. First, both studies are limited by the assumption that applicant and incumbent groups are on average similar in their true test mean scores, which may
not be realistic. Second, although the comparison between applicants and incumbents is common in faking research (see Hough, 1998, for a large scale study), it needs to be noted that intentional distortion is mainly understood as an individual level issue, and not a group level difference. Still, the available research seems to indicate that on average, faking in real-life situations may not be as big of an issue as once thought. However, that does not mean that there is no reason for concern.

In personnel selection literature, faking affects the rank order of job applicants (Rosse et al., 1998). Even when faking has low rates of occurrence – low enough that overall test validity is not threatened —, individuals who do fake may position themselves at the top of the test score distribution and increase their likelihood of being hired. The quality of selection decisions can be affected surprisingly easily: for example, Mueller-Hanson et al. (2003) found that fakers are disproportionally selected in settings where less than 60% of the total number of candidates are to be hired.

In some settings, coaching is another potential problem in real life personnel selection. For example, White, Young, Hunter, and Rumsey (2008) examined intentional distortion in tests developed to be used for entrance into the U.S. army. In this context, recruiters may coach applicants on how to answer the self-report items so that they pass the screening. Landers, Sackett, and Previsor (2011) also described the spread of coaching rumors among internal candidates within a nationwide retailer company. Some books and websites are also famous (infamous?) for offering advise on how to obtain ideal scores in psychometric tests (e.g., Hoffman, 2000; Parkinson, 2008; Personality Test Questions – Answers That Get You Hired!, 2022).
1.3.2 Is it possible to deter faking?

Given that participants can lie in self-report measures, and at least some participants will lie, it is of interest to determine if any methods are effective deterrents to faking efforts. Below I review research on some proposed methods.

1.3.2.1 Warnings and consequences. One of the earliest systematic investigations on the use of warnings and consequences is Doll’s (1971) experiment with Aviation Officer Candidates. The study design followed the instructed faking approach with three different conditions on the second time point. In addition, items had two categories: objective versus subjective (i.e., independently verifiable or not), and continuous versus noncontinuous (i.e., Likert or dichotomous response alternatives). At the first opportunity, all participants were oriented to answer a battery of noncognitive tests as honestly as possible. On the second administration, subjects were given one of three sets of instructions. The first was to “make yourself look good but keep in mind that you could be placed in a position to defend your answers in an interview situation”. The second set of instructions informed the participants that a lie score key would identify people who exaggerate their answers. The third set instructed subjects to make themselves look as good as possible. In the first two sets (denoted “subtle faking”), participants faked 17% and 20% of the objective items. For the subjective items, 35% and 24% faking rates were observed. In contrast, the unsubtle condition saw 97% of faking in objective items and 81% in subjective items.

Since Doll’s experiment, other research has investigated the effects of warning applicants of faking detection measures and of using verifiable items. Stanush (1997) found in her meta-analysis that warnings were effective in situations of real-life
motivational distortion (i.e., job applicants and incumbents), with mean difference scores of $\delta = -0.07$ when warnings were used, versus $\delta = 0.32$ when they were not\(^2\). In contexts of instructionally induced distortion, no significant effects were found. Another meta-analysis by Dwight and Donovan (2003) found that effect sizes can vary substantially depending on the type of warning. They identified three main types of warnings: identification-only, consequences-only, and identification combined with consequences. Effect sizes were $d = 0.01$, $d = 0.30$, and $d = 0.25$, respectively.

There is evidence that the combination of warning of lie scales and consequences has considerable upsides and not many disadvantages. Landers et al. (2011) found that warnings were effective in counteracting coaching rumors that spread within an organization. Vasilopoulos et al. (2005) found stronger correlations between personality measures and cognitive ability tests when using warnings in both field and laboratory studies. They also found slower item response times, suggesting an increase in response decision complexity. Some studies found no change in criterion related validity (Converse et al., 2008; Fox & Dinur, 1988; Robson et al., 2008), but McFarland (2003) found less multicollinearity among personality variables in conditions where respondents were warned that a social desirability scale was embedded in the test.

1.3.2.2 Written elaboration. Research on written elaboration has traditionally focused on biographical data (“biodata”). In the past this type of item was relatively objective and verifiable (e.g., “what is the highest educational degree you have obtained?”). Over time, biodata became often indistinguishable from personality

\(^2\) $\delta$ is a measure of effect size adjusted for unequal or unbalanced sample sizes. It is given by $\delta = d / A$, where “$A$” is bias multiplier defined as $A = 1 + (.75 / (N - 3))$, and $N$ is the average sample size across studies.
questions, and has now seemingly come full-circle with the use of information from CVs and application forms (Ramos-Villagrasa et al., 2022). Nevertheless, past research on written elaboration has found that this requirement results in significantly lower test scores by about .6 standard deviations on elaborated items alone and .46 on the overall test (Schmitt & Kunce, 2002). The items used were hard to verify, however. With verifiable items, Ramsay et al. (2006) found more modest effect sizes ($d = .23$).

1.3.2.3 Grouped items. In research contexts, it has been suggested that grouping and labeling items of the same construct can help the factor structure of an instrument (Schriesheim, Kopelman, et al., 1989; Schriesheim, Solomon, et al., 1989). In the context of personnel selection, however, it is possible that noncognitive measures would be considerably easier to fake (Morgeson et al., 2007). McFarland et al. (2002) investigated the hypothesis that grouping items together results in more faking in a Big Five measure. This was the case for some constructs, but not all. Under instructions to fake or to simulate an ideal candidate, participants distorted their responses more easily for measures of Emotional Stability (i.e., Neuroticism) and Conscientiousness when the corresponding items were grouped together. For the other constructs, the effects were non-significant.

1.3.3 Is it possible to detect faking?

One of the most studied operationalizations of response distortion is in the use of social desirability scales. In general terms, socially desirable responding is typically conceptualized as having a self-deception component and an impression management component (Paulhus, 1984). The former would represent “unconscious” distortion and the latter would correspond to the intentional distortion that here I call “faking.” As the
name implies, social desirability scales aim to capture the tendency to respond not to the item content, but to its socially desirable characteristics (Kuncel, Borneman, et al., 2011). An example of self-deception item would be “Have you ever enjoyed your bowel movements?” and an example of impression management item would be “When you take sick-leave from work or school, are you as sick as you say you are?”

Unfortunately, social desirability scales do not seem to be able to distinguish distorted variance from valid trait variance in applied settings. They seem to perform reasonably well in directed faking studies (e.g., Viswesvaran & Ones, 1999) but in practice correlate highly with emotional stability and conscientiousness. Consequently, attempting to partial out social desirability would also remove true variability from personality measures. Further, the scale scores do not correlate with subsequent job performance (Ones et al., 1996). Over the years, efforts in this line of investigation were drastically reduced (Jackson, 1970, 1989; Tellegen et al., 2006) and other alternatives have been proposed. I review some of them below.

1.3.3.1 Unlikely virtues or “lie scales”. Admittedly, unlikely virtue items have been discussed in the literature both as equivalent to social desirability items and as an alternative to them (e.g., Hough et al., 1990). Unlikely virtue scales are composed of items that in theory are extremely unlikely to be endorsed by a respondent that is engaged and thinking logically. A classic example is the True-or-false item “I never lie”, where an honest respondent is assumed to think “Everyone lies, including me, so I cannot endorse this item” (Kuncel, Borneman, et al., 2011). There is some evidence from directed faking studies that these items can detect intentional distortion. One meta-analysis found a

---

3 I hope the reader find these items as amusing as I do.
correlation of .09 between unlikely-virtues scales and other personality scales in honest conditions, and a correlation of .34 in faking conditions (Stanush, 1997). Another study found that in directed faking conditions, unlikely virtue scale scores increase above and beyond mean scores for other scales (Hough et al., 1990; Stanush, 1997). Overall, however, most research on social desirability and/or unlikely virtues scales – often called “lie scales” – has not provided sufficient justification for the use of these measures in high-stakes contexts (Ellingson et al., 1999, 2001; Hough, 1998; Hough et al., 1990; A. Li & Bagger, 2006; McGrath et al., 2010; Moorman & Podsakoff, 1992; Ones et al., 1996; Schmitt & Oswald, 2006).

One notable finding in a related line of research may help explain the limited success of this approach. In two studies, Kuncel and Tellegen (2009) found that social desirability is often nonlinearly associated with trait level. Further, perceived desirability changes across different work and personal life contexts. Even when instructed to fake in a maximally desirable way, participants do not always view the highest options (e.g., “strongly agree”) as the most desirable options and may instead select a more moderate alternative. For example, in work settings, people reported being maximally “well-organized” as the most desirable alternative, following a monotonic relationship. In general life settings, however, the desirability of being more highly organized generally declines at the high end in general life settings (Kuncel & Tellegen, 2009).

1.3.3.2 Over-claiming or “bogus statements”. The over-claiming technique (OCT) is a seemingly little-utilized method. Phillips and Clancy (1972) coined the term in a study where they asked participants to rate their familiarity with consumer items. In reality, none of the products existed, and the rated familiarities were taken to represent
self-serving response distortion. Stanovich and Cunningham (1992) performed a similar study using a list of authors where only half of the names were real. Randall and Fernandes (1991) went a step further and used over-claiming to control for bias in self-report measure of unethical behavior, and their over-claiming index correlated significantly with an inventory of socially desirable responding. More recently, Paulhus (2003) systematized the approach, but more research is necessary to understand the psychological mechanism behind the over-claiming responses and correlations with cognitive abilities.

1.3.3.3 Idiosyncratic item responses. Kuncel and Borneman (2007) proposed a method for detecting faking based on idiosyncratic item response patterns. This technique was based on scoring items based on their response distributions in a simulated applicant setting. They also used three pieces of information from a study that was later published in Kuncel and Tellegen (2009): (1) the relationship between trait level and desirability is nonlinear, (2) people do not universally agree on whether some items and traits are desirable or not, and (3) the relationship between trait level and desirability changes across different contexts. It was reasoned that during intentional response distortion, participants would avoid intermediate responses and focus on extreme endorsements if the desirability of the trait was ambiguous – in other words, participants guess what extreme response is more desirable. The scale had some success in detecting deliberate faking and was uncorrelated with valid trait scores. However, it should be noted that the authors used an adjective rating scale, and it is unknown whether other item-response options would perform satisfactorily. In addition, because the study participants were introductory undergraduate psychology students, it may be the case that real job
applicants may not perceive some of the items as ambiguous.

1.3.4 Measurement approaches to overcome faking

The concerns with response distortions have motivated many test development efforts. In this section, I describe several lines of research that aim to increase the validity and the accuracy of predictions using SRIs.

1.3.4.1 Item characteristics under criterion validity. Research in this area is indirectly related to faking, but looks into a pertinent question: what are the item characteristics for the measures that do predict relevant criteria such as job performance and learning in training? In line with studies discussed previously, Schmidt and Hunter (1998) found in their meta-analysis that verifiable measures such as tests of general mental ability, work sample tests, and integrity tests had the largest correlations with overall job performance and performance in job training. In a meta-analysis of studies on self-evaluation of abilities (for example, scholastic, athletic, or interpersonal skills), Mabe and West (1982) found that test instructions that use social comparison terminology (e.g., “compared to your friends, do you...”) tend to improve the predictive validity of test scores. These two approaches differentiate themselves in that unlike some of the other methods, they involve changes in the logistics of the assessment (verifying information) and test instructions, and not so much on the items themselves.

1.3.4.2 Forced-choice measurement. Forced-choice items are a response format alternative to the more common Likert (1932) format. Likert items typically consist of a statement (e.g., I am the life of the party) to which respondents need to indicate their degree of agreement in a rating scale – most commonly, five points from strongly disagree to strongly agree. Forced-choice items on the other hand present more than one
statement and ask the respondent to choose the statement that is most (or least) descriptive of them. An example of a forced-choice item is “Which is more descriptive of you? I make friends easily or I follow a schedule?

Early research on forced-choice tests suggested that pairing equally desirable statements would effectively reduce susceptibility to faking (Gordon, 1951). However, there is some evidence from directed faking studies with forced-choice instruments that respondents can successfully fake when instructed to do so (Waters, 1965), and to a degree similar to the Likert format (Stanush, 1997). Also under directed faking, forced-choice scale scores retain predictive validity (Christiansen et al., 2005; Hirsh & Peterson, 2008), but at the cost of increased correlations across personality factors and cognitive general cognitive ability (Christiansen et al., 2005; Vasilopoulos et al., 2006). Thanks to new developments on forced-choice scoring (Brown & Maydeu-Olivares, 2011b, 2012) there has been renewed interest in this type of scale (Pavlov et al., 2019), but this response format will not be the focus of this study.

1.3.4.3 Subtle item content. Subtle items are those for which the trait that they are supposed to measure is not obvious. It has become a common technique to provide some frame of reference to items with the goal of increasing predictive and criterion-related validity as well as reliability (Bing et al., 2004; Hunthausen et al., 2003; Lievens et al., 2008; Schmit et al., 1995). However, there is also research indicating that subtle items have higher validity in high-stakes applicant settings (such as high security clearance military personnel), potentially due to the difficulty in distorting responses to those items given the ambiguity of what is being assessed (White et al., 2008).

1.3.4.4 Conditional reasoning measurement. The conditional reasoning framework was
first proposed by James (1998) and further developed in James et al. (2004). They were primarily interested in aggressive behaviors and proposed the term *justification mechanisms* (JMs), which are understood as implicit/unconscious bias that enhance the rational appeal of aggression. For example, the authors propose a *Retribution bias* that represents the “tendency to confer logical priority to reparation or retaliation over reconciliation”, which would be used by aggressive individuals to make their behavior seem justifiable. The items on the measure consist of scenarios that prime these JMs. In two studies with undergraduate students and patrol officers, the average criterion validity of the Conditional Reasoning-Aggression Scale was of $r = .44$ while comparable self-report instruments had on average $r = .30$ (James et al., 2004). One metanalysis has found more conservative estimates (average $r = .26$ when predicting counterproductive work behavior; Berry et al., 2010), but it is an interesting approach that may prove useful in countering faking through the evaluation of implicit biases.

All in all, there are varying degrees of success in dealing with faking, but no universal solution exists. In the next section, we will lay the foundation for the statistical method proposed for the current study.

### 1.4 Statistical background

In any field of science, the populations under investigation are often heterogeneous. The source of that heterogeneity is often known because it is artificially created (e.g., treatment and control groups in an experimental design), or because it is observable (e.g., grade level, country of origin). From a statistical point of view, these groups can be compared on their mean levels of a dependent variable using classic methods such as multivariate analysis of variance (MANOVA) and multiple regression.
with dummy predictors, for example. If the dependent variables are measured with error, a latent variable approach such as multiple group confirmatory factor analysis (MGFA) can be used, where observed scores are decomposed into residual variance and variance that is predicted by the latent variable that represents the construct of interest. Each of these measurement models is then fit to their respective groups. If group membership itself is the dependent variable and we want to determine the probability of belonging to each observed group given certain predictors, logistic regression (LR) or discriminant analysis (DA) are suitable approaches.

If, however, heterogeneity is inferred but unobserved, it is impossible to directly allocate sample participants into groups. The proportion of individuals within each class needs to be estimated. In the social and behavioral sciences, these discrete unobserved groups are often called latent classes\(^4\). Popular methods to handle unobserved heterogeneity include K-means clustering, latent class analysis (LCA), latent profile analysis (LPA), and factor mixture models (FMM). The K-means method minimizes the within-cluster variability while maximizing between-cluster variability, with the number of clusters being a pre-specified, arbitrary number. The non-model based nature of K-means is an advantage in the sense that no a priori theory is required, but it is also a drawback in the sense that it can be difficult to perform model selection by comparing alternative models. This method also does not account for measurement error. LCA and LPA do not have these limitations, allowing easy model comparison and accounting for measurement error – however, they introduce a local independence assumption, meaning that all covariance amongst observed variables is assumed to be due to class differences.

\(^4\) I use the terms unobserved and latent interchangeably.
FMM then emerges as the ideal candidate to model faking, as it can account for measurement error in the substantive factor(s) of interest (for example, the Big Five) in addition to handling unobserved heterogeneity through a latent class variable. Further, FMM can also incorporate covariates so that observed variables, substantive latent factor, and the latent class can each account for the same or different background variables.

1.4.1 Factor analysis

A factor analysis (FA) model is simply a multivariate multiple regression model where the predictor (in this context called factor) is an unobserved continuous variable. The model can be written algebraically as

\[ y_{ij} = \lambda_i \eta_j + \epsilon_{ij}, \]  

(1)

where \( y_{ij} \) is the observed response to item \( i \) by respondent \( j \), \( \eta \) represents the unobserved attribute, \( \lambda_i \) is a slope (factor loading), \( \nu_i \) is the item intercept, and \( \epsilon_{ij} \) the variance of the error. Equation 1 can be represented in matrix form as

\[ y = \nu + \Lambda \eta + \epsilon, \]  

(2)

where \( y, \nu, \) and \( \epsilon \) are \( p \times 1 \) vectors. In addition, \( \Lambda \) is a \( p \times m \) matrix of factor loadings and \( \eta \) a \( m \times 1 \) vector of factor scores, where \( m \) is the number of dimensions (factors) in the model. For simplicity, here I discuss only the unidimensional case. Under certain assumptions\(^5\), Equation 2 implies a mean vector

\[ \mu_y = \mu, \]  

(3)

and a covariance matrix of the items

---

\(^5\)The assumptions of the common factor model are (1) the mean of each factor is zero, (2) the mean of the errors is zero, (3) error terms have zero autocorrelation, and (4) error terms are uncorrelated with the factors (Maydeu-Olivares & Coffman, 2006). In CFA, there may be correlated errors.
\[ \Sigma_y = \Lambda \Psi \Lambda' + \Theta, \]  

where \( \Psi \) is the \( m \times m \) covariance matrix of the latent factors and \( \Theta \) is the \( p \times p \) covariance matrix of error terms. In confirmatory factor analysis (CFA) models, exclusion constraints are imposed on \( \Lambda \) and \( \Psi \) (i.e., some loadings are fixed to zero based on substantive theory).

Because the factor is latent and only inferred from observed information, it is necessary to arbitrarily impose restrictions to set the scale of measurement. Two approaches are the unit loading identification (ULI), where one item has its factor loading fixed to one, and unit variance identification (UVI), where we fix the factor’s mean and variance to zero and one, respectively. In the UVI approach, this means that we are performing multivariate regression with a single predictor where the outcomes (items) are unstandardized, and the predictor (factor) is standardized. After we estimate the model, we can obtain factor scores. Because the factor has been standardized, the factor scores are standardized. In this case, the unstandardized loadings are interpreted as the item score change for a change of one standard deviation in the factor. In addition, a factor score of one (for example) indicates that the respondent is one standard deviation of the factor above the mean of factor.

An alternative is to estimate unstandardized factor scores, giving the factor the same metric of the items. One approach for that is the ULI mentioned previously, where one item’s loading is fixed to one. Another approach that can be used is to compute the average of the items for all respondents, and then obtain the mean and variance of that average. These values are then used as the mean and variance of the factor. In that case, the factor scores are on the same scale as the items and can be interpreted as the items,
which can be convenient in test development. For example, if items for Extroversion are in a 1-5 Likert scale, a factor score close to one indicates that the respondent’s average Extroversion is represented by the item responses labeled “1”.

1.4.2 Multiple group factor analysis

A multiple group factor analysis model is simply a CFA model with an additional variable: group membership. Equations 3 and 4 are modified to denote group membership:

\[
\mu_g = \nu_g, \quad \Sigma_g = \Lambda_g \Psi_g \Lambda_g' + \Theta_g
\]

Where all matrices are the same as defined previously, with the \(g\) subscript indicating matrices derived from the \(g^{th}\) sample. The simultaneous estimation of structural equation models (SEMs) across groups that have their own mean vector and covariance matrices allows for the investigation of measurement invariance, by testing if parameters can be set equal across groups.

1.4.3 Latent class and latent profile analysis

Latent class analysis (LCA) and latent profile analysis (LPA) are used to identify unobserved subgroups (classes) of a study population. They are similar to a factor analysis model, except the latent variable is an unordered, categorical variable with \(K\) classes. The counterpart of factor scores in LCA and LPA are the estimated class probabilities of each individual, called posterior probabilities. Respondents are assigned a class membership based on their highest probability of being in a certain class, given their observed response to the items. Once the assignment is done, individuals are assumed to belong exclusively to that class. Items are categorical in LCA and continuous in LPA.
Critically, LCA and LPA assume local independence. This means that any correlation among the observed responses is explained by the latent class. In other words, there are no residual correlations between the indicators within a class and individuals within a class are homogeneous. When the study of unobserved class membership is of interest and the assumption of conditional independence is overly restrictive, it is possible to put forward a model where a within-group factor structure exists, such as the factor mixture model.

1.4.4 Factor Mixture Model

In its simplest form, the factor mixture model (FMM) incorporates a latent subgroup variable and an FA model. The sources of population heterogeneity are represented by the latent class variable, while within each latent class the factor model accounts for intraclass variability. Factor scores on the trait of interest can also be obtained. For this reason, FMMs have been described as a combination of LCA and FA (Clark, 2010; Clark et al., 2013). Historically, however, FMM has been developed more as an extension of univariate (Pearson, 1894) and multivariate (Wolfe, 1970, 1971) normal mixture models. In finite normal mixture models, homogeneity is defined at the distributional level (Bauer & Curran, 2004), via a common mean vector and covariance matrix. Blåfield (1980) and Yung (1997) then took the approach of structuring that mean and covariance structure using a continuous latent variable model, where a within-class covariance matrix \( \Sigma_k \) is used following Equation 4. In other words, an FMM can also be seen as a multiple group factor analytic model where the grouping variable is unknown.

In an FMM, the observed variables relate to each other through two unobserved variables: the \( m \)-dimensional vector \( \eta \) of latent continuous variables previously
described, and a $K$-dimensional vector of vector $c$ of latent unordered categorical variables (Muthén & Shedden, 1999). For a model with two classes, the variable $C$ reduces to a Bernoulli variable, so that $c_{ik} = 1$ if respondent $i$ belongs to class $k$, or $c_{ik} = 0$ otherwise. The least restrictive measurement and structural models without covariates can then be expressed by

$$y_{ik} = \mathbf{v}_k + A_{yk} \eta_{ik} + \varepsilon_{ik},$$

(6)

$$\eta_{ik} = A_k c_i + \zeta_{ik},$$

(7)

Where the $k$ subscript indicates how parameters may vary across classes, $A$ is a $m \times k$ matrix that contains the factor intercepts for each class, and $\zeta_{ik}$ is a $m$-dimensional vector of residual factor scores, or the factor variance not explained by the latent class (Lubke & Muthén, 2005). Estimation is typically performed by maximum likelihood (Muthén & Shedden, 1999).

Class membership can be parameterized as the log odds of the probability of belonging to a given class compared to the probability of belonging to the other class. Specifically, they are the intercepts of a multinomial regression model. If no covariates are used, the multinomial regression equation is composed of intercepts exclusively; in other words, the class weights are determined by the intercepts alone:

$$\ln \left[ \frac{P(c_{ik} = 1)}{1 - P(c_{ik} = 1)} \right] = \frac{\exp(\alpha_k)}{\sum_{k=1}^{K} \exp(\alpha_k)},$$

(8)

where $\alpha_k$, the intercept for the last class, is set to zero. Setting all starting values for the class specific intercepts to zero corresponds to equal population class proportions.

After all parameters are estimated, factor scores and estimated class membership
are computed. Given the model, the posterior probability of belonging to each class is computed, and individuals are assigned to the class with the highest probability. This assumes equal prior probability of belonging to each class.

As for factor scores, the most widely used method with continuous items is the Regression Method (Thomson, 1934; Thurstone, 1935). Under normality assumptions for the factors, these factor scores are also the maximum of the posterior distribution of the factor (aka maximum a posteriori, MAP), as well as the mean of the posterior distribution of the factor (aka expected a posteriori, EAP). Class specific factor scores are estimated for individual $i$ using

$$\hat{\eta}_i = (y_i - \nu)' W, \quad W = \Theta^{-1}(\Psi^{-1} + \Lambda'\Theta^{-1}\Lambda)^{-1}. \quad (9)$$

The class specific scores for the class with highest posterior probability are provided, along with a score obtained by weighting the class specific scores using the posterior probabilities.

1.4.5 Previous uses of mixture modeling on the faking literature

The potential of mixture modeling to account for faking has not gone unnoticed, but few studies investigated this type of application. Zickar et al. (2004) used a mixture model within an item response theory framework to analyze honest, incumbent, and instructed faking responses. More specifically, polytomous ordinal data was analyzed with a partial credit model (Masters, 1982), an extension of the Rasch model. Rasch models have the benefit of simplicity as selected items have equal discrimination, but this restriction also severely limits the choice of items. Critically, in the context of faking this implies that faking behavior is consistent across all items.

Ziegler et al. (2015) investigated a latent difference score model (LDSM), an
FMM, and a combination of the two to analyze honest and applicant responses. The first model represents a quantitative view of faking, representing it as a change in scores from one time of point to another (i.e., from honest to faking), and the second represents faking as a qualitative difference in terms of distinct classes (i.e., honest respondents and faking respondents). While insightful, this study was limited by the restriction of equal loadings across classes (similarly to a Rasch model). In addition, the authors did not propose any method to extract honest scores when only assessment scores are available.

More recently, Brown and Böckenholt (2022) proposed the use of a grade-of-membership model (Asparouhov & Muthen, 2017), a specific FMM that allows for partial class membership, to model responses under a Retrieve-Edit-Select framework (Böckenholt, 2014). Their findings show that this method has promise, but estimation can be demanding because at least three latent variables and multilevel modeling requiring integration over the latent variables are needed. In addition, more simulation results are needed to demonstrate the generalizability of the method.
Figure 1.1 Selected contributions in the history of the Big Five model of personality.
CHAPTER 2: METHODS

2.1 Preliminary data analysis

To investigate the feasibility of FMM to counter faking, a preliminary analysis was performed using Pavlov et al.’s data (2019). Participants consisted of 180 undergraduate students in a large public university in Spain that responded to the same personality test (Brown & Maydeu-Olivares, 2011a) twice: first under honest instructions (“baseline” or “honest”), and then under instructions to imagine that they were applying for a particular job opening (“assessment” or “faking”). Participants were also asked to report how much they believe they distorted their responses at the second time point in a 1-10 scale, among other measures. While Pavlov et al. used both forced choice and Likert items, here I only use the Likert data.

As shown in Figure 2.1, a regression model clearly shows a multiplicative relationship between honest scores, faking scores, and tendency to fake. Given this finding and the hypothesis that FMM is a reasonable approximation to the data generating process, the main objective of this research is to answer the question: *In the presence of faking, does FMM provide more accurate scores than scores that ignore faking?* In other words, does the effort to fit an FMM to assessment data result in scores that have higher correlation with true scores?

2.1.1 What is the effect of faking on test scores?

To understand the effect of faking on estimated factor scores, the correlation between baseline and assessment scores was computed under different models. For all
analysis, two groups were arbitrarily defined: honest and fakers. The honest group is composed of respondents that on average changed their responses (from baseline to assessment) by less than .5 in a Likert scale with range [-2, 2]. Fakers are all remaining participants. This is represented in Figure 2.1, with honest respondents represented by data points within the diagonal grey band, and fakers outside of the same band. I used Extroversion mean scores for this demonstration.

When considering only the honest respondents, the correlation between baseline and assessment scores was $r = .94$. The same analysis for the subset of fakers revealed a correlation of $r = .45$, indicating a much weaker relationship between their true scores and what is suggested by the participants’ responses. Alarmingly, the overall correlation of baseline and assessment scores across participants (honest and fakers) was not much higher at $r = .48$.

The next step was to consider scores obtained from FMM. A few points should be noted. The available items are ordinal, and the latent factor is continuous. Therefore, the model is strictly incorrect because the relationship between ordinal outcomes and continuous predictors cannot be linear. The FMM was used as an approximation only and future work should consider how to extend this approach to a non-linear model (i.e., using ordinal factor analysis instead of linear FA). In addition, the interaction with tendency to fake is continuous (Pavlov et al., 2019; Ziegler et al., 2015) so that an infinite number of lines could be fit. The use of two classes is only a simplification.

When fitting the FMM, the limited sample size ($n = 180$) proved to be a challenge to convergence. To ameliorate this issue, an MGFA model was first fit to the data, using the arbitrary definition of honest and faking respondents as the grouping variable. Then,
the parameter estimates from the MGFA model were used as starting values for the mixture model. The correlation between class specific factor scores (obtained from the assessment responses) and the baseline responses was \( r = .63 \). When considering the \( r = .48 \) obtained when ignoring faking effects, this suggests that FMM scores are an improvement over ignoring faking effects.

**2.1.2 How well does a mixture model work relative to a multiple group approach?**

If faking is known to exist, the MGFA model can be considered the best-case scenario in obtaining accurate factor scores because group membership is known. Therefore, it is interesting to determine how close the FMM scores (where group membership is unknown) approximate the MGFA values. As a first effort in this direction, I generated a single data set under an MGFA model with \( N = 500 \). Of the total number of observations, 440 (88\%) were of the “honest” group. The parameter values used were similar to those obtained in the analysis of Pavlov et al.’s data. For all items in the honest group, \( \lambda_{i1} = .6 \), \( \nu_{i1} = .2 \), and \( \epsilon_{i1} = .3 \); in the faking group, \( \lambda_{i2} = .25 \), \( \nu_{i2} = 1.2 \), and \( \epsilon_{i2} = .3 \). Then, both an MGFA and a FMM were fit to the data without misspecifications. In the FMM case, the true population values were used as starting values. For the faking-only subset, the correlation between true and faking scores was \( r = .80 \); while the correlation for the honest-only group was \( r = .97 \).

The correlation across all observations (ignoring group membership) was \( r = .85 \). For the MGFA, as expected factor scores were highly correlated with the true values, \( r = .94 \). However, FMM scores had a correlation of only \( r = .74 \) with honest scores, indicating worse performance than the lack of any control for faking at all.

These relationships can be visualized in Figures 2.2 and 2.3: note that the line for
MGFA scores is very close to the hypothetical, “perfectly honest”, diagonal dashed line. This indicates that MGFA scores recover the true scores from the potentially faked responses very well. In the FMM case, the model line is substantially far from the ideal. On Figure 2.3, it can also be seen how the model misclassified several observations (represented by the blue points close to the red line). This suggests that the source of the inaccuracy in scores is the misclassification of individuals. Therefore, to fully address my research questions, it is important to perform a Monte Carlo study manipulating different variables to determine under what conditions FMM may prove viable to detect faked responses.

2.2 Simulation (Monte Carlo) study

In recent years, many simulation studies have looked into the performance of mixture models. Some of these include studies that investigated regression models (Jaki et al., 2019), growth models (Diallo et al., 2016, 2017; Li & Hser, 2011; Peugh & Fan, 2012), LPA (Peugh & Fan, 2013; Tein et al., 2013) and LCA (Wurpts & Geiser, 2014), and the structural paths of mixture SEMs (Tueller & Lubke, 2010). Others focused on particularities such as the use of covariates (Diallo et al., 2017; Kim et al., 2016; Li & Hser, 2011; Nylund-Gibson & Masyn, 2016), the performance of different estimation approaches (Depaoli, 2013) and of initialization strategies (Shireman et al., 2017).

However, since Yung’s breakthrough work (1994), only a few studies looked into the performance of factor mixture models specifically (Buzick, 2010; Lubke & Muthén, 2007; Lubke & Neale, 2008, 2006; Tueller et al., 2011).

As is usual with simulation studies, only a limited number of conditions can be investigated. Yung’s work (1994) focused on parameter estimation, and a Monte Carlo
study was used to determine if asymptotic results could be applied to finite samples as well. Critically, that study only considered “k-Mixtures of Common-Regression CFA Models”, which assume item intercepts and factor loadings invariant across classes. Lubke and Neale (2006) considered only exploratory factor models with continuous outcomes, focusing on model selection. A follow-up study considered similar research questions when the dependent variables were discrete (Lubke & Neale, 2008). Lubke and Muthén (2007) reported three different studies, but generalizability of their results to faking applications is limited because in each study either (a) item loadings, item intercepts, and residual variances were invariant across classes, or (b) covariates that predict class membership or multidimensional models were used. In addition, only a 50/50 split across two classes was investigated. Buzick (2010) manipulated factor loadings but not intercepts and residuals. Finally, Tueller et al. (2011) addressed the problem of switched class labels in latent variable mixture model simulation studies, and developed an R (R Core Team, 2019) package to detect such occurrences. While relevant to the present study, it is unrelated to applied problems.

2.2.1 Simulation study design

To investigate the feasibility of using FMM to counter faking, I used simulation studies where the data generating model approximates data that contains faked responses, as inferred from the preliminary analysis. All data was generated in R (R Core Team, 2019), with estimation in Mplus (Muthén & Muthén, 2017a) using maximum likelihood (Muthén & Shedden, 1999), and scoring again done in R. Only continuous normal outcomes will be considered. All code is available in Appendices A through E.

In the first part of the simulation study, I investigate how well the factor mixture
model recovers the data-generating model. Data is generated under a two-group one factor model with the total sample size split 70% honest/30% faking. This split corresponds to the split found empirically in Pavlov et al.’s study. Generating data with a two-group approach has the advantage of keeping the group sizes constant across all replications. Each dataset will be used to estimate two models. The first model is a mixture model where the grouping variable that identifies the classes is not used, meaning that class membership is estimated as described in the Introduction. The second model is a multiple group CFA where the known group membership is incorporated in the analysis. The factor scores of the multiple group CFA model and their correlations with the true factor scores will be used as the ceiling or “gold standard” when evaluating the results of the mixture model. The baseline will be mean scores: these scores are representative of choosing to ignore faking effects. This means that for factor mixture scores to be worthwhile, they must have higher correlation with true scores compared to mean scores.

In both multiple group CFA and factor mixture approaches, the models are correctly specified. Only one dimension is used, with factor mean and variance fixed to (0, 1) in both groups and in all models. Within the simulation setup, the mixture model only has one more parameter than the multiple group CFA model, the proportion of “honest” respondents in a logit scale. Also, in the mixture models, the true data-generating parameters are used as starting values. Parameter values were obtained by approximating parameter estimates of a multiple group CFA model fit to Pavlov et al.’s (2019) data.

The following conditions will be manipulated (unstandardized values):
1. Number of items: $p = 6, 12, \text{ or } 18.$

2. Total sample size: $N = 500, 1000, \text{ or } 2000.$

3. Difference in loadings: .2 or .6 ($\lambda_{\text{honest}} = .5 \text{ or } .9; \lambda_{\text{faking}} = .3$).

4. Difference in item intercepts: 0 or 1.0 ($\psi_{\text{honest}} = 0 \text{ or } .5; \psi_{\text{faking}} = 1.5$).

5. Difference in item residual variances: .3 or .6 ($\varepsilon_{\text{honest}} = .3 \text{ or } .6; \varepsilon_{\text{faking}} = .9$).

The correspondent standardized values are the following. For the faking class, $\lambda_{\text{faking}} = .303$ and $\varepsilon_{\text{faking}} = .908.$ For the honest class, we have the following combinations:

a. $\lambda_{\text{honest}} = .5$ with $\varepsilon_{\text{honest}} = .3$ (standardized loading = .909 and standardized uniqueness = .174).

b. $\lambda_{\text{honest}} = .5$ with $\varepsilon_{\text{honest}} = .6$ (standardized loading = .588 and standardized uniqueness = .654).

c. $\lambda_{\text{honest}} = .9$ with $\varepsilon_{\text{honest}} = .3$ (standardized loading = .811 and standardized uniqueness = .342).

d. $\lambda_{\text{honest}} = .9$ with $\varepsilon_{\text{honest}} = .6$ (standardized loading = .638 and standardized uniqueness = .593).

All in all, $3 \times 3 \times 2 \times 2 \times 2 = 72$ conditions are evaluated. For all models, 1000 data sets (i.e., Monte Carlo replications) will be generated. In the optimization process for model estimation, the number of initial random starts and final stage optimizations are set to 1000 and 250, respectively.

2.3 Outcomes

FMM factor scores (for most likely class and weighted) will be correlated with
the true simulated scores. The goal of such analyzes is to determine if FMM-scores – which in theory controls for faking effects – perform better than the current practice of ignoring faking effects (i.e., mean scores). Because correct class assignment is critical to obtaining valid factor scores, the performance of the classification algorithm will also be evaluated.

2.4 Empirical application

To investigate the feasibility of applying factor mixture modeling to faking data, I use data collected by Wetzel, Frick, and Brown (2021). It contains responses from 622 participants to a rating scale version of the Big Five Triplets (BFT; Wetzel & Frick, 2020). Each participant filled out the BFT under honest instructions and again under faking instructions.
Figure 2.1. Regression model for Extroversion mean scores with interaction by faking (a dummy variable). The dashed line is a true diagonal (i.e., a theoretical one-to-one relationship between honest and assessment scores). The grey band represents a .5 deviation in mean scale scores from honest to faking scores in a Likert scale with range [-2, 2]. Observations within the grey band are respondents considered honest. “Main effect” represents the regression line of assessment scores on baseline scores where the interaction tendency to fake × honest score is not accounted for.
Figure 2.2. Simulated data example using a two-group factor analysis model and analyzed using an MGFA. The yellow line ("multiple group score") represents the estimated factor scores under the same MGFA model. The dashed line is a true diagonal (i.e., a theoretical one-to-one relationship between honest and assessment scores). Note that here group membership is known.
Figure 2.3. Simulated data example using a two-group factor analysis model and analyzed using an FMM model. The yellow line represents the class specific scores for the class with highest posterior probability. The dashed line is a true diagonal (i.e., a theoretical one-to-one relationship between honest and assessment scores). Note the misclassification of some “faking” observations as “honest” in the left side of the plot.
CHAPTER 3: SIMULATION RESULTS

The results of all simulation conditions are summarized in Tables 3.1-3.6. The average correlation between true factor scores and estimated factor scores is presented for the following methods: mean scores (i.e., ignoring the effect of faking), multiple group factor analysis scores, factor mixture scores (most likely class), and factor mixture scores (class probability-weighted). In addition, several classification indices are discussed: accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. Results in Tables 3.1-3.6 show that overall, factor mixture scores have higher correlations with true factor scores in models of larger size ($p > 18$) and when larger sample sizes are available ($N = 2000$); classification also tends to improve across the board under the same conditions. FMM performance also generally tends to improve under higher class separation conditions: higher discrepancy between factor loadings, item intercepts, and uniqueness between the two classes. However, interpretation of these findings requires caution due to interactions between the simulation conditions. An analysis of variance (ANOVA) showed significant two-way interactions for model size $\times N (p < .001)$, $N \times \varepsilon_{honest} (p < .05)$, $\lambda_{honest} \times \nu_{honest} (p < .001)$, and $\lambda_{honest} \times \varepsilon_{honest} (p < .01)$.

3.1 Mean scores

3.1.1 Model with $p = 6$

In the small-size model ($p = 6$), the correlation between true factor scores and mean scores is seemly independent of sample size in the studied range ($N = 500 – 2000$). On the other hand, the standard deviation of the estimated correlation across 1000
replications is reduced approximately in half when sample size increases from 500 to 2000 and all else remains constant. For example, when $\lambda_{\text{honest}} = .5$, $\psi_{\text{honest}} = 0$, and $\varepsilon_{\text{honest}} = .3$, the correlation between true scores and mean scores is $r = .506$ for $N = 500$ and $N = 1000$, and $r = .507$ for $N = 2000$. Meanwhile, their respective standard deviations are .034, .024, and .016. For this reason, the discussion below focuses on the conditions where $N = 2000$.

With higher factor loadings such that the separation between honest and faking classes is also higher, the correlations with true scores also increase. These mean scores were obtained ignoring group membership (as in applications this is not available) so this increase is simply a reflection of stronger relationships between item and factor for that 70% of the sample size. For conditions where the item intercepts $\psi_{\text{honest}} = 0$ so that separation from the faking class is higher, correlations increase by about .17 when $\lambda_{\text{honest}}$ increases from .5 to .9: when $\varepsilon_{\text{honest}} = .3$, we observe $r = .507$ increase to $r = .673$, and when $\varepsilon_{\text{honest}} = .6$ we observe $r = .495$ increase to $r = .662$. For conditions where item intercepts $\psi_{\text{honest}} = .5$ so that class separation is lower, correlations increase by about .15 when $\lambda_{\text{honest}}$ increases from .5 to .9: when $\varepsilon_{\text{honest}} = .3$, we observe $r = .628$ increase to $r = .766$, and when $\varepsilon_{\text{honest}} = .6$ we observe $r = .606$ increase to $r = .751$.

In terms of item intercepts, we observe the same pattern as with the loadings: an increase in intercepts leads to an increase in true-mean score correlations. In conditions where $\lambda_{\text{honest}} = .5$, correlations increase by about $r = .12$ when $\varepsilon_{\text{honest}} = .3$. Increasing the intercept from $\psi_{\text{honest}} = 0$ to $\psi_{\text{honest}} = .5$ leads to an increase in correlations from an
average of $r = .507$ to $r = .628$; for $\varepsilon_{honest} = .6$, correlations increase from $r = .495$ to $r = .606$. In conditions where $\lambda_{honest} = .9$, we see an increase of around $r = .09$ when $\varepsilon_{honest} = .3$. When the intercept increases from $\psi_{honest} = 0$ to $\psi_{honest} = .5$, correlations increase from average of $r = .673$ to $r = .766$; for $\varepsilon_{honest} = .6$, correlations increase from $r = .662$ to $r = .751$.

Finally, we see an inverse relation as residual variance (aka uniqueness) increases: correlations between true factor scores and mean scores are higher when residuals are lower ($\varepsilon_{honest} = .3$). The largest discrepancy occurs when class separation is the lowest ($\lambda_{honest} = .5$, $\psi_{honest} = .5$) where score correlation decrease from $r = .628$ to .606. For other combinations of loadings and intercepts, the changes in correlation are generally below .10.

### 3.1.2 Model with $p = 12$

In line with the results for the small model size ($p = 6$), for the model of intermediate size ($p = 12$) the correlation between true factor scores and mean scores is seemingly independent of sample size in the studied range ($N = 500 – 2000$). All correlations are however larger than their counterparts in the small model, and their associated standard deviations slightly smaller though not in a meaningful way (for example, changing from .011 to .010). The discussion focuses again on conditions where $N = 2000$.

With higher factor loadings such that the separation between honest and faking classes is also higher, the correlations with true scores also increase. When the item intercepts $\psi_{honest} = 0$ so that separation from the faking class is higher, correlations
increase by about .16 when $\lambda_{honest}$ increases from .5 to .9: when $\epsilon_{honest} = .3$, we observe $r = .520$ increase to $r = .684$, and when $\epsilon_{honest} = .6$ we observe $r = .514$ increase to $r = .679$. When the item intercepts $\nu_{honest} = .5$ so that class separation is lower, correlations increase by about .13 when $\lambda_{honest}$ increases from .5 to .9: when $\epsilon_{honest} = .3$, we observe $r = .655$ increase to $r = .783$, and when $\epsilon_{honest} = .6$ we observe $r = .643$ increase to $r = .775$.

In terms of item intercepts, we observe the same pattern as with the loadings: an increase in intercepts leads to an increase in the correlations between true factor scores and mean scores. In conditions where $\lambda_{honest} = .5$, correlations increase by about $r = .13$ when $\epsilon_{honest} = .3$. Increasing the intercept from $\nu_{honest} = 0$ to $\nu_{honest} = .5$ leads to an increase in correlations from an average of $r = .520$ to $r = .655$; for $\epsilon_{honest} = .6$, correlations increase from $r = .514$ to $r = .643$. In conditions where $\lambda_{honest} = .9$, we see an increase of around $r = .09$ when $\epsilon_{honest} = .3$. When the intercept increases from $\nu_{honest} = 0$ to $\nu_{honest} = .5$, correlations increase from average of $r = .684$ to $r = .783$; for $\epsilon_{honest} = .6$, correlations increase from $r = .679$ to $r = .775$.

Finally, with item residuals we see again an inverse relationship: correlations between true factor scores and mean scores are higher when residuals are lower ($\epsilon_{honest} = .3$). The largest discrepancy - for the lowest class separation ($\lambda_{honest} = .5$, $\nu_{honest} = .5$) – is a correlation decrease from $r = .655$ to .643. Note that this difference ($\Delta = .012$) is smaller than what we previously found ($\Delta = .022$) in the models where $p = 6$. For other combinations of loadings and intercepts, the changes in correlation are again below .10.
3.1.3 Model with \( p = 18 \)

Once again in line with previous results, for the model of large size \((p = 18)\) the correlation between true factor scores and mean scores is largely independent of sample size in the studied range \((N = 500 – 2000)\). While correlations between mean and true scores are larger than their counterparts in the intermediate-sized model, this difference is of \( r = .012 \) at most. Their associated standard deviations are again slightly smaller though not in a meaningful way (at most .002). The discussion focuses again on conditions where \( N = 2000 \).

With higher factor loadings such that the separation between honest and faking classes is also higher, the correlations with true scores also increase. When the item intercepts \( \psi_{honest} = 0 \) so that separation from the faking class is higher, correlations increase by about .16 when \( \lambda_{honest} \) increases from .5 to .9; when \( \epsilon_{honest} = .3 \), we observe \( r = .525 \) increase to \( r = .688 \), and when \( \epsilon_{honest} = .6 \) we observe \( r = .521 \) increase to \( r = .685 \). When the item intercepts \( \psi_{honest} = .5 \) so that class separation is lower, correlations increase by about .13 when \( \lambda_{honest} \) increases from .5 to .9; when \( \epsilon_{honest} = .3 \), we observe \( r = .664 \) increase to \( r = .790 \), and when \( \epsilon_{honest} = .6 \) we observe \( r = .655 \) increase to \( r = .784 \).

In terms of item intercepts, we observe the same pattern as with the loadings: an increase in intercepts leads to an increase in true-mean score correlations. In conditions where \( \lambda_{honest} = .5 \), correlations increase by about \( r = .13 \) when \( \epsilon_{honest} = .3 \). Increasing the intercept from \( \psi_{honest} = 0 \) to \( \psi_{honest} = .5 \) leads to an increase in correlations from an average of \( r = .525 \) to \( r = .664 \); for \( \epsilon_{honest} = .6 \), correlations increase from \( r = .521 \) to \( r = \)
.655. In conditions where \( \lambda_{honest} = .9 \), we see an increase of around \( r = .10 \) when \( \varepsilon_{honest} = .3 \). When the intercept increases from \( \nu_{honest} = 0 \) to \( \nu_{honest} = .5 \), correlations increase from average of \( r = .688 \) to \( r = .790 \); for \( \varepsilon_{honest} = .6 \), correlations increase from \( r = .685 \) to \( r = .784 \).

Finally, with item residuals we see again an inverted pattern: correlations between true factor scores and mean scores are higher when residuals are lower (\( \varepsilon_{honest} = .3 \)). The largest discrepancy - for the lowest class separation (\( \lambda_{honest} = .5 \), \( \nu_{honest} = .5 \)) – is a correlation decrease from \( r = .664 \) to \( r = .655 \). Note that this difference (\( \Delta = .009 \)) is smaller than what we previously found (\( \Delta = .012 \)) in the models where \( p = 12 \) or in the models where \( p = 6 \) (\( \Delta = .022 \)). For other combinations of loadings and intercepts, the changes in correlation are again below .10.

### 3.2 Mixture scores

In the mixture models, two factor scores are investigated. The first factor score is computed after each individual has been classified as part of the honest or faking group based on a posterior probability higher than 50%. Then, that individual is considered to be 100% part of that class and its score is obtained based on the estimated parameters of that class. For the purpose of the discussion, this factor score is referred to as “mixClass”. The other factor score available under the mixture model is a weighted score based on the posterior probability of that observation belonging to each class. This is referred to as “mixWeight”.

#### 3.2.1 Model with \( p = 6 \)

Unlike what was observed in the case of mean scores, in the model of
intermediate size \((p = 6)\) the correlation between true factor scores and mixture factor scores changes substantially depending on sample size. Consider first the conditions with smallest class separation: \(\lambda_{\text{honest}} = .5, \ \psi_{\text{honest}} = .5, \ \text{and} \ \varepsilon_{\text{honest}} = .6.\) As sample size increases from 500 to 1000 and 2000, \(r_{\text{mixClass}}\) increases from .391 to .447 and .520. Alternatively, \(r_{\text{mixWeight}}\) increases from .232 to .302 and .412.

A similar but less pronounced pattern is observed when there is a better distinction between class parameters. Consider now the conditions where \(\lambda_{\text{honest}} = .9, \ \psi_{\text{honest}} = 0, \ \text{and} \ \varepsilon_{\text{honest}} = .3.\) As sample size increases from 500 to 1000 and 2000, \(r_{\text{mixClass}}\) increases from .613 to .697 and .736. Alternatively, \(r_{\text{mixWeight}}\) increases from .587 to .701 and .763.

As seen previously, the variability of the correlation estimates for factor mixture scores decreases as sample size increases. For example, consider \(r_{\text{mixClass}}\) in the conditions where \(\lambda_{\text{honest}} = .5, \ \psi_{\text{honest}} = .5, \ \text{and} \ \varepsilon_{\text{honest}} = .6.\) As sample size increases from 500 to 1000 and 2000, the standard deviation decreases from .204 to .179 and .131, for a range of .073. In the same conditions, the standard deviation of \(r_{\text{mixWeight}}\) is less predictable and changes from .260 to .285 and .284. Note that the standard deviation of \(r_{\text{mixWeight}}\) in this case is not only higher than the equivalent conditions of \(r_{\text{mixClass}}\), but also much higher relative to the correlation coefficient itself. For example, the standard deviation associated with \(r_{\text{mixWeight}} = .520\) is \(SD = .131\) while the standard deviation for \(r_{\text{mixClass}} = .412\) is \(SD = .284.\) This pattern is observed repeatedly in the conditions \(\varepsilon_{\text{honest}} = .6.\)

Now let us turn to the effects of factor loadings. Correlations between true factor scores and \(r_{\text{mixClass}}\) change predictably, and higher loadings for the honest group –
implying higher separation between honest and faking classes – are associated with higher correlations. For example, consider the conditions where \( N = 2000, \ \nu_{honest} = 0, \) and \( \epsilon_{honest} = .3. \) When \( \lambda_{honest} = .5, \ r_{mixClass} = .682, \) and when \( \lambda_{honest} = .9, \ r_{mixClass} = .736. \) This pattern is observed throughout the \( p = 6 \) conditions for \( r_{mixClass}. \) In general, this is also the case for \( r_{mixWeight}, \) with the curious exception of conditions where \( \nu_{honest} = .0, \) and \( \epsilon_{honest} = .6. \) In those cases, performance under \( \lambda_{honest} = .5 \) seems to be more favorable though not by a meaningful margin (an average difference in \( r \) of around .02).

In terms of item intercepts, interactions with the factor loadings make interpretation of the simulation results less straightforward. For example, let us look at conditions where \( N = 500. \) Where \( \lambda_{honest} = .5 \) and \( \epsilon_{honest} = .3, \) increasing the intercept from \( \nu_{honest} = 0 \) to .5 leads to a decrease in \( r_{mixClass} \) from .589 to .580 and a decrease in \( r_{mixWeight} \) from .606 to .544. Where \( \epsilon_{honest} = .6, \) raising the intercept leads to a decrease in \( r_{mixClass} \) from .404 to .391 and a decrease in \( r_{mixWeight} \) from .346 to .232. Now, what happens when factor loadings are higher? Where \( \lambda_{honest} = .9 \) and \( \epsilon_{honest} = .3, \) increasing the intercept from \( \nu_{honest} = 0 \) to .5 leads to an increase in \( r_{mixClass} \) from .613 to .635 and a decrease in \( r_{mixWeight} \) from .587 to .576. Where \( \epsilon_{honest} = .6, \) raising the intercept leads to an increase in \( r_{mixClass} \) from .456 to .524 and an increase in \( r_{mixWeight} \) from .347 to .360. To summarize, in general increasing the intercept decreases the correlation \( r_{mixClass} \) with true factor scores when there is less difference between the class in terms of factor loadings. When the factor loadings of the two classes are further apart, then in general increasing the intercept also increases the correlation \( r_{mixClass} \) with true factor scores. Nevertheless,
there are unusual cases that do not subscribe to this simplification \((N = 1000, \lambda_{\text{honest}} = .5,\)
both \(\varepsilon_{\text{honest}} = .3\) and \(\varepsilon_{\text{honest}} = .6\). With a few exceptions, this is also the case for \(r_{\text{mixWeight}}\).

With item residuals, we see the same pattern as with mean scores: correlations between true factor scores and mixture factor scores (both \(r_{\text{mixClass}}\) and \(r_{\text{mixWeight}}\)) are higher when residuals are lower \((\varepsilon_{\text{honest}} = .3)\). For the lowest class separation \((\lambda_{\text{honest}} = .5,\)
\(\psi_{\text{honest}} = .5)\) and \(N = 2000\), \(r_{\text{mixClass}}\) decreases from .690 to .620 and \(r_{\text{mixWeight}}\) decreases from .714 to .412. Note that these differences are considerably larger than what was found for mean scores, where for \(p = 6\) we saw \(\Delta = .022\).

### 3.2.2 Model with \(p = 12\)

As in the smallest models \((p = 6)\), in the model of intermediate size \((p = 12)\) the correlation between true factor scores and mixture factor scores changes substantially depending on sample size. Consider first the conditions with smallest class separation:

\(\lambda_{\text{honest}} = .5, \psi_{\text{honest}} = .5,\) and \(\varepsilon_{\text{honest}} = .6\). As sample size increases from 500 to 1000 and 2000, \(r_{\text{mixClass}}\) increases from .544 to .601 and .631. Alternatively, \(r_{\text{mixWeight}}\) increases from .381 to .456 and .562. This pattern is also observed when there is a better distinction between class parameters. Consider the conditions where \(\lambda_{\text{honest}} = .9, \psi_{\text{honest}} = 0,\) and \(\varepsilon_{\text{honest}} = .3\). As sample size increases from 500 to 1000 and 2000, \(r_{\text{mixClass}}\) increases from .813 to .831 and .834. Alternatively, \(r_{\text{mixWeight}}\) increases from .826 to .849 and .852.

Once again, the variability of the correlation estimates for factor mixture scores decreases as sample size increases. In the case of \(r_{\text{mixClass}}\) where \(\lambda_{\text{honest}} = .5, \psi_{\text{honest}} = .5,\) and \(\varepsilon_{\text{honest}} = .6,\) as sample size increases from 500 to 1000 and 2000 the standard
deviation decreases from .142 to .079 and .026, for a range of .116. In the same conditions, the standard deviation of $r_{\text{mixWeight}}$ is less predictable and changes from .292 to .296 and .254. As in the $p = 6$ case, we see that the standard deviations of $r_{\text{mixWeight}}$ are not only higher than the equivalent conditions of $r_{\text{mixClass}}$, but also much higher relative to the correlation coefficient itself. For example, in the smallest sample size we see $r_{\text{mixClass}} = .544$ with SD = .142 and $r_{\text{mixWeight}} .381$ with SD = .292. This pattern is more pronounced in the conditions where $\varepsilon_{\text{honest}} = .6$.

Next, the effects of factor loadings. Correlations between true factor scores and $r_{\text{mixClass}}$ again change predictably, and higher loadings for the honest group – implying higher separation between honest and faking classes – are associated with higher correlations. For example, consider the conditions where $N = 2000$, $\nu_{\text{honest}} = 0$, and $\varepsilon_{\text{honest}} = .3$. When $\lambda_{\text{honest}} = .5$, $r_{\text{mixClass}} = .799$, and when $\lambda_{\text{honest}} = .9$, $r_{\text{mixClass}} = .834$. This pattern is observed throughout the $p = 12$ conditions for $r_{\text{mixClass}}$. The relationships above also generally apply for $r_{\text{mixWeight}}$, but there are again two exceptions for conditions where $\nu_{\text{honest}} = .0$, $\varepsilon_{\text{honest}} = .6$, and $N = 500$ or $N = 1000$.

In terms of item intercepts, interpretation is further complicated due to the interaction mentioned previously. In these conditions with more indicators ($p = 12$), the correlation $r_{\text{mixClass}}$ with true factor scores tends to increase when item intercepts are raised regardless of factor loading magnitude - provided that $\varepsilon_{\text{honest}} = .3$. Otherwise, in the cases where $\varepsilon_{\text{honest}} = .6$, $r_{\text{mixClass}}$ tends to decrease when item intercepts increase. On the other hand, the behavior of $r_{\text{mixWeight}}$ is similar to what was seen in the $p = 6$ conditions. In other words, when the factor loadings of the honest class is higher ($\lambda_{\text{honest}}$
= .9), then in general increasing the intercept also increases the correlation \( r_{mixWeight} \) with true factor scores.

Finally, with item residuals, we see again the same pattern described in other conditions: correlations between true factor scores and mixture factor scores (both \( r_{mixClass} \) and \( r_{mixWeight} \)) are higher when residuals are lower (\( \varepsilon_{honest} = .3 \)). For the lowest class separation (\( \lambda_{honest} = .5, \ \psi_{honest} = .5 \)) and \( N = 2000 \), \( r_{mixClass} \) decreases from .801 to .631 (\( \Delta = .17 \)) and \( r_{mixWeight} \) decreases from .822 to .562 (\( \Delta = .26 \)). Note that in the \( p = 12 \) conditions, both \( r_{mixClass} \) and \( r_{mixWeight} \) consistently reach or surpass the .800s in the \( \varepsilon_{honest} = .3 \) cases but tend to perform poorly in the \( \varepsilon_{honest} = .6 \) cases.

3.2.3 Model with \( p = 18 \)

In the models of largest size (\( p = 18 \)), the correlation between true factor scores and mixture factor scores also increases with larger sample sizes. However, this effect of sample size is stronger on \( r_{mixWeight} \) than on \( r_{mixClass} \). To illustrate, consider again the conditions with smallest class separation: \( \lambda_{honest} = .5, \ \psi_{honest} = .5, \) and \( \varepsilon_{honest} = .6 \). As sample size increases from 500 to 1000 and 2000, \( r_{mixClass} \) averages increase from .618 to .654 and .668. Alternatively, \( r_{mixWeight} \) increases from .433 to .548 and .653. In models of this size, the effect of sample size on average correlations can be negligible when there is a better distinction between class parameters. Consider the conditions where \( \lambda_{honest} = .9, \ \psi_{honest} = 0, \) and \( \varepsilon_{honest} = .3 \). As sample size increases from 500 to 1000 and 2000, \( r_{mixClass} \) increases from .878 to .885 and .888. Alternatively, \( r_{mixWeight} \) increases from .889 to .896 and .898.

The variability of the correlation estimates also benefits from sample size.
increases. However, the standard deviations of $r_{mixClass}$ in $p = 18$ conditions are often close to half of their value in $p = 12$ conditions. On the other hand, the standard deviations of $r_{mixWeight}$ are similar to what was observed previously under $p = 12$. In the case of $r_{mixClass}$ where $\lambda_{honest} = .5$, $\psi_{honest} = .5$, and $\varepsilon_{honest} = .6$, as sample size increases from 500 to 1000 and 2000 the standard deviation decreases from .073 to .027 and .019 (compare to .142, .079, and .026 under $p = 12$). In the same conditions, the standard deviation of $r_{mixWeight}$ changes from .294 to .278 and .204 (compare to .292, .296 and .254 under $p = 12$). As in the $p = 6$ and $p = 12$ cases, we see that the standard deviations of $r_{mixWeight}$ are not only higher than the equivalent conditions of $r_{mixClass}$, but also much higher relative to the correlation coefficient itself.

For $p = 18$, we find again that higher loadings for the honest group (i.e., greater class separation) are associated with higher correlations. For example, consider the conditions where $N = 2000$, $\psi_{honest} = 0$, and $\varepsilon_{honest} = .3$. When $\lambda_{honest} = .5$, $r_{mixClass} = .865$, and when $\lambda_{honest} = .9$, $r_{mixClass} = .888$. This pattern is observed throughout the $p = 12$ conditions for $r_{mixClass}$. The relationships above also generally apply for $r_{mixWeight}$, with exceptions again for conditions where $\psi_{honest} = .0$, $\varepsilon_{honest} = .6$, and $N = 500$ or $N = 1000$.

In terms of item intercepts, the correlation $r_{mixClass}$ with true factor scores tends to increase when item intercepts are raised regardless of factor loading magnitude for conditions where $\varepsilon_{honest} = .3$. In conditions where $\varepsilon_{honest} = .6$, $r_{mixClass}$ tends to decrease when item intercepts increase given $\lambda_{honest} = .5$, but increases when $\lambda_{honest} = .9$. The behavior of $r_{mixWeight}$ is similar provided $\varepsilon_{honest} = .3$. In conditions where $\lambda_{honest} = .5$ and $\varepsilon_{honest} = .6$, $r_{mixWeight}$ decreases when intercepts increase; in conditions where $\lambda_{honest} = .9$
and $\varepsilon_{honest} = .6$, $r_{mixWeight}$ is inconsistent across sample sizes.

### 3.2.4 Comparison of mixture scores and multiple group FA

In multiple group factor analysis (MGFA), class membership is known and not estimated. Therefore, factor scores obtained under this approach represent an ideal ceiling for what can be computed under FMM. In Tables 3.1 through 3.3, $r_{mg}$ represents the correlation between honest mean scores and MGFA factor scores across the two groups (honest and fakers). In general, $r_{mg}$ increases with larger $p$, $N$, $\lambda_{honest}$, and $\nu_{honest}$, and decreases with higher $\varepsilon_{honest}$. One exception is that for $p = 6$ and $N = 500$, intercept increases lead to lower $r_{mg}$.

As expected, $r_{mg}$ is generally much higher than both $r_{mixClass}$ and $r_{mixWeight}$. At the extreme, in the smallest class separation example with $p = 6$ in Table 3.1 ($N = 500$, $\lambda_{honest} = .5$, $\nu_{honest} = .5$, and $\varepsilon_{honest} = .6$) we have $r_{mg} = .752$, $r_{mixClass} = .391$, and $r_{mixWeight} = .232$.

In fact, the factor mixture scores are far below even the mean scores that do not account for faking in any way ($r_{mean} = .606$). On the other end of the results, in Table 3.3 ($p = 18$) we see factor mixture scores performing very closely to the ideal. For example, given $N = 2000$, $\lambda_{honest} = .9$, $\nu_{honest} = .5$, and $\varepsilon_{honest} = .3$, we have $r_{mg} = .936$, $r_{mixClass} = .906$, and $r_{mixWeight} = .912$. Factor mixture scores perform both close to the MGFA ideal and above ignoring faking effects ($r_{mean} = .784$).

### 3.3 Classification

The quality of the classification of individuals in the two classes is critical to obtaining the proper factor mixture scores. In this section, we go over a few indices derived from contingency tables with two dimensions – “actual class” and predicted
class”. In each contingency table we then have four cells: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The “positive” outcomes refer to the honest class, such that TP represents the number of observations that were generated under the model correspondent to the honest class which were correctly identified as such by the mixture model. Alternatively, the “negative” outcomes refer to the faking class, such that TN represents the number of observations that were generated under the model correspondent to the faking class which were correctly identified as such by the mixture model. By extension, FP represents observations generated under the faking model that were classified as honest, and FN represents observations generated under the honest model that were classified as faking.

Given these four cells (TP, TN, FP, FN) and the sample size (N), a few indices are of interest. Accuracy is given by (TP + TN)/N and is interpreted as the proportion of correct predictions among the total number of cases. The true positive rate, also known as sensitivity, recall, or hit rate, is given by TP/(TP + FN) and is interpreted as the probability of a positive classification, conditioned on the observation truly being positive. Similarly, the true negative rate (also known as specificity or selectivity) is given by TN/(TN + FP) and represents the probability of a negative classification, conditioned on the observation truly being negative. Finally, the positive predictive value (also known as precision) and the negative predictive value represent the proportions of positive and negative classifications that are true positive and true negative results, respectively. The former is given by TP/(TP + FP) and the latter by TN/(TN + FN).

Classification results are given in Tables 3.4, 3.5, and 3.6. All else held constant,

\[^{6}\text{In the faking literature it often the opposite: TP are the correctly identified faking responses, and TN are the correctly identified honest responses. In this study, this was inverted for computational convenience.}\]
in general increasing both the number of observed variables and sample size improves accuracy, true positive rate, and negative predictive value, with higher averages and lower standard deviations. For example, accuracy hits its lowest point at .695, in the condition where \( p = 6, N = 500, \lambda_{\text{honest}} = .9, \nu_{\text{honest}} = .5, \) and \( \lambda_{\text{honest}} = .6. \) For the same parameters but with \( p = 18 \) and \( N = 2000, \) accuracy =.779 and is above .80 in all other conditions within this model size and sample size. On the other hand, true negative rate and positive predictive value – both conditioned on false positives – on average tend to decrease as sample size increases. In other words, with larger sample sizes the mixture model misidentifies faking observations as honest at a higher rate.

All classification indices behave similarly given changes in model parameters. However, each parameter has unique effects. Factor loading was the only parameter with an exact opposite effect to what was expected: higher loadings for the honest group, implying larger overall class separation, are associated with worse performance for all indices tested (lower averages and higher standard deviations). Item intercepts and residual had the same effects and in the same direction, which was predictable because as they increase, class separation becomes less pronounced.

For the purpose of identifying fake responses, the true negative rate (specificity) might be the most important classification index. Unfortunately, it is also the most sensitive to changes in model parameters. In the \( p = 18 \) and \( N = 2000 \) conditions, specificity hits its minimum at .489 (\( SD = .195 \)), while true positive rate (sensitivity) has a minimum of .949 (\( SD = .042 \)). This is a positive finding in the sense that honest participants have a high probability of being correctly identified but undermines the potential of factor mixture modeling to identify fake responses.
3.4 Summary of findings

To summarize the findings of the simulation study, it is useful to visualize the correlations between scale scores under the most extreme conditions. In Figures 3.1-3.5, each row represents different sample sizes (\( N = 500, 1000, \) and \( 2000 \)), and each column represents different model sizes (\( p = 6, 12, \) and \( 18 \)). Each bar represents the average correlation between true factor scores and scale scores of each method: from left to right, mean scores, factor mixture scores (most likely class), factor mixture scores (weighted by class probabilities), and multiple group factor scores. Mean scores represent the baseline as they completely ignore faking effects – i.e., mixture scores represent an improvement over ignoring faking if their correlation with true scores are higher than mean scores-true scores correlation. On the other hand, multiple group factor scores directly incorporate group membership (faking or honest) and represent what we would ideally recover with factor mixture modeling.

In Figure 3.1, we see that when class separation is at its highest (\( \lambda_{honest} = .9, \)
\( \nu_{honest} = 0, \) and \( \varepsilon_{honest} = .3 \)), factor mixture scores represent an improvement over ignoring faking in all but one condition (small sample size and small sample size).

Alternatively, Figure 3.2 shows that when class separation is at its lowest (\( \lambda_{honest} = .5, \)
\( \nu_{honest} = .5, \) and \( \varepsilon_{honest} = .6 \)), factor mixture scores at best perform as well as ignoring faking effects, and much worse when model size is small. This shows that within the simulation conditions we found cases where factor mixture scores perform well, but alarmingly also cases where it performs much worse than completely ignoring faking.

Given this worst case scenario of lowest class separation, how does manipulating model parameters change the scale score correlations? These effects are shown in Figure
3.3 (increasing loading from $\lambda_{honest} = .5$ to .9), Figure 3.4 (decreasing item intercept from $\psi_{honest} = .5$ to 0), and Figure 3.5 (decreasing item uniqueness from $\epsilon_{honest} = .6$ to .3). The faded areas on top of each bar represent the increase in scale score correlations for each of these parameter changes, all else held constant. Overall, we see that increasing factor loadings improves both mean scores and factor mixture scores to a similar degree. In other words, there is no improvement in the discrepancy between mean scores and factor mixture scores. Decreasing item intercepts does little to change factor mixture scores (most likely class) but improves mean scores and factor mixture scores (weighted by class probabilities) again to a similar degree. Finally, decreasing uniqueness has the most impact on factor mixture scores, but negligible impact on mean scores. In these cases alone, particularly for models of intermediate ($p = 12$) and large size ($p = 18$), mixture modeling represents a very substantial improvement over ignoring faking effects.

In terms of classification, we observed high rates of false positives - incorrectly identifying an observation as “honest”. This in turn leads to small true negative rates (i.e., specificity), the probability of correctly identifying fake responses conditioned on those responses being fake. In the examined conditions, this index can be as low as .377 and is greatly influenced by uniqueness. This is problematic as identifying fake responses is the entire motivation to use factor mixture modeling in this application.
Table 3.1
Simulation results for models of small size ($p = 6$): average and standard deviation (in parentheses) of correlations between true and estimated factor scores

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\lambda$</th>
<th>$\upsilon$</th>
<th>$\Theta$</th>
<th>$r_{\text{mean}}$</th>
<th>$r_{\text{mg}}$</th>
<th>$r_{\text{mixClass}}$</th>
<th>$r_{\text{mixWeight}}$</th>
<th>Rep. Co.</th>
<th>Rep. S.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td></td>
<td>.3</td>
<td>.5</td>
<td>.506 (.034)</td>
<td>.809 (.051)</td>
<td>.589 (.174)</td>
<td>.606 (.200)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.494 (.033)</td>
<td>.755 (.056)</td>
<td>.404 (.175)</td>
<td>.346 (.252)</td>
<td></td>
<td>1000</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.625 (.029)</td>
<td>.807 (.054)</td>
<td>.580 (.197)</td>
<td>.544 (.263)</td>
<td></td>
<td>1000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.606 (.030)</td>
<td>.752 (.059)</td>
<td>.391 (.204)</td>
<td>.232 (.260)</td>
<td></td>
<td>987</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>.9</td>
<td>.3</td>
<td>.674 (.027)</td>
<td>.854 (.054)</td>
<td>.613 (.221)</td>
<td>.587 (.277)</td>
<td></td>
<td>1000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.662 (.028)</td>
<td>.834 (.046)</td>
<td>.456 (.216)</td>
<td>.347 (.301)</td>
<td></td>
<td>993</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>.3</td>
<td>.763 (.022)</td>
<td>.851 (.059)</td>
<td>.635 (.264)</td>
<td>.576 (.316)</td>
<td></td>
<td>1000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.751 (.021)</td>
<td>.835 (.036)</td>
<td>.524 (.242)</td>
<td>.360 (.327)</td>
<td></td>
<td>984</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>.3</td>
<td>.506 (.024)</td>
<td>.825 (.014)</td>
<td>.649 (.112)</td>
<td>.678 (.119)</td>
<td></td>
<td>1000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.495 (.024)</td>
<td>.773 (.016)</td>
<td>.478 (.133)</td>
<td>.446 (.237)</td>
<td></td>
<td>1000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.627 (.020)</td>
<td>.825 (.026)</td>
<td>.658 (.116)</td>
<td>.658 (.182)</td>
<td></td>
<td>1000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.606 (.021)</td>
<td>.771 (.026)</td>
<td>.447 (.179)</td>
<td>.302 (.285)</td>
<td></td>
<td>999</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>.9</td>
<td>.3</td>
<td>.671 (.019)</td>
<td>.870 (.013)</td>
<td>.697 (.134)</td>
<td>.701 (.190)</td>
<td></td>
<td>1000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.662 (.019)</td>
<td>.848 (.027)</td>
<td>.532 (.184)</td>
<td>.430 (.311)</td>
<td></td>
<td>1000</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>.3</td>
<td>.766 (.015)</td>
<td>.870 (.015)</td>
<td>.742 (.166)</td>
<td>.710 (.240)</td>
<td></td>
<td>1000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.751 (.016)</td>
<td>.849 (.018)</td>
<td>.578 (.227)</td>
<td>.410 (.341)</td>
<td></td>
<td>1000</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>.3</td>
<td>.507 (.016)</td>
<td>.831 (.008)</td>
<td>.682 (.025)</td>
<td>.713 (.022)</td>
<td></td>
<td>1000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.495 (.016)</td>
<td>.779 (.010)</td>
<td>.529 (.072)</td>
<td>.542 (.166)</td>
<td></td>
<td>999</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.628 (.014)</td>
<td>.830 (.008)</td>
<td>.690 (.031)</td>
<td>.714 (.077)</td>
<td></td>
<td>1000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.606 (.014)</td>
<td>.779 (.009)</td>
<td>.520 (.131)</td>
<td>.412 (.284)</td>
<td></td>
<td>1000</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>.9</td>
<td>.3</td>
<td>.673 (.014)</td>
<td>.875 (.007)</td>
<td>.736 (.025)</td>
<td>.763 (.060)</td>
<td></td>
<td>1000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.662 (.014)</td>
<td>.854 (.007)</td>
<td>.590 (.140)</td>
<td>.521 (.288)</td>
<td></td>
<td>999</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>.3</td>
<td>.766 (.011)</td>
<td>.875 (.007)</td>
<td>.793 (.040)</td>
<td>.801 (.092)</td>
<td></td>
<td>1000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.751 (.011)</td>
<td>.854 (.007)</td>
<td>.647 (.185)</td>
<td>.516 (.336)</td>
<td></td>
<td>1000</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Notes. $N$ = sample size, $\lambda$ = factor loading, $\upsilon$ = item intercept, $\Theta$ = error variance (uniqueness), Rep. Co. = number of completed replications out of 1000, Rep. S.C. = replications where all observations were assigned to a single class.
Table 3.2
Simulation results for models of intermediate size ($p = 12$): average and standard deviation (in parentheses) of correlations between true and estimated factor scores

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\lambda$</th>
<th>$\upsilon$</th>
<th>$\Theta$</th>
<th>$r_{\text{mean}}$</th>
<th>$r_{\text{mg}}$</th>
<th>$r_{\text{mixClass}}$</th>
<th>$r_{\text{mixWeight}}$</th>
<th>Rep. Co.</th>
<th>Rep. S.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.520 (.032)</td>
<td>.809 (.147)</td>
<td>.780 (.055)</td>
<td>.800 (.053)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.6</td>
<td>.513 (.032)</td>
<td>.852 (.016)</td>
<td>.567 (.111)</td>
<td>.523 (.245)</td>
<td>996</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>.3</td>
<td>.653 (.027)</td>
<td>.883 (.013)</td>
<td>.782 (.042)</td>
<td>.790 (.097)</td>
<td></td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.6</td>
<td>.642 (.026)</td>
<td>.853 (.016)</td>
<td>.544 (.142)</td>
<td>.381 (.292)</td>
<td>971</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>.3</td>
<td>.683 (.027)</td>
<td>.899 (.057)</td>
<td>.813 (.070)</td>
<td>.826 (.093)</td>
<td></td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.6</td>
<td>.678 (.027)</td>
<td>.897 (.013)</td>
<td>.598 (.167)</td>
<td>.429 (.318)</td>
<td>991</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>.3</td>
<td>.783 (.019)</td>
<td>.908 (.012)</td>
<td>.848 (.055)</td>
<td>.848 (.101)</td>
<td></td>
<td></td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.6</td>
<td>.776 (.020)</td>
<td>.897 (.015)</td>
<td>.681 (.179)</td>
<td>.480 (.393)</td>
<td>967</td>
<td>3</td>
</tr>
<tr>
<td>1000</td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.520 (.023)</td>
<td>.797 (.159)</td>
<td>.795 (.023)</td>
<td>.815 (.018)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.6</td>
<td>.513 (.024)</td>
<td>.859 (.009)</td>
<td>.615 (.056)</td>
<td>.632 (.154)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.6</td>
<td>.642 (.019)</td>
<td>.859 (.009)</td>
<td>.601 (.079)</td>
<td>.456 (.296)</td>
<td>999</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>.3</td>
<td>.684 (.019)</td>
<td>.906 (.039)</td>
<td>.831 (.018)</td>
<td>.849 (.015)</td>
<td></td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.6</td>
<td>.679 (.019)</td>
<td>.902 (.008)</td>
<td>.662 (.085)</td>
<td>.554 (.295)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>.3</td>
<td>.784 (.015)</td>
<td>.913 (.008)</td>
<td>.863 (.014)</td>
<td>.873 (.028)</td>
<td></td>
<td></td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.6</td>
<td>.775 (.014)</td>
<td>.902 (.007)</td>
<td>.728 (.129)</td>
<td>.557 (.332)</td>
<td>1000</td>
<td>4</td>
</tr>
<tr>
<td>2000</td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.520 (.016)</td>
<td>.829 (.138)</td>
<td>.799 (.016)</td>
<td>.819 (.012)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.6</td>
<td>.514 (.016)</td>
<td>.861 (.006)</td>
<td>.636 (.033)</td>
<td>.684 (.026)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.6</td>
<td>.643 (.013)</td>
<td>.863 (.006)</td>
<td>.631 (.026)</td>
<td>.562 (.254)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>.3</td>
<td>.684 (.013)</td>
<td>.916 (.005)</td>
<td>.834 (.012)</td>
<td>.852 (.010)</td>
<td></td>
<td>1000</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.6</td>
<td>.679 (.013)</td>
<td>.905 (.005)</td>
<td>.687 (.017)</td>
<td>.689 (.181)</td>
<td>1000</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>.3</td>
<td>.783 (.010)</td>
<td>.916 (.005)</td>
<td>.866 (.009)</td>
<td>.878 (.008)</td>
<td></td>
<td></td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.6</td>
<td>.775 (.010)</td>
<td>.905 (.005)</td>
<td>.770 (.038)</td>
<td>.697 (.250)</td>
<td>1000</td>
<td>0</td>
</tr>
</tbody>
</table>

Notes: $N =$ sample size, $\lambda =$ factor loading, $\upsilon =$ item intercept, $\Theta =$ error variance (uniqueness), Rep. Co. = number of completed replications out of 1000, Rep. S.C. = replications where all observations were assigned to a single class.
Table 3.3
Simulation results for models of large size \((p = 18)\): average and standard deviation (in parentheses) of correlations between true and estimated factor scores

<table>
<thead>
<tr>
<th>N</th>
<th>(\lambda)</th>
<th>(\upsilon)</th>
<th>(\Theta)</th>
<th>(r_{\text{mean}})</th>
<th>(r_{\text{mg}})</th>
<th>(r_{\text{mixClass}})</th>
<th>(r_{\text{mixWeight}})</th>
<th>Rep. Co.</th>
<th>Rep. S.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.522 (.031)</td>
<td>.915 (.010)</td>
<td>.853 (.026)</td>
<td>.865 (.021)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.6 .520 (.030)</td>
<td>.893 (.010)</td>
<td>.642 (.069)</td>
<td>.632 (.193)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.661 (.025)</td>
<td>.914 (.009)</td>
<td>.854 (.021)</td>
<td>.865 (.018)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.6 .655 (.027)</td>
<td>.892 (.010)</td>
<td>.618 (.073)</td>
<td>.433 (.294)</td>
<td>984</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>.9</td>
<td>0</td>
<td>.3</td>
<td>.687 (.026)</td>
<td>.902 (.104)</td>
<td>.878 (.020)</td>
<td>.889 (.017)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.6 .683 (.026)</td>
<td>.922 (.018)</td>
<td>.681 (.071)</td>
<td>.511 (.310)</td>
<td>988</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.790 (.020)</td>
<td>.930 (.008)</td>
<td>.897 (.015)</td>
<td>.904 (.015)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.6 .784 (.019)</td>
<td>.923 (.009)</td>
<td>.755 (.085)</td>
<td>.533 (.331)</td>
<td>976</td>
<td>2</td>
</tr>
<tr>
<td>2000</td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.525 (.017)</td>
<td>.920 (.004)</td>
<td>.865 (.013)</td>
<td>.877 (.010)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.6 .521 (.017)</td>
<td>.898 (.005)</td>
<td>.694 (.027)</td>
<td>.737 (.021)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.664 (.013)</td>
<td>.920 (.004)</td>
<td>.864 (.010)</td>
<td>.876 (.008)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.6 .655 (.014)</td>
<td>.898 (.005)</td>
<td>.668 (.019)</td>
<td>.653 (.204)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>.9</td>
<td>0</td>
<td>.3</td>
<td>.688 (.013)</td>
<td>.930 (.047)</td>
<td>.888 (.009)</td>
<td>.898 (.008)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.6 .685 (.013)</td>
<td>.929 (.004)</td>
<td>.723 (.018)</td>
<td>.748 (.135)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.790 (.010)</td>
<td>.936 (.004)</td>
<td>.906 (.007)</td>
<td>.912 (.006)</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.6 .784 (.010)</td>
<td>.928 (.004)</td>
<td>.797 (.013)</td>
<td>.779 (.182)</td>
<td>1000</td>
<td>0</td>
</tr>
</tbody>
</table>

Notes. \(N\) = sample size, \(\lambda\) = factor loading, \(\upsilon\) = item intercept, \(\Theta\) = error variance (uniqueness), Rep. Co. = number of completed replications out of 1000, Rep. S.C. = replications where all observations were assigned to a single class.
Table 3.4
Simulation results for models of small size \((p = 6)\): average and standard deviation (in parentheses) of classification indices

<table>
<thead>
<tr>
<th>(N)</th>
<th>(\lambda)</th>
<th>(\nu)</th>
<th>(\Theta)</th>
<th>Acc</th>
<th>TPR</th>
<th>TNR</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.940 (.024), .983 (.014), .840 (.097), .937 (.034), .958 (.031)</td>
<td>.885 (.048), .954 (.066), .721 (.189), .896 (.061), .902 (.090)</td>
<td>.882 (.031), .968 (.033), .681 (.150), .880 (.047), .918 (.065)</td>
<td>.746 (.108), .820 (.236), .575 (.274), .844 (.083), .709 (.176)</td>
<td>.878 (.037), .964 (.036), .676 (.177), .880 (.056), .911 (.070)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.847 (.031), .958 (.049), .587 (.173), .849 (.051), .892 (.089)</td>
<td>.695 (.084), .775 (.239), .509 (.334), .824 (.100), .605 (.159)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.9</td>
<td>0</td>
<td>.3</td>
<td>.946 (.017), .985 (.009), .857 (.068), .942 (.024), .961 (.021)</td>
<td>.890 (.043), .968 (.033), .708 (.190), .891 (.060), .922 (.060)</td>
<td>.885 (.033), .974 (.022), .679 (.146), .880 (.045), .928 (.048)</td>
<td>.780 (.089), .871 (.189), .569 (.248), .844 (.075), .765 (.153)</td>
<td>.800 (.046), .921 (.092), .538 (.297), .839 (.089), .791 (.106)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.853 (.031), .968 (.029), .586 (.152), .848 (.044), .907 (.064)</td>
<td>.719 (.068), .836 (.209), .444 (.331), .809 (.097), .684 (.162)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.951 (.007), .985 (.006), .870 (.033), .947 (.013), .962 (.014)</td>
<td>.898 (.039), .973 (.021), .724 (.165), .896 (.052), .929 (.042)</td>
<td>.893 (.022), .976 (.015), .700 (.102), .885 (.032), .931 (.035)</td>
<td>.806 (.046), .916 (.098), .548 (.244), .837 (.069), .805 (.116)</td>
<td>.807 (.049), .926 (.068), .529 (.297), .836 (.086), .813 (.094)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.891 (.022), .973 (.017), .701 (.105), .885 (.033), .923 (.037)</td>
<td>.738 (.047), .893 (.153), .377 (.298), .789 (.083), .733 (.138)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1800</td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.860 (.024), .973 (.020), .598 (.115), .851 (.033), .913 (.047)</td>
<td>.738 (.047), .893 (.153), .377 (.298), .789 (.083), .733 (.138)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: in the descriptions below, \(n\) = sample size, TP = true positives, TN = true negatives, FP = false positives, FN = false negatives.

Table abbreviations:
- Acc: accuracy, \((TP + TN)/n\).
- TPR: true positive rate (aka sensitivity, recall, or hit rate). \(TP/(TP + FN)\).
- TNR: true negative rate (aka specificity, selectivity). \(TN/(TN + FP)\).
- PPV: positive predictive value (aka precision). \(TP/(TP + FP)\).
- NPV: negative predictive value. \(TN/(TN + FN)\).
Table 3.5
Simulation results for models of intermediate size ($p = 12$): average and standard deviation (in parentheses) of classification indices

<table>
<thead>
<tr>
<th>N</th>
<th>$\lambda$</th>
<th>$\upsilon$</th>
<th>$\Theta$</th>
<th>Acc</th>
<th>TPR</th>
<th>TNR</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.975 (.008)</td>
<td>.991 (.007)</td>
<td>.938 (.032)</td>
<td>.974 (.013)</td>
<td>.979 (.015)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td></td>
<td></td>
<td>.918 (.036)</td>
<td>.976 (.023)</td>
<td>.783 (.151)</td>
<td>.917 (.051)</td>
<td>.941 (.045)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td></td>
<td>.3</td>
<td>.942 (.020)</td>
<td>.983 (.013)</td>
<td>.848 (.080)</td>
<td>.939 (.028)</td>
<td>.957 (.030)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td></td>
<td></td>
<td>.812 (.064)</td>
<td>.896 (.134)</td>
<td>.614 (.224)</td>
<td>.857 (.069)</td>
<td>.788 (.128)</td>
</tr>
<tr>
<td></td>
<td>.9</td>
<td>0</td>
<td>.3</td>
<td>.943 (.017)</td>
<td>.983 (.012)</td>
<td>.851 (.070)</td>
<td>.940 (.025)</td>
<td>.957 (.027)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td></td>
<td></td>
<td>.822 (.042)</td>
<td>.926 (.074)</td>
<td>.579 (.247)</td>
<td>.849 (.076)</td>
<td>.815 (.091)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td></td>
<td>.3</td>
<td>.926 (.020)</td>
<td>.980 (.015)</td>
<td>.799 (.085)</td>
<td>.920 (.029)</td>
<td>.948 (.035)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td></td>
<td></td>
<td>.754 (.054)</td>
<td>.873 (.151)</td>
<td>.477 (.261)</td>
<td>.813 (.076)</td>
<td>.705 (.133)</td>
</tr>
<tr>
<td>1000</td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.977 (.006)</td>
<td>.992 (.004)</td>
<td>.942 (.020)</td>
<td>.976 (.008)</td>
<td>.980 (.010)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td></td>
<td></td>
<td>.931 (.030)</td>
<td>.976 (.156)</td>
<td>.823 (.123)</td>
<td>.931 (.041)</td>
<td>.942 (.032)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td></td>
<td>.3</td>
<td>.947 (.010)</td>
<td>.985 (.008)</td>
<td>.859 (.044)</td>
<td>.943 (.017)</td>
<td>.962 (.019)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td></td>
<td></td>
<td>.831 (.047)</td>
<td>.933 (.083)</td>
<td>.592 (.218)</td>
<td>.852 (.065)</td>
<td>.841 (.099)</td>
</tr>
<tr>
<td></td>
<td>.9</td>
<td>0</td>
<td>.3</td>
<td>.947 (.010)</td>
<td>.984 (.008)</td>
<td>.861 (.044)</td>
<td>.943 (.017)</td>
<td>.960 (.019)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td></td>
<td></td>
<td>.831 (.046)</td>
<td>.941 (.053)</td>
<td>.574 (.251)</td>
<td>.848 (.075)</td>
<td>.847 (.081)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td></td>
<td>.3</td>
<td>.931 (.013)</td>
<td>.982 (.010)</td>
<td>.812 (.057)</td>
<td>.925 (.020)</td>
<td>.952 (.023)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td></td>
<td></td>
<td>.779 (.035)</td>
<td>.910 (.102)</td>
<td>.473 (.241)</td>
<td>.813 (.067)</td>
<td>.765 (.118)</td>
</tr>
<tr>
<td>2000</td>
<td>.5</td>
<td>0</td>
<td>.3</td>
<td>.977 (.004)</td>
<td>.992 (.003)</td>
<td>.942 (.015)</td>
<td>.976 (.006)</td>
<td>.981 (.007)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td></td>
<td></td>
<td>.940 (.015)</td>
<td>.978 (.010)</td>
<td>.849 (.067)</td>
<td>.939 (.024)</td>
<td>.945 (.022)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td></td>
<td>.3</td>
<td>.949 (.007)</td>
<td>.986 (.006)</td>
<td>.861 (.029)</td>
<td>.943 (.011)</td>
<td>.964 (.013)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td></td>
<td></td>
<td>.838 (.043)</td>
<td>.953 (.042)</td>
<td>.569 (.220)</td>
<td>.845 (.063)</td>
<td>.870 (.074)</td>
</tr>
<tr>
<td></td>
<td>.9</td>
<td>0</td>
<td>.3</td>
<td>.949 (.007)</td>
<td>.986 (.006)</td>
<td>.863 (.029)</td>
<td>.944 (.011)</td>
<td>.963 (.013)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td></td>
<td></td>
<td>.840 (.048)</td>
<td>.946 (.041)</td>
<td>.592 (.240)</td>
<td>.853 (.070)</td>
<td>.856 (.069)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td></td>
<td>.3</td>
<td>.932 (.008)</td>
<td>.984 (.007)</td>
<td>.813 (.037)</td>
<td>.925 (.013)</td>
<td>.956 (.015)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td></td>
<td></td>
<td>.782 (.034)</td>
<td>.943 (.060)</td>
<td>.406 (.229)</td>
<td>.795 (.059)</td>
<td>.816 (.098)</td>
</tr>
</tbody>
</table>

Note: in the descriptions below, $n$ = sample size, TP = true positives, TN = true negatives, FP = false positives, FN = false negatives.

Table abbreviations:
- Acc: accuracy. ($TP + TN)/n$.
- TPR: true positive rate (aka sensitivity, recall, or hit hate). $TP/(TP + FN)$.
- TNR: true negative rate (aka specificity, selectivity). $TN/(TN + FP)$.
- PPV: positive predictive value (aka precision). $TP/(TP + FP)$.
- NPV: negative predictive value. $TN/(TN + FN)$. 
Table 3.6
Simulation results for models of large size ($p = 18$): average and standard deviation (in parentheses) of classification indices

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\lambda$</th>
<th>$\nu$</th>
<th>$\Theta$</th>
<th>Acc</th>
<th>TPR</th>
<th>TNR</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>.5</td>
<td>.3</td>
<td></td>
<td>.986 (.006)</td>
<td>.995 (.005)</td>
<td>.967 (.019)</td>
<td>.986 (.008)</td>
<td>.988 (.011)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.936 (.030)</td>
<td>.976 (.019)</td>
<td>.843 (.123)</td>
<td>.938 (.043)</td>
<td>.942 (.039)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.3</td>
<td></td>
<td>.970 (.009)</td>
<td>.990 (.007)</td>
<td>.923 (.033)</td>
<td>.968 (.013)</td>
<td>.976 (.017)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.833 (.051)</td>
<td>.924 (.099)</td>
<td>.621 (.213)</td>
<td>.861 (.066)</td>
<td>.830 (.110)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.3</td>
<td></td>
<td>.970 (.009)</td>
<td>.990 (.007)</td>
<td>.924 (.032)</td>
<td>.968 (.013)</td>
<td>.975 (.017)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.833 (.045)</td>
<td>.949 (.048)</td>
<td>.564 (.235)</td>
<td>.844 (.071)</td>
<td>.859 (.078)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.3</td>
<td></td>
<td>.960 (.011)</td>
<td>.988 (.009)</td>
<td>.897 (.040)</td>
<td>.957 (.016)</td>
<td>.970 (.021)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.779 (.041)</td>
<td>.904 (.113)</td>
<td>.487 (.244)</td>
<td>.817 (.070)</td>
<td>.762 (.123)</td>
</tr>
<tr>
<td>1000</td>
<td>.5</td>
<td>.3</td>
<td></td>
<td>.987 (.004)</td>
<td>.995 (.003)</td>
<td>.969 (.013)</td>
<td>.987 (.005)</td>
<td>.989 (.007)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.947 (.020)</td>
<td>.980 (.011)</td>
<td>.869 (.082)</td>
<td>.947 (.029)</td>
<td>.951 (.024)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.3</td>
<td></td>
<td>.971 (.006)</td>
<td>.991 (.005)</td>
<td>.925 (.023)</td>
<td>.969 (.009)</td>
<td>.978 (.011)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.853 (.039)</td>
<td>.950 (.045)</td>
<td>.624 (.203)</td>
<td>.862 (.061)</td>
<td>.872 (.076)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.3</td>
<td></td>
<td>.972 (.006)</td>
<td>.991 (.005)</td>
<td>.927 (.022)</td>
<td>.970 (.009)</td>
<td>.978 (.011)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.851 (.044)</td>
<td>.950 (.041)</td>
<td>.622 (.222)</td>
<td>.862 (.067)</td>
<td>.869 (.070)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.3</td>
<td></td>
<td>.963 (.007)</td>
<td>.989 (.005)</td>
<td>.901 (.027)</td>
<td>.959 (.010)</td>
<td>.974 (.013)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.800 (.036)</td>
<td>.939 (.065)</td>
<td>.475 (.225)</td>
<td>.815 (.062)</td>
<td>.823 (.097)</td>
</tr>
<tr>
<td>2000</td>
<td>.5</td>
<td>.3</td>
<td></td>
<td>.988 (.003)</td>
<td>.996 (.002)</td>
<td>.970 (.009)</td>
<td>.987 (.004)</td>
<td>.990 (.005)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.954 (.008)</td>
<td>.981 (.007)</td>
<td>.890 (.038)</td>
<td>.954 (.014)</td>
<td>.954 (.016)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.3</td>
<td></td>
<td>.972 (.004)</td>
<td>.992 (.003)</td>
<td>.926 (.016)</td>
<td>.969 (.006)</td>
<td>.980 (.008)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.859 (.043)</td>
<td>.960 (.031)</td>
<td>.624 (.200)</td>
<td>.862 (.059)</td>
<td>.889 (.060)</td>
</tr>
<tr>
<td></td>
<td>.6</td>
<td>.3</td>
<td></td>
<td>.973 (.004)</td>
<td>.992 (.003)</td>
<td>.928 (.015)</td>
<td>.970 (.006)</td>
<td>.980 (.007)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.865 (.038)</td>
<td>.955 (.029)</td>
<td>.655 (.181)</td>
<td>.871 (.055)</td>
<td>.877 (.053)</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.3</td>
<td></td>
<td>.964 (.005)</td>
<td>.990 (.004)</td>
<td>.901 (.019)</td>
<td>.959 (.008)</td>
<td>.976 (.009)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.811 (.035)</td>
<td>.949 (.042)</td>
<td>.489 (.195)</td>
<td>.818 (.052)</td>
<td>.841 (.078)</td>
</tr>
</tbody>
</table>

Note: in the descriptions below, $n = \text{sample size}$, TP = true positives, TN = true negatives, FP = false positives, FN = false negatives.

Table abbreviations:
- Acc: accuracy, $(TP + TN)/n$.
- TPR: true positive rate (aka sensitivity, recall, or hit rate), $TP/(TP + FN)$.
- TNR: true negative rate (aka specificity, selectivity), $TN/(TN + FP)$.
- PPV: positive predictive value (aka precision), $TP/(TP + FP)$.
- NPV: negative predictive value, $TN/(TN + FN)$. 
Figure 3.1 Simulation results for the conditions with highest class separation.
Figure 3.2 Simulation results for the conditions with lowest class separation.
Figure 3.3 Simulation results: lowest class separation with increased factor loadings.
Figure 3.4 Simulation results: lowest class separation with decreased item intercepts.
Figure 3.5 Simulation results: lowest class separation with decreased uniquenesses.
CHAPTER 4: EMPIRICAL APPLICATION

To investigate the feasibility of applying factor mixture modeling to faking data, I used data collected by Wetzel, Frick, and Brown (2021). That study aimed to compare the validity of multidimensional forced choice (MFC) and rating scale (RS) versions of a newly developed Big Five instrument, the Big Five Triplets (BFT; Wetzel & Frick, 2020). The BFT is available in German and English from https://osf.io/ft9ud/. Each participant filled out the BFT under honest instructions and again under fake-good instructions.

4.1 Sample

The data for Wetzel et al.’s study (2021) came from two subsamples. The first is a laboratory sample of 1,042 respondents from two German universities. The second subsample is an Internet access panel sample that consisted of 1,217 participants. Only respondents whose first language was German and who were between 18 and 30 years old were selected. Following the same exclusion criteria and quality checks outlined in Wetzel and Frick (2020), 132 participants were excluded from the laboratory sample and 260 participants were excluded from the Internet sample. In total, 622 participants responded to the RS version of the BFT, and that is the data used here. Respondents had a mean age of 23.64 (SD = 4.39). Out of the 622 that completed the RS format, 64.7% were female, 34.9% were male, and .3% were transgender.
4.2 Measure

The BFT consists of seven agreeableness items, 10 openness items, 13 extroversion items, 14 conscientiousness items, and 16 neuroticism items, for a total of 60 items. Participants were instructed to answer each item on a four-point rating scale with categories *strongly disagree*, *disagree*, *agree*, and *strongly agree*. Items were presented in triplets which in turn were matched based on social desirability ratings. A sample of 33 undergraduate psychology students performed the rating prior to test construction following the instruction “Socially desirable means that the trait or behavior described in the statement fulfills societal norms and expectations” (Wetzel et al., 2021). Under honest instructions, empirical reliability ranged from .79 (agreeableness) to .91 (neuroticism and extroversion).

4.3 Faking instructions

After responding to the BFT under honest instructions, participants were asked to complete the same measure in a hypothetical but realistic scenario. Respondents were instructed to imagine that they were applying to a Master’s program in Psychology in Germany (a common goal for undergraduate students in that country, which composed 38% of the laboratory sample). The faking instructions are copied below (Wetzel et al., 2021):

> Universities have to select students into their programs based on certain criteria. Besides cognitive abilities, personality profiles also play a major role in the selection of candidates. Personality instruments such as the following are administered to candidates during the selection process. Please imagine that your goal is to be admitted as a student to the Master’s
program of the Department of Psychology at the University of Konstanz. The Department of Psychology is looking for students who are conscientious, reliable, extraverted, and emotionally stable. An ideal candidate would, for example, be punctual and assertive, and would complete assignments on time. He or she should be interested in people and be characterized by curiosity and creativity. Furthermore, the Department of Psychology is looking for students who are optimistic and purposefully pursue their academic goals. At the same time, they should be balanced, gregarious and helpful, and show compassion with others.

Please fill out the following questionnaire in a way that fulfills these criteria to increase your chances of being admitted to the Master’s program. (p. 160)

It is of note that the participants were instructed specifically on what direction to fake their responses for each dimension, aiming to decrease their scores in neuroticism and increase their scores in extroversion, openness, agreeableness, and conscientiousness.

4.4 Model fitting: honest

Items for each Big Five factor must have reasonably good fit to unidimensional models, as this is the level that will be focus of the factor mixture modeling. Confirmatory factor analyses were conducted using the software R (R Core Team, 2019) and the package lavaan (Rosseel, 2012) following a series of steps for each Big Five dimension. In the first step, a CFA model is fit individually to a Big Five dimension and its associated items in the BFT. The second step was to eliminate one at a time any items
with small standardized loadings (< .3). A new model is fit after each item removal. If all loadings are adequate, we proceed to the third step where we investigate the matrix of standardized residuals. If any items had residual correlation above an absolute value of .10, they were more closely investigated. To consider items for removal, I also computed the average of the residual correlations, with the highest values indicating items with worst fit.

Throughout all analyses, data was treated as continuous, maximum likelihood with robust standard errors was used, and model fit was verified after each modification. The scale for all models was set using unit variance identification (UVI), meaning that the factor variance was set to one (\( \psi = 1 \)). As recommended by Kline (2015), I used both the “chi-square test” of exact fit and goodness-of-fit indices to interpret the results of the analyzes. Specifically, to test exact fit I use the mean-adjusted likelihood ratio test statistic described in by Asparouhov and Muthén (2005) and commonly referenced as \( X^2_{MLR} \). The goodness-of-fit indices used were the root mean square error of approximation (RMSEA; Steiger, 1990) with its 90% confidence interval, the standardized root mean square residual (SRMR; Bentler, 1990), the comparative fit index (CFI; Bentler, 1990), and the Tucker-Lewis Index (TLI; Tucker & Lewis, 1973).

The interpretation of these indices is as follows. The model chi-square tests the null hypothesis that the model fit exactly. If the model is rejected, the size of model misfit is evaluated. The SRMR and the RMSEA evaluate the absolute size of misfit, whereas the CFI and TLI assess the size of model misfit relative to an independence model. The RMSEA and TLI assess the size of the misfit per degree of freedom, introducing a penalty to prevent overfitting, whereas the SRMR and CFI do not. An RMSEA between 0
and .05 indicates a close fit to the data generating process, values between .05 and .08 indicate an acceptable fit, and greater than or equal to .10 suggest a poor fit (Browne & Cudeck, 1993). Values of the SRMR of less than .08 indicate a close fit, whereas values between .08 and .10 indicate acceptable fit (Hu & Bentler, 1999). The CFI values of greater than roughly .90 indicate an adequate fit (Hu & Bentler, 1999) and values of greater than or equal to .95 suggest an excellent fit. Values of TLI ≥ .95 indicate good fit (Hu & Bentler, 1999).

Model fit indices of the final models are provided in Table 4.1. The Neuroticism and Extroversion subscales were not modified. The modifications to the Openness, Agreeableness, and Conscientiousness subscales are described below. Overall, although exact fit was not attained after removing items from these three scales, fit statistically improved ($p < .05$).

In the Openness subscale, the fifth item (number 53 in the BFT), *I am very sensitive*, was removed due to a nonsignificant ($p > .10$) and small standardized loading (< .3). Chi-square difference testing for the scaled test statistic (Satorra & Bentler, 2010) showed improved fit for the new model, $\chi^2(8) = 45.024, p < .001$.

In the Agreeableness subscale, the third item (number 16 overall), *I don’t allow others to use me*, was removed due to a nonsignificant ($p > .10$) and small standardized loading (< .3). After this modification, some large residual correlations were observed ($r > .20$). The items in question are *I am not interested in other people’s problems* (second item, number 4 overall), *I pay attention to the needs of others* (sixth item, number 55 overall), and *I sympathize with others’ feelings* (seventh item, number 59 overall). It is plausible that the multiple occurrences of the word “other(s)” is the cause of the issue. To
avoid the removal of more items from an already short subscale, these residual correlations were incorporated in the model. Chi-square difference testing showed improvement of the third model over the second model, $\chi^2(3) = 496.264, p < .001$. See Table 4.1 for further details.

In the Conscientiousness subscale, the eleventh item (number 51 overall), I get chores done right away, was removed due to large residual correlations with other items. The second model had another item, I like when everything has its place (eighth, 41 overall), removed for the same reason. Chi-square difference testing showed improvement of the third model over the second model, $\chi^2(11) = 133.456, p < .001$. See Table 4.1

4.5 Model fitting: faking

Using Pavlov et al.’s data as a reference, we can draw lines that are representative of the average response of honest and faking participants. The slope and intercept of those lines can then be used as starting values for the factor mixture model. Those values will change depending on the construct, as respondents can aim to increase their assessment scores (as in the case of Extroversion, Openness, Agreeableness, and Conscientiousness in this application), or they may aim to decrease their assessment scores (as in the case of Neuroticism). For example, in Figure 4.1 we have a 1-4 scale (as in Wetzel et al.’s data) mapped in both X and Y axes. Ideal honest respondents will fall close to the grey diagonal line going through points (1, 1) and (4, 4). The mean assessment scores of participants “faking high” will vary around the dashed line going through the points (1, 3.5) and (4, 4). Alternatively, the mean assessment scores of participants “faking low” will vary around the dashed line going through the points (1, 1).
and (4, 1.5). Note that this is an approximation, and the .5 point difference from the maximum and minimum values is used to accommodate for variability both above and below the lines as most respondents do not fake maximally or minimally.

Given two \((x, y)\) points, for each linear function we can determine the intercept and slope. The slope is given by \(b = \Delta y / \Delta x\) and the intercept by solving \(y = a + bx\) with any of the two points. In this application, it is more interesting to use a [-3, 3] range for the y-axis because the latent trait is standardized. Then, the honest group has slope

\[
b = \frac{(4 - 1)}{(3 - (-3))} = .5
\]

and intercept \(a = 4 - (.5 \times 3) = 2.5\). The faking group in the “faking high” dimensions will have slope

\[
b = \frac{(4 - 3.5)}{(3 - (-3))} = .083
\]

and intercept \(a = 4 - (.083 \times 3) = 3.751\). The faking group in the “faking low” dimensions will have the same slope (.083) and intercept \(a = 1.5 - (.083 \times 3) = .751\).

I addition to creating plausible starting values, I define the criterion to determine what degree of deviation in the assessment scores (from the honest scores) is taken to be “faking”. This is recorded as a grouping variable and used as reference value when judging the quality of the classification. The grouping variable was also recoded as dummy variable to be used as training data (see section 4.5.2 for details). In the exploratory analysis of Pavlov et al.’s data, this was done through visual analysis and an arbitrary difference of .5 in mean scores was chosen. Here, I choose instead to take into account the reliability (or lack thereof) of the scores. First, I calculated the standard error
of measurement (SEm) = \( SD \times \sqrt{1 - r_{xy}} \) where \( SD \) refers to the standard deviation of mean test scores and is the reliability of scores. Cronbach’s alpha was used in lieu of \( r_{xy} \).

Then, around the ideal diagonal I used the SEm to build a 95% confidence interval within which are the “honest” observations. In some dimensions, some participants had faking scores outside of this interval but in the opposite direction of what is expected. For example, in the case of extroversion the participants were instructed to inflate their scores (“fake good”) in the second assessment; yet, their scores were much lower than their original responses. In those cases, those data points were also classified as honest responses. The group-coded observations are shown in Figures 4.2-4.6 for each Big Five dimension.

4.5.1 FMM: starting values

Two FMMs were fitted to each Big Five dimension, with starting values for factor loadings and item intercepts chosen using the procedure described in the previous section. The difference between each FMM is that in the first model item residual variances were constrained to be equal across classes, and in the second model they were freely estimated. Information criteria and entropy values are provided in Table 4.2, namely: the Akaike information criterion (AIC; Akaike, 1974, 1987), the Consistent AIC (CAIC; Bozdogan, 1987), the Bayesian information criterion (BIC; Schwarz, 1978), and the sample size adjusted BIC (saBIC; Sclove, 1987). Models that have the lowest information criteria in terms of absolute value are selected. Entropy is a statistic that ranges from 0.00 to 1.00 (in Mplus) and indicates the degree of confidence with which class assignment was performed. Entropy values > .80 indicate adequate separation between the latent classes.
Next, I compare the correlation of faking mean scores and factor mixture scores (of the most likely class) with honest mean scores. I also report specificity, the probability that the test can identify faked responses conditioned on the individual truly having faked, and sensitivity, the probability that the test can identify honest responses conditioned on the individual truly responding honestly. As 13 individuals did not complete the faking component of this study, \( N = 609 \). The number of random starts was 5000.

Neuroticism is the only “faking low” dimension in this dataset. Mean honest scores correlated \( r = .270 \) \((p < .001)\) with mean faking scores. In the constrained model, mean honest scores and factor mixture scores correlated \( r = .069 \) \((p < .10)\). For the constrained model, specificity = .761 and sensitivity = .799. In the model with unconstrained residual variances, mean honest scores correlated \( r = .160 \) \((p < .001)\) with factor mixture scores. In the unconstrained model, specificity = .971 and sensitivity = .114. Information criteria suggests the selection of the first model. Figure 4.7 shows classification results for Neuroticism (compare to Figure 4.2).

Extroversion is a “faking high” construct in this application. Mean honest scores correlated \( r = .318 \) \((p < .001)\) with mean faking scores. In the constrained model, mean honest scores and factor mixture scores correlated \( r = -.236 \) \((p < .001)\), with specificity = .610 and sensitivity = .744. Given the negative correlation coefficient, I experimented with changing the signs of model coefficients, but results were identical in magnitude. In the model with unconstrained residual variances, mean honest scores correlated \( r = .247 \) \((p < .001)\) with factor mixture scores, with specificity = .864 and sensitivity = .492. Information criteria is slightly higher in the first model, but entropy is much lower in the
second model (.696 versus .987 in the first model), so Model 1 is selected. Figure 4.8 shows classification results for Extroversion (compare to Figure 4.3).

Openness is another “faking high” dimension. Mean honest scores correlated $r = .343$ ($p < .001$) with mean faking scores. In the constrained model, mean honest scores and factor mixture scores correlated $r = -.283$ ($p < .001$), with specificity = .741 and sensitivity = .528. In the model with unconstrained residual variances, mean honest scores correlated $r = .217$ ($p < .001$) with factor mixture scores, with specificity = .353 and sensitivity = .669. Information criteria suggests the selection of Model 1. Figure 4.9 shows classification results for Openness (compare to Figure 4.4).

The third “faking high” dimension is Agreeableness. Mean honest scores correlated $r = .393$ ($p < .001$) with mean faking scores. In the constrained model, mean honest scores and factor mixture scores correlated $r = -.166$ ($p < .001$), with specificity = .071 and sensitivity = .781. In the model with unconstrained residual variances, mean honest scores correlated $r = .324$ ($p < .001$) with mixture factor scores, with specificity = .929 and sensitivity .219. Information criteria suggests the selection of the first model. Figure 4.10 shows classification results for Agreeableness (compare to Figure 4.5). Note the “switch” of the classes.

The last “faking high” construct is Conscientiousness. Mean honest scores correlated $r = .212$ ($p < .001$) with mean faking scores. In the constrained model, mean honest scores and factor mixture scores did not correlate significantly, $r = -.047$ ($p = .243$). Specificity = .906 and sensitivity = .552 in the constrained model. Mean honest scores also did not correlate significantly with factor mixture scores in the unconstrained models, $r = -.008$ ($p = .851$), with specificity = .997 and sensitivity = .10. Information
criteria suggests the selection of the first model. Figure 4.11 shows classification results for Conscientiousness (compare to Figure 4.6).

4.5.2 Additional analyses

Two additional approaches were tested. The first one involves the TRAINING option in Mplus, which is used to incorporate variables that have information about the latent class membership. This option was used in the following way. Half of the observations were randomly selected to be used as training data. Because there are two classes, there are two dummy variables: c1 and c2. Individuals in the honest group were assigned values of 1 for c1 and values of 0 for c2. Individuals in the faking group were assigned values of 0 for c1 and values of 1 for c2. The remaining individuals that were not selected as training data were assigned values of 1 for both c1 and c2 so that their class membership is estimated.

The Extroversion dimension was used to test the TRAINING feature using the previously selected model with item uniqueness constrained equal across classes. Honest scores and factor mixture scores correlated $r = .355 (p < .001)$, with specificity = .958 and sensitivity = .644. This represents an improvement over ignoring faking effects, as mean faking scores correlated $r = .318 (p < .001)$ with mean honest scores. However, this difference would not be meaningful in practice.

The second additional analysis was an application of the faking-as-grade-of-membership (F-GOM) approach delineated by Brown and Böckenholt (2022). The main distinction of that approach from the one investigated here is that they propose considering “intermittent” faking, where respondents may distort their responses to some items but not others. The F-GOM model may also not work with measures of a single
attribute. Unfortunately, the F-GOM models fit to this data failed to converge given the provided starting values in both unidimensional and multidimensional cases. In communication with the corresponding author, it was indicated that both the limited range of the ordinal responses and the instructed faking setting represent significant challenges.

4.6 Summary of findings

Given the simulation results showing that in many conditions factor mixture scores perform worse than mean scores (which do not take faking into account), it was critical to determine if those conditions are also observed in practice. The data used in the empirical application was obtained under the instructed-faking settings of an experiment which does not strictly mimic a real-world application. However, to judge classification performance it was important to have access to both “honest” and “faking” scores.

Overall, factor mixture model scores had low correlations with honest scores across all Big Five dimensions and did not represent an improvement over mean scores which ignore faking distortions. The source of these low correlations is likely the poor classification performance of the mixture models which can be visualized in Figures 4.7 through 4.11. It appears that the model relies almost exclusively on item intercepts to classify responses. In other words, responses are separated by high and low scores without distinguishing honest respondents (who truly have high scores) from faking respondents (who only have high scores because they intentionally distorted their answers in that direction)\(^7\).

One alarming finding in the empirical application relates to Entropy, a statistic

\(^7\) This applies to the “faking high” dimensions in this application (Extroversion, Openness, Agreeableness, and Conscientiousness). In the case of Neuroticism, faking respondents have artificially low scores.
ranging from 0.00 to 1.00 where higher values indicate a higher degree of confidence with which class assignment was performed. In all except one of the models tested, Entropy was very close to 1.00 despite the poor classification of observations. In a real-world, high-stakes assessment, blind reliance on this statistic would lead to dangerously misleading conclusions.
Table 4.1
CFA models and goodness-of-fit indexes for the rating scale version of the BFT (n = 622)

<table>
<thead>
<tr>
<th>Dimension (Model #)</th>
<th>$\chi^2(df)$</th>
<th>rRMSEA [90% CI]</th>
<th>SRMR</th>
<th>rCFI</th>
<th>rTLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuroticism - Model 1 ($\alpha = .91$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original sixteen items</td>
<td>602.776(104)*</td>
<td>.094[.086, .101]</td>
<td>.055</td>
<td>.860</td>
<td>.838</td>
</tr>
<tr>
<td>Extroversion – Model 1 ($\alpha = .90$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original thirteen items</td>
<td>424.706(65)*</td>
<td>.101[.092, .110]</td>
<td>.052</td>
<td>.882</td>
<td>.858</td>
</tr>
<tr>
<td>Openness – Model 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original ten items</td>
<td>192.192(35)*</td>
<td>.089[.077, .102]</td>
<td>.051</td>
<td>.910</td>
<td>.884</td>
</tr>
<tr>
<td>Openness – Model 2 ($\alpha = .82$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 53 removed</td>
<td>147.013(27)*</td>
<td>.089[.075, .104]</td>
<td>.045</td>
<td>.924</td>
<td>.899</td>
</tr>
<tr>
<td>Agreeableness – Model 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original seven items</td>
<td>293.955(14)*</td>
<td>.174[.155, .193]</td>
<td>.089</td>
<td>.726</td>
<td>.589</td>
</tr>
<tr>
<td>Agreeableness – Model 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 16 removed</td>
<td>293.385(9)*</td>
<td>.215[.191, .239]</td>
<td>.099</td>
<td>.729</td>
<td>.548</td>
</tr>
<tr>
<td>Agreeableness – Model 3 ($\alpha = .75$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlated residuals</td>
<td>29.034(6)*</td>
<td>.079[.053, .106]</td>
<td>.033</td>
<td>.973</td>
<td>.932</td>
</tr>
<tr>
<td>Conscientiousness – Model 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original fourteen items</td>
<td>624.700(77)*</td>
<td>.112[.104, .121]</td>
<td>.074</td>
<td>.763</td>
<td>.720</td>
</tr>
<tr>
<td>Conscientiousness – Model 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 51 removed</td>
<td>356.547(65)*</td>
<td>.089[.080, .098]</td>
<td>.061</td>
<td>.842</td>
<td>.811</td>
</tr>
<tr>
<td>Conscientiousness – Model 3 ($\alpha = .82$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 41 removed</td>
<td>219.529(54)*</td>
<td>.070[.061, .080]</td>
<td>.050</td>
<td>.898</td>
<td>.875</td>
</tr>
</tbody>
</table>

Notes. $\alpha = $ Cronbach’s alpha, $\chi^2(df) = $ Scaled chi-square test of model fit and associated degrees of freedom, rRMSEA [90% CI] = Robust root-mean-square error of approximation and 90% confidence interval, SRMR = Standardized root-mean-square residual, rCFI = Robust comparative fit index, rTLI = Robust Tucker-Lewis Index.

*p < .05
<table>
<thead>
<tr>
<th>Dimension (Model #)</th>
<th>logL-value</th>
<th>AIC</th>
<th>BIC</th>
<th>saBIC</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuroticism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>-6121.812</td>
<td>12405.624</td>
<td>12762.981</td>
<td>12505.825</td>
<td>.998</td>
</tr>
<tr>
<td>Uniqueness constrained equal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td>-6397.543</td>
<td>12989.086</td>
<td>13417.032</td>
<td>13109.079</td>
<td>.995</td>
</tr>
<tr>
<td>Uniqueness freely estimated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extroversion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>-5687.582</td>
<td>11497.164</td>
<td>11766.285</td>
<td>11572.624</td>
<td>.987</td>
</tr>
<tr>
<td>Uniqueness constrained equal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td>-5648.261</td>
<td>11442.521</td>
<td>11764.584</td>
<td>11532.825</td>
<td>.696</td>
</tr>
<tr>
<td>Uniqueness freely estimated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Openness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>-3456.373</td>
<td>6994.746</td>
<td>7175.630</td>
<td>7045.464</td>
<td>.999</td>
</tr>
<tr>
<td>Uniqueness constrained equal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td>-3746.419</td>
<td>7590.838</td>
<td>7807.017</td>
<td>7651.453</td>
<td>.789</td>
</tr>
<tr>
<td>Uniqueness freely estimated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreeableness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>-2689.045</td>
<td>5440.090</td>
<td>5576.857</td>
<td>5478.439</td>
<td>.981</td>
</tr>
<tr>
<td>Uniqueness constrained equal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td>-2685.127</td>
<td>5444.255</td>
<td>5607.492</td>
<td>5490.025</td>
<td>.978</td>
</tr>
<tr>
<td>Uniqueness freely estimated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conscientiousness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>-4059.988</td>
<td>8241.977</td>
<td>8511.098</td>
<td>8317.436</td>
<td>1.000</td>
</tr>
<tr>
<td>Uniqueness constrained equal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td>-4479.943</td>
<td>9105.886</td>
<td>9427.948</td>
<td>9196.190</td>
<td>.976</td>
</tr>
<tr>
<td>Uniqueness freely estimated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes. logL-val = log likelihood value; AIC = Akaike information criterion; BIC = Bayesian information criterion; saBIC = sample size adjusted BIC.
Figure 4.1 Visual guide to determine plausible starting values. They will depend on whether respondents are expected to fake high scores or low scores. The diagonal in both cases provides plausible starting values for the honest respondents.
Figure 4.2 Neuroticism scale scores in Wetzel et al.’s data, group-coded.
Figure 4.3 Extroversion scale scores in Wetzel et al.’s data, group-coded.
Figure 4.4 Openness scale scores in Wetzel et al.’s data, group-coded.
Figure 4.5 Agreeableness scale scores in Wetzel et al.’s data, group-coded.
Figure 4.6 Conscientiousness scale scores in Wetzel et al.’s data, group-coded.
Figure 4.7 FMM classification of Neuroticism scale scores in Wetzel et al.’s data.
Figure 4.8 FMM classification of Extroversion scale scores in Wetzel et al.’s data.
Figure 4.9 FMM classification of Openness scale scores in Wetzel et al.’s data.
Figure 4.10 FMM classification of Agreeableness scale scores in Wetzel et al.’s data.
Figure 4.11. FMM classification of Conscientiousness scale scores in Wetzel et al.’s data.
CHAPTER 5: DISCUSSION

The problem of faking in self-report instruments has been studied for quite some time (Doll, 1971; Gordon, 1951; Paulhus, 1984; Phillips & Clancy, 1972; Waters, 1965). A recent approach to address the issue consists of using a statistical model that may approximate the process used by individuals to respond to items in high stakes assessments. This approach often involves the use of mixture modeling in combination with other frameworks, such as item response theory (Zickar et al., 2004), latent difference score models (Ziegler et al., 2015), and multilevel models (Brown & Böckenholt, 2022). This dissertation was positioned within this framework.

The model investigated in this study was motivated by recent research by Pavlov et al. (2019). Using a within-subjects experimental design gathering baseline (honest) scale scores and high-stakes scale scores, they showed that job applicant’s scale scores are best modeled as a function of a) their propensity to fake, b) their true (i.e., honest) scores, and c) the interaction of these two terms. In this context, scores free of distortions due to faking can be accurately estimated. However, Pavlov et al. did not propose any method to extract honest scores in “real-life” assessment settings, when only high stakes data is available.

As a potential solution, I proposed to treat the tendency to fake as a discrete unobserved latent variable, with item responses assumed to depend on a common factor. Measurement models involving both discrete and continuous latent variables are referred to as factor mixture models (FMM). Mixture models are, however, difficult to reliably
estimate (Hipp & Bauer, 2006; McLachlan & Peel, 2000), and exploratory analyzes of Pavlov et al.’s were challenging.

To further investigate the feasibility of using FMM to counter faking, I designed a Monte Carlo study with 72 conditions where the data generating model approximates data that contains faked responses. Only two classes – “honest” and “faking” – were considered. Two outcomes were of particular interest: the correlation of scores estimated using a factor mixture models with the true, generated, scores, and the performance of the classification algorithm. In addition, this modeling approach was tested in an empirical application analyzing data from a instructed faking study by Wetzel, Frick, and Brown (2021).

5.1 Simulation findings

Two types of factor mixture scores were studied. The first are class specific scores for the class with highest posterior probability, and the second are scores obtained by weighting the class specific scores using the posterior probabilities. Both types of estimated factor scores had higher correlations with true factor scores when larger sample sizes where available ($N > 1000$) and when models of larger sizes where used ($p = 12$ or $18$). Interpreting changes in factor score correlations due to model parameters was less straightforward because several significant two-way interactions were found: model size $\times N$, $N \times \varepsilon_{\text{honest}}$, $\lambda_{\text{honest}} \times \nu_{\text{honest}}$, and $\lambda_{\text{honest}} \times \varepsilon_{\text{honest}}$. In general, however, the correlation between true scores and factor mixture scores tends to increase when class discrepancy increases – in particular, as the discrepancy of factor loadings and uniquenesses between classes increases. More importantly, the mixture factor scores represent an improvement over ignoring faking effects in many conditions, particularly
for models of large size ($p = 18$).

Given that correct class assignments are critical in obtaining correct factor scores for each group, several measures of classification performance were considered. Of focal interest are the true positive rate (aka sensitivity, recall, or hit rate), interpreted as the probability of a positive classification conditioned on the observation truly being positive (honest responses), and the true negative rate (aka specificity or selectivity), interpreted as the probability of a negative classification, conditioned on the observation truly being negative (faking responses).

Here, two findings help contextualize the results in the empirical application. The first finding is that specificity is very low in some simulation conditions where we observed high correlations between true factor scores and factor mixture scores. In other words, those correlations are likely elevated because a disproportional number of observations are classified as “honest”, inflating the false positive rate. While this assures that most honest respondents are identified as such (and avoids most adverse effects to job applicants, for example), it also defeats the goal of identifying fake responses.

The second finding is that lower factor loadings for the honest group were associated with worse classification performance. This was at first surprising, because a higher loading discrepancy between the two classes was expected to improve class assignments. However, it is also true that if the two classes have similar factor loadings (so that lines representing factor scores of the two classes are parallel) and large differences in intercept, the model tends to identify classes based on those intercepts.

**5.2 Empirical application findings**

The FMM model was fit to instructed faking scale scores from the study by
Wetzel, Frick, and Brown (2021). In this application, participants were instructed to fake low scores in Neuroticism items and high scores in the remaining Big Five dimensions. Pavlov et al.’s data was used as a reference for obtaining sensible starting values. In the analyses where only starting values were incorporated in the models, factor mixture scores correlated very poorly with baseline (honest) mean scores and did not represent an improvement over ignoring faking effects. Curiously, high entropy values indicate high confidence in the class assignments, while a visual inspection of the scatterplots of each dimension readily suggests that the classification is relying on item intercepts. Another analysis was done using training data for the Extroversion dimension, with a modest improvement in the honest score-factor mixture score correlation but very low sensitivity.

5.3 General considerations

The simulation findings demonstrate that there are select conditions where factor mixture scores perform well, but unfortunately also many cases where it performs worse than ignoring faking distortions. Further, empirical findings indicate that the theoretical potential of factor mixture modeling is also not realized in real-world settings, where class separations are not as defined.

Given the current results, I would recommend against the use of FMM scores in real-world, high-stakes settings such as personality measurement in personnel selection. It is not possible to implement such a model without incurring the risk of poor scoring practices as it was unable to clearly delineate the different classes even on the simplest cases of unidimensional measurement. Moreover, one of the appealing characteristics of personality measures is that they do not have adverse impact. The risk of inaccurate class delineation negates the advantages of that type of measurement from technical and ethical
standpoints.

5.4 Limitations and future directions

In closing, in this study we have investigated whether factor mixture scores represent an improvement over ignoring the effects of faking on scale scores. Within some of the investigated simulation conditions, FMM scores do provide substantially higher correlations with true factor scores compared to mean scores. However, it may be the case that this theoretical potential is not realized in practice, where class separations are not as defined. The Monte Carlo study is also limited by very strict design choices, namely: only normal data was used, model parameters were uniform within each class, and only two classes with a fixed proportion were investigated. Future simulation studies should consider new combinations of all these factors.

The data used in our empirical study shows significant ceiling effects, likely due to the very detailed (and directional) faking instructions that the participants received. Future experiments investigating FMM applications should consider using designs with more natural job applicant prompts. Data obtained under that approach may also prove to be more fruitful as training data.

Alternative analytical strategies should also be considered. Studies on applications of machine learning techniques to the problem of faking are also promising (Calanna et al., 2020; Mazza et al., 2019), although potentially at the cost of interpretability. Additionally, the F-GOM approach recently developed by Brown and Böckenholt (2022) can be implemented with standard software packages and seems capable to identify “real” and “ideal” response profiles. Their method also represents an improvement over previous “ideal-employee factor” approaches (Cellar et al., 1996; Paulhus et al., 1995;
Schmit & Ryan, 1993). The F-GOM method was designed to provide control for intermittent faking (where individuals do not fake all items) and requires multiple dimensions with clear factor structure. Finally, it would be interesting to explore frameworks that can leverage information from process data (Kuncel, Goldberg, et al., 2011) such as response times (Vasilopoulos et al., 2005).
REFERENCES


https://doi.org/10.1037/0021-9010.90.6.1185


https://doi.org/10.1111/j.1744-6570.2010.01173.x


https://doi.org/10.1037/0021-9010.89.1.150


https://doi.org/10.1111/j.1468-2389.2006.00354.x


https://doi.org/10.3389/fpsyg.2010.00165

https://doi.org/10.1111/ijsa.12279


https://doi.org/10.1037/h0054116


https://doi.org/10.2307/1417576


Cattell, R. B. (1948). The primary personality factors in women compared with those in


https://doi.org/10.1080/08959285.2010.501047


Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several Five-Factor Models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality and Individual Differences* (pp. 7–28). Tilburg University Press.


Heymans, G., & Wiersma, E. (1906). Beiträge zur speziellen Psychologie auf Grund einer Massenuntersuchung [Special contributions to psychology from a large-scale investigation]. *Zeitschrift Fur Psychologie Mit Zeitschrift Fur Angewandte*


https://doi.org/10.1177/0013164418791673

https://doi.org/10.1177/109442819812001


https://doi.org/10.1037/0033-2909.96.1.201

https://doi.org/10.1080/10705511.2016.1158655


preparation with practice questions from leading test providers. Kogan Page.


https://www.personalitytestinc.com/


https://doi.org/10.1080/10705511.2013.824780


https://doi.org/10.1086/225231


https://doi.org/10.1371/journal.pone.0274878


https://www.researchgate.net/publication/26514499


https://doi.org/10.1007/BF00383696


https://doi.org/10.1007/s11336-009-9135-y


https://doi.org/10.1037/0033-2909.124.2.262

121


https://doi.org/10.1177/1094428104263674


https://doi.org/10.1037/0021-9010.84.4.551


https://doi.org/10.1177/1094428115574518
APPENDIX A: INSTRUCTIONS FOR MONTE CARLO STUDY

1. Use the first R script to a) generate the data, b) save the .dat files to be analyzed in Mplus as well as a C_list file, and c) save the data as an R object (which I call “simulated_data.RData”) to do the scoring (step 4) without having to read the .dat back into R.

2. The simulation is run with the Mplus input file.

3. Run the second R script to obtain factor scores and perform classification.
APPENDIX B: EXAMPLE R SCRIPT FOR DATA GENERATION

# Set working directory
curDir <- dirname(rstudioapi::getSourceEditorContext()$path)
setwd(curDir)

# Load packages
library(mvtnorm)
library(dplyr)

# Set seed
set.seed(3574)

######################################################################## Write function
generateData <- function (nvars, 
  # Parameters for Honest
  iH, #intercept honest
  lH, #loading honest
  eH, #errors honest

  # Parameters for Faking
  iF, #intercept faking
  lF, #loading faking
  eF, #errors faking

  factorMean, # factor mean for BOTH groups
  factorVariance, # factor variance for BOTH groups

  # Proportion in group 1 (honest)
  probhonest,
  # sample size
  subj,
  # replications (number of datasets)
  reps){
  #### Define parameters
  intHonest <- rep(iH, each = nvars) #vector of intercepts Honest
  laHonest <- rep(lH, each = nvars) #vector of loadings Honest
  psiHonest <- rep(eH, each = nvars) #vector of errors Honest

  intFaking <- rep(iF, each = nvars) #vector of intercepts Faking
  laFaking <- rep(lF, each = nvars) #vector of loadings Faking
  psiFaking <- rep(eF, each = nvars) #vector of errors Faking

  probfaking <- 1 - probhonest #probability/proportion of faking

  honestsubj <- probhonest*subj #sample size honest group
  fakingsubj <- probfaking*subj #sample size faking group

  zeromean <- rep(0, each = nvars) #zero mean to be used for
errors
mean <- rep(zeromean, each = nvars) # vector of factor mean

#### Loops

# True factor scores honest
fHonest <- list()
for (i in 1:reps) fHonest[[i]] <- matrix(rnorm(honestsubj, mean = factorMean, sd = sqrt(factorVariance)))

# Error honest
eHonest <- list()
for (i in 1:reps) eHonest[[i]] <- rmvnorm(honestsubj, mean = zeromean, sigma = diag(psiHonest, nvars, nvars))

# Data honest
honest <- list()
for (i in 1:reps) honest[[i]] <- matrix(rep(intHonest, each = honestsubj), byrow = TRUE, nrow = honestsubj) + 
  fHonest[[i]] %*% laHonest + 
  eHonest[[i]]

# True factor score faking
fFaking <- list()
for (i in 1:reps) fFaking[[i]] <- matrix(rnorm(fakingsubj, mean = factorMean, sd = sqrt(factorVariance)))

# Error faking
eFaking <- list()
for (i in 1:reps) eFaking[[i]] <- rmvnorm(fakingsubj, mean = zeromean, sigma = diag(psiFaking, nvars, nvars))

# Data faking
faking <- list()
for (i in 1:reps) faking[[i]] <- matrix(rep(intFaking, each = fakingsubj), byrow = TRUE, nrow = fakingsubj) + 
  fFaking[[i]] %*% laFaking + 
  eFaking[[i]]

# Combine honest and faking true scores
f <- list()
for (i in 1:reps) f[[i]] <- rbind(fHonest[[i]], fFaking[[i]])

#### Join honest and faking data
data <- list()
for (i in 1:reps) data[[i]] <- rbind(honest[[i]], faking[[i]])

#### Create dummy variable
honestdummy <- list()
for (i in 1:reps) honestdummy[[i]] <- c(rep(1, each = honestsubj), rep(2, each = fakingsubj))

#### Create mean score
obsmean <- list()
for (i in 1:reps) obsmean[[i]] <- rowMeans(data[[i]])
### Combine everything

```r
simulated_data <- list()
for (i in 1:reps) simulated_data[[i]] <- cbind(data[[i]], f[[i]], obsmean[[i]], honestdummy[[i]])
```

### Change datasets from matrix format to dataframe format

```r
for (i in 1:reps) simulated_data[[i]] <- data.frame(simulated_data[[i]])
```

### Change column names

```r
for (i in 1:reps) names(simulated_data[[i]]) <- c('y1', 'y2', 'y3', 'y4', 'y5', 'y6', 'y7', 'y8', 'y9', 'y10', 'y11', 'y12', 'truef', 'obs_mean', 'honest1')
```

# Set number of digits to be equal across the board

```r
is.num <- list()
for (i in 1:reps) is.num[[i]] <- sapply(simulated_data[[i]], is.numeric)
simulated_data[is.num[[i]]] <- lapply(simulated_data[is.num[[i]]], round, 5)
```

# End of loops

# End of function

```r
}
```

### Use the function to create the data

```r
simulated_data <- generateData(  
  nvars = 12,
  iH = 0, lH = .5, eH = .3,
  iF = 1.5, lF = .3, eF = .9,
  factorMean = 0, factorVariance = 1,
  probhonest = .7,
  subj = 500,
  reps = 1000
)
```

# This saves each replication as a dataframe within a list. To see one of the datasets, you can for example do the following:

```
#try <- simulated_data[[1]]
#View(try)
```

### Save data as R object to disk so that it can be easily scored after estimation (instead of reading the .dat files back into R)

save(simulated_data, file = "simulated_data.RData")

### Save .dat files externally to be analyzed in Mplus (could use write.csv instead)

```r
names(simulated_data) <- paste("C", c(1:length(simulated_data)), sep = ".")
lapply(1:length(simulated_data), function(i) write.table(simulated_data[[i]],
```

130
paste0(names(simulated_data[i]), ".dat"),
row.names = FALSE, col.names = FALSE)

### Save list of .dat files (also for Mplus)
C_list <- vector()
for (i in 1:length(simulated_data)) C_list[i] <- paste('C', i, "dat", sep = ".")

write.table(C_list, file = "C_list.dat", row.names = FALSE, col.names = FALSE, quote = FALSE)
APPENDIX C: EXAMPLE MPLPLUS SYNTAX FOR EXTERNAL SIMULATION OF FACTOR MIXTURE MODEL

DATA:
  file = C_list.dat;
  type = montecarlo;

VARIABLE:
  names are y1-y12 truef obs_mean honest1;
  usevariables are y1-y12;
  auxiliary = truef obs_mean honest1;
  classes = c (2); ! class 1 = honest, class 2 = fakers

ANALYSIS:
  type = mixture;
  process = 4;

MODEL:
  %overall%
  f BY y1-y12*;
  [f@0];
  f@1;

  %c#1% !honest
  f by y1-y12*.5;
  y1-y12*.3;
  [y1-y12*0];

  %c#2% !fakers
  f by y1-y12*.3;
  y1-y12*.9;
  [y1-y12*1.5];

SAVEDATA:
RESULTS IS external sim mixture fa results.sav;
APPENDIX D: EXAMPLE R SCRIPT FOR SCORING AND CLASSIFICATION

# Set working directory
curDir <- dirname(rstudioapi::getSourceEditorContext()$path)
setwd(curDir)

# Package to read .sav file
# Open the file "estimates.sav" in SPSS first for correct formatting
(trthere is probably a better way)
library(haven)

## Load package to obtain density function of MVN distribution
library(mvtnorm)

## Used to remove missing data from replications that did not converge
library(dplyr)
library(purrr)

# Set seed
set.seed(3574)

 ###################### Function start
scoreData <- function (nvars, # number of indicators
                      subj, # Total sample size
                      reps, # replications (number of datasets)
                      nf){ # number of factors

  ## Open simulated data
  load("simulated_data.RData")

  # Read parameter estimates
  estimates <- read_sav("estimates.sav")

  # Remove first column (replication number) so that the parameter
  numbers match TECH1 output of Mplus
  estimates <- estimates[, -1]

  ### Class proportions, parameter 73
  probh <- as.data.frame(estimates[, 73])
  profb <- 1 - probh

  #### Class 1 (honest)
  muh <- list()


for (i in 1:reps) muh[[i]] <- t(estimates[i, 1:nvars]) #item intercepts (parameters 1 through 12)

lah <- list()
for (i in 1:reps) lah[[i]] <- t(estimates[i, (nvars+1):(nvars*2)])
#item loadings (pars 13:24)

psih <- list()
for (i in 1:reps) psih[[i]] <- t(estimates[i, ((nvars*2)+1):(nvars*3)]) #item residuals (pars 25:36)

psih.diag <- list()
for (i in 1:reps) psih.diag[[i]] <- diag(x = as.vector(psih[[i]]),
nrow = nvars, ncol = nvars) #diagonal

#test <- lah[[1]] %*% t(lah[[1]]) + psih.diag[[1]]
sigmah <- list()
for (i in 1:reps) sigmah[[i]] <- lah[[i]] %*% t(lah[[i]]) +
psih.diag[[i]]

#### Class 2 (faking)

muf <- list()
for (i in 1:reps) muf[[i]] <- t(estimates[i, ((nvars*3)+1):(nvars*4)]) #item intercepts (pars 37:48)

laf <- list()
for (i in 1:reps) laf[[i]] <- t(estimates[i, ((nvars*4)+1):(nvars*5)]) #item loadings (pars 49:60)

psif <- list()
for (i in 1:reps) psif[[i]] <- t(estimates[i, ((nvars*5)+1):(nvars*6)]) #item residuals (pars 61:72)

psif.diag <- list()
for (i in 1:reps) psif.diag[[i]] <- diag(x = as.vector(psif[[i]]),
nrow = nvars, ncol = nvars) #diagonal

sigmahf <- list()
for (i in 1:reps) sigmahf[[i]] <- laf[[i]] %*% t(laf[[i]]) +
psif.diag[[i]]

### General

mu <- list()
for (i in 1:reps) mu[[i]] <- probh*muh[[i]] + probf*muf[[i]]

sigma <- list()
for (i in 1:reps) sigma[[i]] <- probh*sigmah[[i]] + probh*sigmahf[[i]]

gh <- list()
for (i in 1:reps) gh[[i]] <- solve((lah[[i]] %*% t(lah[[i]]) +
psih.diag[[i]]) %*% lah[[i]])
gf <- list()
for (i in 1:reps) gf[[i]] <- solve((laf[[i]] %*% t(laf[[i]]) +
psif.diag[[i]]) %*% laf[[i]])
### Solution

# Obtain density function of MVN distribution with mean equal to mean and
covariance matrix sigma

```r
## tmph
resulth <- list()
for (i in 1:reps) resulth[[i]] <- dmvnorm(simulated_data[[i]][, 1:nvars], mean = muh[[i]], sigma = sigmah[[i]])

tmph <- list()
for (i in 1:reps) tmph[[i]] <- sapply(resulth[[i]], FUN="*", probh[[i]])

## tmpf
resultf <- list()
for (i in 1:reps) resultf[[i]] <- dmvnorm(simulated_data[[i]][, 1:nvars], mean = muf[[i]], sigma = sigmaf[[i]])

tmpf <- list()
for (i in 1:reps) tmpf[[i]] <- sapply(resultf[[i]], FUN="*", probf[[i]])

## sumprob
sumprob <- list()
for (i in 1:reps) sumprob[[i]] <- tmph[[i]] + tmpf[[i]]

## iprobh
iprobh <- list()
for (i in 1:reps) iprobh[[i]] <- (tmph[[i]])/(sumprob[[i]])

## iprobf
iprobf <- list()
for (i in 1:reps) iprobf[[i]] <- (tmpf[[i]])/(sumprob[[i]])

## dh
dh <- list()
for (i in 1:reps) dh[[i]] <- t(simulated_data[[i]][, 1:nvars] - muh[[i]])

## fh
fh <- list()
for (i in 1:reps) fh[[i]] <- t(gh[[i]]) %*% dh[[i]]  # honest factor score

## df
df <- list()
for (i in 1:reps) df[[i]] <- t(simulated_data[[i]][, 1:nvars] - muf[[i]])

## ff
ff <- list()
for (i in 1:reps) ff[[i]] <- t(gf[[i]]) %*% df[[i]]  # faking factor score
```
# The f_C parameter from Mplus. The respondent's class-specific factor score. Assigned to each person depending on which class is more likely - in the case of two classes, which one has probability > 50%

fc <- list()
for (i in 1:reps) fc[[i]] <- ifelse(iprobh[[i]] >= iprobf[[i]], fh[[i]], ff[[i]])

# f-hat, the weighted sum of honest f-hat and faking f-hat
fweighted <- list()
for (i in 1:reps) fweighted[[i]] <- as.numeric((iprobh[[i]]*fh[[i]]) + (iprobf[[i]]*ff[[i]]))

# Classification, 1 for honest and 2 for faking
class <- list()
for (i in 1:reps) class[[i]] <- ifelse(iprobh[[i]] >= iprobf[[i]], 1, 2)

####### Combine everything
results <- list()
for (i in 1:reps) results[[i]] <- cbind(simulated_data[[i]]$honest1, class[[i]], simulated_data[[i]]$truef, simulated_data[[i]]$obs_mean, fc[[i]], fweighted[[i]])

####### Change datasets from matrix format to dataframe format
for (i in 1:reps) results[[i]] <- data.frame(results[[i]])

####### Change column names
for (i in 1:reps) names(results[[i]]) <- c('honest1', 'class', 'truef', 'obs_mean', 'fc', 'fweighted')

# Set number of digits to be equal across the board
is.num <- list()
for (i in 1:reps) is.num[[i]] <- sapply(results[[i]], is.numeric)
results[is.num[[i]]] <- lapply(results[is.num[[i]]], round, 5)

# End of loops

}
'truef', 'obs_mean', 'fc', 'fweighted')

# For replications that do not converge, missing data can mess up the
# function to create the confusion matrix below
# Create function to select column without any missing value
# remove dataset with missing based on values of "class" - if it NA, take out
not_all_na <- function(x) any(!is.na(x))

# remove column with missing values from the datasets (replications) that had any
for (i in 1:length(test)) test[[i]] <- test[[i]] %>%
  select(where(not_all_na))

# now remove the replications that had missing values (because they did not converge) based on number of columns
test.2 <- list()
for (i in 1:length(test)) test.2[[i]] <- if (ncol(test[[i]]) != 6) {
  test.2[[i]] <- NULL
} else {
  test.2[[i]] <- test[[i]]
}

# In some replications, all observations may be classified as the same group. Tag those as "NULL".
test.3 <- list()
for (i in 1:length(test.2)) test.3[[i]] <- if (length(unique(test.2[[i]]$class)) != 1) {
  test.3[[i]] <- test.2[[i]]
} else {
  next
}

# Remove NULL reps from the list (may take a few seconds)
test.4 <- test.3 %>% compact()

# Correlation tests
correlations <- data.frame()
for (i in 1:length(test.4)) correlations[i, 1:3] <-
cbind(cor(test.4[[i]]$truef, test.4[[i]]$obs_mean),
cor(test.4[[i]]$truef, test.4[[i]]$fc),
cor(test.4[[i]]$truef, test.4[[i]]$fweighted))
names(correlations) <- c('r.obs_mean', 'r.fc', 'r.fweighted')

# Average and SD of the correlations
mean(correlations$r.obs_mean)
sd(correlations$r.obs_mean)

mean(correlations$r.fc)
sd(correlations$r.fc)

mean(correlations$r.fweighted)
sd(correlations$r.fweighted)
### Cross-tabulation

# Load function
source("http://pcwww.liv.ac.uk/~william/R/crosstab.r")

# Save the true class (dummy variable) and classification
confusion.matrix <- list()
for (i in 1:length(test.4)) confusion.matrix[[i]] <- crosstab(test.4[[i]], row.vars = c("honest1"), col.vars = "class", type = "f", addmargins = TRUE)

# Save relevant values from the confusion matrix (True positive, true negative, false positive, false negative, sample size)
conf.matrix.values <- list()
for (i in 1:length(confusion.matrix)) conf.matrix.values[[i]] <- data.frame(TP = confusion.matrix[[i]]$table[1,1],
                                             TN = confusion.matrix[[i]]$table[2,2],
                                             FP = confusion.matrix[[i]]$table[2,1],
                                             FN = confusion.matrix[[i]]$table[1,2],
                                             n = confusion.matrix[[i]]$table[3,3])

# Derivations from the confusion matrix
conf.matrix.results <- data.frame()
for (i in 1:length(test.4)) conf.matrix.results[i, 1:5] <- cbind(Acc = (conf.matrix.values[[i]]$TP + conf.matrix.values[[i]]$TN)/conf.matrix.values[[i]]$n,
                                                                     TPR = conf.matrix.values[[i]]$TP/(conf.matrix.values[[i]]$TP + conf.matrix.values[[i]]$FN),
                                                                     TNR = conf.matrix.values[[i]]$TN/(conf.matrix.values[[i]]$TN + conf.matrix.values[[i]]$FP),
                                                                     PPV = conf.matrix.values[[i]]$TP/(conf.matrix.values[[i]]$TP + conf.matrix.values[[i]]$FP),
                                                                     NPV = conf.matrix.values[[i]]$TN/(conf.matrix.values[[i]]$TN + conf.matrix.values[[i]]$FN))
names(conf.matrix.results) <- c('Acc', 'TPR', 'TNR', 'PPV', 'NPV')

# Averages & Standard deviations
mean(conf.matrix.results$Acc)
sd(conf.matrix.results$Acc)
mean(conf.matrix.results$TPR)
sd(conf.matrix.results$TPR)
mean(conf.matrix.results$TNR)
sd(conf.matrix.results$TNR)
mean(conf.matrix.results$PPV)
sd(conf.matrix.results$PPV)

mean(conf.matrix.results$NPV)
sd(conf.matrix.results$NPV)
APPENDIX E: R SCRIPT FOR CROSS TABULATION

crosstab <- function (..., dec.places = NULL,
                    type = NULL,
                    style = "wide",
                    row.vars = NULL,
                    col.vars = NULL,
                    percentages = TRUE,
                    addmargins = TRUE,
                    subtotals=TRUE)

# Function created by Dr Paul Williamson, Dept. of Geography and Planning,
# School of Environmental Sciences, University of Liverpool, UK.
# Adapted from the function ctab() in the catspec package
# Version: 12th July 2013
# Output best viewed using the companion function print.crosstab()

#Declare function used to convert frequency counts into relevant type
# of proportion or percentage
{
  mk.pcnt.tbl <- function(tbl, type) {
    a <- length(row.vars)
    b <- length(col.vars)
    mrgn <- switch(type, column.pct = c(row.vars[-a], col.vars),
                   row.pct = c(row.vars, col.vars[-b]),
                   joint.pct = c(row.vars[-a], col.vars[-b]),
                   total.pct = NULL)
    tbl <- prop.table(tbl, mrgn)
    if (percentages) {
      tbl <- tbl * 100
    }
    tbl
  }
}

#Find no. of vars (all; row; col) for use in subsequent code
n.row.vars <- length(row.vars)
n.col.vars <- length(col.vars)
n.vars <- n.row.vars + n.col.vars

#Check to make sure all user-supplied arguments have valid values
stopifnot(as.integer(dec.places) == dec.places, dec.places > -1)
#type: see next section of code
stopifnot(is.character(style))
stopifnot(is.logical(percentages))
stopifnot(is.logical(addmargins))
stopifnot(is.logical(subtotals))
stopifnot(n.vars>=1)

#Convert supplied table type(s) into full text string (e.g. "f" becomes "frequency")
#If invalid type supplied, failed match gives user automatic error message

types <- NULL
choices <- c("frequency", "row.pct", "column.pct", "joint.pct", "total.pct")
for (tp in type) types <- c(types, match.arg(tp, choices))
type <- types

#If no type supplied, default to 'frequency + total' for univariate tables and to
#'frequency' for multi-dimensional tables

#For univariate table....
if (n.vars == 1) {
  if (is.null(type)) {
    # default = freq count + total.pct
    type <- c("frequency", "total.pct")
    #row.vars <- 1
  } else {
    #and any requests for row / col / joint.pct must be changed into requests for 'total.pct'
    type <- ifelse(type == "frequency", "frequency", "total.pct")
  }
  #For multivariate tables...
} else if (is.null(type)) {
  # default = frequency count
  type <- "frequency"
}

#Check for integrity of requested analysis and adjust values of function arguments as required
if ((addmargins==FALSE) & (subtotals==FALSE)) {
  warning("WARNING: Request to suppress subtotals (subtotals=FALSE) ignored because no margins requested (addmargins=FALSE)")
  subtotals <- TRUE
}
if ((n.vars>1) & (length(type)>1) & (addmargins==TRUE)) {
  warning("WARNING: Only row totals added when more than one table type requested")
  #Code lower down selecting type of margin implements this...
}
if ((length(type)>1) & (subtotals==FALSE)) {
  warning("WARNING: Can only request supply one table type if requesting suppression of subtotals; suppression of subtotals not executed")
  subtotals <- TRUE
}
if ((length(type)==1) & (subtotals==FALSE)) {
    choices <- c("frequency", "row.pct", "column.pct", "joint.pct", "total.pct")
    tp <- match.arg(type, choices)
    if (tp %in% c("row.pct","column.pct","joint.pct")) {
        warning("WARNING: subtotals can only be suppressed for tables of type 'frequency' or 'total.pct'")
        subtotals<- TRUE
    }
}

if ((n.vars > 2) & (n.col.vars>1) & (subtotals==FALSE))
    warning("WARNING: suppression of subtotals assumes only 1 col var; table flattened accordingly")

if ({ (subtotals==FALSE) & (n.vars>2) } {  
    #If subtotals not required AND total table vars > 2  
    #Reassign all but last col.var as row vars  
    #[because, for simplicity, crosstabs assumes removal of subtotals uses tables with only ONE col var]  
    #N.B. Subtotals only present in tables with > 2 cross-classified vars...  
    if (length(col.vars)>1) {  
        row.vars <- c(row.vars,col.vars[-length(col.vars)])  
        col.vars <- col.vars[length(col.vars)]  
        n.row.vars <- length(row.vars)  
        n.col.vars <- 1  
    }
}

#if dec.places not set by user, set to 2 unlesss only one table of type frequency requested,  
#in which case set to 0. [Leaves user with possibility of having frequency tables with > 0 dp]  
if (is.null(dec.places)) {  
    if ((length(type)==1) & (type[1]=="frequency")) {  
        dec.places <- 0  
    } else {  
        dec.places <-2  
    }
}

#if dec.places not set by user, set to 2 unlesss only one table of type frequency requested,  
#in which case set to 0. [Leaves user with possibility of having frequency tables with > 0 dp]  
if (is.null(dec.places)) {  
    if ((length(type)==1) & (type[1]=="frequency")) {  
        dec.places <- 0  
    } else {  
        dec.places <-2  
    }
}

#Take the original input data, whatever form originally supplied in,  
#convert into table format using requested row and col vars, and save as 'tbl'  
args <- list(...)  
if (length(args) > 1) {  
    if (!all(sapply(args, is.factor)))  
        stop("If more than one argument is passed then all must be factors")  
    tbl <- table(...)  
}
else {
    if (is.factor(...)) {
        tbl <- table(...)
    }
    else if (is.table(...)) {
        tbl <- eval(...)
    }
    else if (is.data.frame(...)) {
        #tbl <- table(...)
        if (is.null(row.vars) && is.null(col.vars)) {
            tbl <- table(...)
        } else {
            var.names <- c(row.vars, col.vars)
            A <- (...)
            tbl <- table(A[var.names])
            if (length(var.names==1)) names(dimnames(tbl)) <-
                var.names
                #[table()] only autocompletes dimnames for multivariate
                crosstabs of dataframes
        }
    } else if (class(...)) == "ftable") {
        tbl <- eval(...)
        if (is.null(row.vars) && is.null(col.vars)) {
            row.vars <- names(attr(tbl, "row.vars"))
            col.vars <- names(attr(tbl, "col.vars"))
        }
        tbl <- as.table(tbl)
    } else if (class(...)) == "ctab") {
        tbl <- eval(...)
        if (is.null(row.vars) && is.null(col.vars)) {
            row.vars <- tbl$row.vars
            col.vars <- tbl$col.vars
        }
        for (opt in c("dec.places", "type", "style", "percentages",
                       "addmargins", "subtotals")) if (is.null(get(opt)))
            assign(opt, eval(parse(text = paste("tbl$", opt,
                        sep = ""))))
        tbl <- tbl$table
    } else {
        stop("first argument must be either factors or a table
                object")
    }
}

#Convert supplied table style into full text string (e.g. "l"
becomes "long")
style <- match.arg(style, c("long", "wide"))

#Extract row and col names to be used in creating 'tbl' from
supplied input data
nms <- names(dimnames(tbl))
z <- length(nms)
if (!is.null(row.vars) && !is.numeric(row.vars)) {
row.vars <- order(match(nms, row.vars), na.last = NA)
}
if (!is.null(col.vars) & is.numeric(col.vars)) {
    col.vars <- order(match(nms, col.vars), na.last = NA)
}
if (!is.null(row.vars) & is.null(col.vars)) {
    col.vars <- (1:z)[-row.vars]
}
if (!is.null(col.vars) & is.null(row.vars)) {
    row.vars <- (1:z)[-col.vars]
}
if (is.null(row.vars) & is.null(col.vars)) {
    col.vars <- z
    row.vars <- (1:z)[-col.vars]
}

# Take the original input data, converted into table format using supplied row and col vars (tbl)
# and create a second version (crosstab) which stores results as percentages if a percentage table type is requested.

if (type[1] == "frequency")
    crosstab <- tbl
else
    crosstab <- mk.pcnt.tbl(tbl, type[1])

# If multiple table types requested, create and add these to
if (length(type) > 1) {
    tbldat <- as.data.frame.table(crosstab)
    z <- length(names(tbldat)) + 1
    tbldat[z] <- 1
    pcntlab <- type
    pcntlab[match("frequency", type)] <- "Count"
    pcntlab[match("row.pct", type)] <- "Row %"
    pcntlab[match("column.pct", type)] <- "Column %"
    pcntlab[match("joint.pct", type)] <- "Joint %"
    pcntlab[match("total.pct", type)] <- "Total %"
    for (i in 2:length(type)) {
        if (type[i] == "frequency")
            crosstab <- tbl
        else crosstab <- mk.pcnt.tbl(tbl, type[i])
        crosstab <- as.data.frame.table(crosstab)
        crosstab[z] <- i
        tbldat <- rbind(tbldat, crosstab)
    }
    tbldat[[z]] <- as.factor(tbldat[[z]])
    levels(tbldat[[z]]) <- pcntlab
    crosstab <- xtabs(Freq ~ ., data = tbldat)
    names(dimnames(crosstab))[z - 1] <- ""
}

# Add margins if required, adding only those margins appropriate to user request
if (addmargins==TRUE) {
    vars <- c(row.vars, col.vars)
if (length(type)==1) {
  if (type=="row.pct")
  { crosstab <- addmargins(crosstab,margin=c(vars[n.vars]))
    tbl <- addmargins(tbl,margin=c(vars[n.vars]))
  } else
  { if (type=="column.pct")
    { crosstab <-
      addmargins(crosstab,margin=c(vars[n.row.vars]))
      tbl <- addmargins(tbl,margin=c(vars[n.row.vars]))
    } else
    { if (type=="joint.pct")
      { crosstab <-
        addmargins(crosstab,margin=c(vars[(n.row.vars)],vars[n.vars]))
        tbl <-
        addmargins(tbl,margin=c(vars[(n.row.vars)],vars[n.vars]))
      } else #must be total.pct OR frequency
        { crosstab <-
          addmargins(crosstab)
          tbl <-
          addmargins(tbl)
        } }
  } }
}

#If more than one table type requested, only adding row totals makes any sense...
if (length(type)>1) {
  crosstab <- addmargins(crosstab,margin=c(vars[n.vars]))
  tbl <- addmargins(tbl,margin=c(vars[n.vars]))
}

#If subtotals not required, and total vars > 2, create dataframe version of table, with relevent
#subtotal rows / cols dropped [Subtotals only present in tables with > 2 cross-classified vars]
t1 <- NULL
if ( (subtotals==FALSE) & (n.vars>2) ) { 
  #Create version of crosstab in ftable format
  t1 <- crosstab
  t1 <- ftable(t1,row.vars=row.vars,col.vars=col.vars)

  #Convert to a dataframe
  t1 <- as.data.frame(format(t1),stringsAsFactors=FALSE)

  #Remove backslashes from category names AND colnames
t1 <- apply(t1[,],2, function(x) gsub("\"","",x))
  #Remove preceding and trailing spaces from category names to enable accurate capture of 'sum' rows/cols
  #[Use of grep might extrac category labels with 'sum' as part of a longer one or two word string...]

145
t1 <- apply(t1,2,function(x)
gsub("[[:space:]]*$","",gsub("^[[:space:]]","",x)))

#Reshape dataframe to that variable and category labels display as required
(a) Move col category names down one row; and move col variable name one column to right
  t1[2,(n.row.vars+1):ncol(t1)] <- t1[1,(n.row.vars+1):ncol(t1)]
t1[1,] <- ""
t1[1,(n.row.vars+2)] <- t1[2,(n.row.vars+1)]
(b) Drop the now redundant column separating the row.var labels from the table data + col.var labels
  t1 <- t1[,-(n.row.vars+1)]

#In 'lab', assign category labels for each variable to all rows (to allow identification of sub-totals)
  lab <- t1[,1:n.row.vars]
  for (c in 1:n.row.vars) {
    for (r in 2:nrow(lab)) {
      if (lab[r,c]=="") lab[r,c] <- lab[r-1,c]
    }
  }
lab <- (apply(lab[,1:n.row.vars],2,function(x) x=="Sum"))
lab <- apply(lab,1,sum)
  #Filter out rows of dataframe containing subtotals
  t1 <- t1[((lab==0) | (lab==n.row.vars)),]

  #Move the 'Sum' label associated with last row to the first column; in the process
  #setting the final row labels associated with other row variables to ""
  t1[nrow(t1),1] <- "Sum"
  t1[nrow(t1),(2:n.row.vars)] <- ""

  #set row and column names to NULL
  rownames(t1) <- NULL
  colnames(t1) <- NULL
}

#Create output object 'result' [class: crosstab]
result <- NULL
(a) record of argument values used to produce tabular output
  result$row.vars <- row.vars
  result$col.vars <- col.vars
  result$dec.places <- dec.places
  result$type <- type
  result$style <- style
  result$percentages <- percentages
  result$addmargins <- addmargins
  result$subtotals <- subtotals

(b) tabular output [3 variants]
result$table <- tbl  # Stores original cross-tab frequency counts without margins [class: table]
result$crosstab <- crosstab  # Stores cross-tab in table format using requested style(frequency/pct) and table margins (on/off)
               #[class: table]
result$crosstab.nosub <- t1  # crosstab with subtotals suppressed [class: dataframe; or NULL if no subtotals suppressed]
class(result) <- "crosstab"
# Return 'result' as output of function
result
}

print.crosstab <- function(x,dec.places=x$dec.places,subtotals=x$subtotals,...) {

# Function created by Dr Paul Williamson, Dept. of Geography and Planning,
# School of Environmental Sciences, University of Liverpool, UK.
# Adapted from the function print.ctab() in the catspec package.
# Version: 12th July 2013
# Designed to provide optimal viewing of the output from crosstab()

row.vars <- x$row.vars
col.vars <- x$col.vars
n.row.vars <- length(row.vars)
n.col.vars <- length(col.vars)
n.vars <- n.row.vars + n.col.vars

if (length(x$type)>1) {
    z<-length(names(dimnames(x$crosstab)))
    if (x$style=="long") {
        row.vars<-c(row.vars,z)
    } else {
        col.vars<-c(z,col.vars)
    }
}

if (n.vars==1) {
    if (length(x$type)==1) {
        tmp <- data.frame(round(x$crosstab,x$dec.places))
        colnames(tmp)[2] <- ifelse(x$type=="frequency","Count","%")
        print(tmp,row.names=FALSE)
    } else {
        print(round(x$crosstab,x$dec.places))
    }
}

# If table has only 2 dimensions, or subtotals required for >2 dimensional table,
# print table using ftable() on x$crosstab
if ((n.vars == 2) | ((subtotals==TRUE) & (n.vars>2))) {
    tbl <- ftable(x$crosstab,row.vars=row.vars,col.vars=col.vars)
    if (!all(as.integer(tbl)==as.numeric(tbl))) tbl <- round(tbl,dec.places)
    print(tbl,...)
}

# If subtotals NOT required AND > 2 dimensions, print table using write.table() on x$crosstab.nosub
if ((subtotals==FALSE) & (n.vars>2)) {
    t1 <- x$crosstab.nosub

    # Convert numbers to required decimal places, right aligned
    width <- max( nchar(t1[1,]), nchar(t1[2,]), 7 )
    dec.places <- x$dec.places
    number.format <- paste("%",width,".",dec.places,"f",sep="")
    t1[3:nrow(t1),((n.row.vars+1):ncol(t1))] <- sprintf(number.format,as.numeric(t1[3:nrow(t1),((n.row.vars+1):ncol(t1))])

    # Adjust column variable label to same width as numbers, left aligned, padding with trailing spaces as required
    col.var.format <- paste("%-%",width,"s",sep="")
    t1[1,(n.row.vars+1):ncol(t1)] <- sprintf(col.var.format,t1[1,(n.row.vars+1):ncol(t1)])

    # Adjust column category labels to same width as numbers, right aligned, padding with preceding spaces as required
    col.cat.format <- paste("%",width,"s",sep="")
    t1[2,(n.row.vars+1):ncol(t1)] <- sprintf(col.cat.format,t1[2,(n.row.vars+1):ncol(t1)])

    # Adjust row labels so that each column is of fixed width, using trailing spaces as required
    for (i in 1:n.row.vars) {
        width <- max(nchar(t1[,i])) + 2
        row.lab.format <- paste("%-%",width,"s",sep="")
        t1[,i] <- sprintf(row.lab.format,t1[,i])
    }

    write.table(t1,quote=FALSE,col.names=FALSE,row.names=FALSE)
}
}