The Pennsylvania State University

The Graduate School

College of the Liberal Arts

MULTILEVEL MEASUREMENT MODELS: IMPROVING ASSESSMENT OF TIME-VARYING CONSTRUCTS WITH A LIMITED NUMBER OF OBSERVATIONS PER PERSON

A Dissertation in

Psychology

by

Andrew Athan McAleavey

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2015

The dissertation of Andrew A. McAleavey was reviewed and approved* by the following:

Louis G. Castonguay
Professor of Psychology
Dissertation Advisor
Chair of Committee

Jeffrey A. Hayes
Professor of Counseling Psychology

Benjamin D. Locke
Associate Clinical Professor of Counseling Psychology

Stephen T. Wilson
Assistant Professor of Psychology

Peter C. M. Molenaar
Professor of Human Development and Family Studies

Melvin Mark
Professor of Psychology
Head of the Department of Psychology

*Signatures are on file in the Graduate School.

Abstract

The most common uses of factor analysis (FA) in psychological measurement may leave uncertainty regarding general versus domain-specific factors, frequently cannot distinguish between traits and states, and do not account for differences between persons when estimating the structure of underlying factors. In part, these problems are due to the use of a single timepoint of assessment for individuals rather than multiple observations. Under circumstances of limited but multiple observations per person, it is proposed that multilevel factor analysis (MLFA) may provide a viable alternative to traditional FA, capable of more reliably distinguishing states and traits, and informing the general vs. specific factor question. In addition, an extension of the MLFA model incorporating randomly-varying measurement parameters may provide an acceptable estimate of idiographic and nomothetic factor structures in large samples, with a relatively small number of observations per person. In two studies, these models are investigated and found to be potentially useful, though also with clear weaknesses in their complexity and computational requirements. In particular, it is suggested that the benefits of such models may most outweigh the costs in routine treatment outcome assessment in applied clinical settings, in which each subject presents with different concerns, theoretical factors of interest are thought to be time-varying, and regular but not intensively repeated observations are made within persons (with as few as 3-10 observations per person being common). Though the studies here remain preliminary, the findings suggest that accounting for multiple observations within persons in psychological measurement may be an important direction for future analyses.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

This document is a testament to the good fortune I have had to benefit from so many generous people. I owe a debt and gratitude to several important individuals for their extraordinary assistance.

First, I would like to thank Dr. Louis Georges Castonguay, the most tireless advocate, supportive advisor, and joyful companion I could imagine. Louis, working with you so closely for so many years is truly a great privilege of my life. Thank you for the encouragement, insight, and trust you have given me (not to mention the amazing bottles of wine).

My entire committee has been a source of support and strength, but especially Jeffrey A. Hayes and Benjamin D. Locke, two individuals who have provided the additional structure I needed throughout my graduate career. Without you, I would not have learned half of what I have learned in graduate school.

I have also been tremendously lucky in my family. Over the last several years, the support of the McAleavey, Perry, Pemberton, Barrett, and Littlewood families has been unending, frequent, and varied. You are the people from whom I hope most to continue to learn, and from whom I take the most of my best traits.

Finally, I cannot overstate the importance of my beloved wife, Greta C. Pemberton, has had in completing graduate school and this Dissertation. Without your jokes, love, support, and energy, I am lost. We gave ourselves a tough task for this Dissertation, and I am glad that our long wait is over. Thank you most of all.

MULTILEVEL MEASUREMENT MODELS: IMPROVING ASSESSMENT OF TIME-VARYING CONSTRUCTS WITH A LIMITED NUMBER OF OBSERVATIONS PER PERSON.

Chapter 1: Introduction

There are few questions in psychological research methods that are more important than that of psychological measurement: how we translate the phenomenological, qualitative experiences of the conscious and unconscious mind into quantitatively measured, objectively defined constructs. Perhaps the single most productive method for such tasks in the last century has been factor analysis (FA), which is often used to determine underlying psychological constructs in observed data. However, there have long been discontents with traditional FA methods, particularly in using FA on interpersonal data to ascertain intrapersonal information. Recently, several extensions of traditional FA have been developed that show promise in improving inferences in situations involving multiple assessments within persons.

These models may be particularly helpful in clarifying and improving the structure of multidimensional assessment instruments that are intended to assess time-varying constructs. Specifically, measures of psychopathological symptoms are intuitively unlikely to be well measured with only one assessment per person, because the severity of symptoms is expected to change over time and a given measure may assess either a narrower or broader range of symptoms than any given person experiences. That is, each person's observed severity at a particular moment on a particular symptom factor (e.g., level of depression) is likely to be dependent on both trait (relatively stable dispositions) and state (time-varying) influences. As a consequence, the constructs of interest are inherently comprised of separate between- and within-person differences.

Newer methods of factor analysis may help clarify variability between and within persons, leading to more accurate models of these processes and potentially, a number of beneficial applications. For instance, it may provide treatment providers with more accurate assessments of current levels of problem severity as well as more accurate predictions regarding treatment outcomes. It may also allow for more idiographic quantitative assessment: allowing scoring methods to vary across people as a function of their particular concerns. Such models may permit accurate generalizations to group-level analyses from within-person variation.

The current studies represent steps toward this goal. First, the utility of multi-level factor analysis (MLFA) in assessing psychological symptoms over time is tested and compared to more standard techniques. MLFA, as compared to traditional FA, allows for distinct covariance structure models to explain differences between people and differences within people over time. The primary aim of this study is to determine whether the variation between people on a measure of psychological symptoms is similar to the variation within people, or if trait-like and state-like symptoms are not essentially the same. This study also compares MLFA to more common and traditional methods of factor analysis.

In the second study, certain measurement parameters were allowed to vary randomly across people. Unique factor loadings were estimated for each participant. In this way, a somewhat unique factor structure was calculated for each person based on deviation from a group factor structure, using a random variable framework. Results of both studies are discussed in terms of benefit to individual case assessment, studies of groups of people (samples and populations), with a particular emphasis on applications to naturalistic mental health treatment environments.

Before describing the proposed studies, some more detailed background is required. This document is structured by first reviewing the history, theory, and methods of traditional factor analysis; then discussing some longstanding problems with these traditional FA methods; and the rationale behind a few promising more recently-developed methods for this purpose. Each of two studies is then described with results.

Traditional Factor Analysis

Factor analysis has been one of the most productive statistical methods in psychology research, particularly for measurement of psychological variables. Pioneered by Spearman (1904) and further developed by Cattell (1965) and others, this dimension-reduction strategy has been used most frequently to help determine what underlying (unobserved) psychological processes might contribute to observed differences and similarities between individuals. These "factors" (unobserved latent variables) have a variance within the population of observations, and a set of regression equations (which include the factor loadings, measures of the strength of a relationship between the observed variable and the latent factor score) from the observed variables that determine the level of the factor for each observation. Conceptually, the most common interpretation of factor analysis is that the unobserved factors "cause" the observed variables. That is, in psychology, factor analysis is often used to determine the structure of mental, emotional, and attitudinal phenomena that cannot be directly observed. A single factor model is depicted in Figure 1. I describe here some of the most common methods of FA as they appear in the psychological literature.

Perhaps the most essential (and first-developed) form of factor analysis is an exploratory factor analysis (EFA). In EFA, observed test or item scores are conceived of as being caused by a smaller number of unobserved factors. As a dimension-reduction strategy, the main objective

of EFA is to determine the number of these underlying factors that might explain the observed data. Often, the number of factors is a primary interpretive question, since these factors are frequently intended to represent the "real" psychological features: the actual intrapsychic processes of interest.

Intelligence, which was one of the earliest applications of factor analysis, offers an excellent example of the use and interpretation of factor analysis to determine psychological constructs. An attempt to determine the underlying components of intelligence has immense consequences for applications from early childhood education to job hiring, and helps determine the provision of treatments (e.g., stimulant medications for Attention-Deficit Hyperactivity Disorder) and appropriateness for remedial and advanced education programs. Spearman (1904) contended, on the basis of a FA of many individuals' performances on several putatively different tests of intelligence, that a "general intelligence" factor, rather than several discrete types of intelligence, would sufficiently explain the relationships between these tests. This "*g*" factor has been a primary aim of assessment in intelligence since that time, as many researchers have agreed that one underlying factor of intelligence seems to sufficiently describe variability between persons on intelligence tests (Neisser et al., 1996). However, other researchers have contended that more than one intelligence factor should be considered. These multiple-intelligences are often represented as correlated multiple factor models as in Figure 2. As mentioned below, modern FA techniques have altered the discussion somewhat, though the basic disagreement – one intelligence or several – remains. We return to the topic of intelligence below.

Although it is frequently of primary importance, the number of factors is frequently not obvious based on results of FA. This is partially due to (or in spite of) the fact that are a number

of methods for determining the number of factors to extract, many of which have some empirical support and may diverge in real data. Among the most common methods are Kaiser's (1960) eigenvalue test, Cattell's (1966b) scree test, and parallel analysis (Horn, 1965). In addition, when EFA is conducted in a structural equation modeling program, model indices such as the chi-square, likelihood-ratio tests, and incremental fit indices are also available and often efficient (Fabrigar et al., 1999), and recent suggestions for evaluations of the number of factors have included comparison to simulated data with known factors (Ruscio and Roche, 2012). Thus, there exists a plethora of choices for the researcher to determine the number of factors to extract. Once the number of factors has been determined using whatever method, exploratory factor analysis usually proceeds with factor rotation, a method primarily intended to improve the interpretability of factors. That is, rotated factors should represent understandable concepts or constructs. In the case of instrument design, these factors are usually then scored as subscales: theoretically distinct though frequently correlated scores of a subset of items from a measure. In many applications, the final purpose of factor analysis is precisely to determine the optimal scoring method for a measure: how many subscales should be used for a given measure (represented by the number of factors extracted), and which items or tests should contribute to each subscale (represented by meaningful factor loadings on the factors).

Subsequent to EFA, it is common to conduct a confirmatory factor analysis (CFA). In CFA, the aim is usually not to discover the underlying organization of the observed data, but instead to test a proposed model against data. By definition, CFA models are more restrictive than EFA models: instead of allowing potentially every observed variable to load on every factor, certain factor loadings are fixed at 0: no direct relationship is allowed between an observed variable and a latent factor. Often, the pattern of fixed variables is determined by

examination of the rotated solution of the EFA, with large loadings from the EFA estimated and small loadings from the EFA held at 0. The CFA method allows for tests of specific hypotheses, which an EFA does not.

The method of FA described above has been remarkably productive in psychological research over the last century. Nevertheless, several conceptual and empirical problems with these methods have been raised.

Difficulties with traditional FA regarding covariance between factors: general and specific factors

Some of the most important difficulties of FA have been caused by one of the very problems that FA was intended to resolve: the relationship of general (over-arching or broad-spectrum) factors and specific (focused, theoretically limited) factors. In the study of intelligence, this has meant that though Spearman's suggested *g* factor is indeed widely used and accepted, there have been prominent opponents (e.g., Gardner, 1983; Sternberg, 1985) who have argued that this general factor does not sufficiently capture variability between persons. Instead, they have argued that the more accurate description of intelligence is as a multidimensional construct. For instance, Sternberg (1985) suggested primarily three intelligences: analytical, creative, and practical. There remain two opposing viewpoints on intelligence (crudely defined): one factor versus many factor theories.

This is emblematic of several domains of psychology research in which the underlying number of essential constructs is not empirically clear and several competing possibilities are espoused by different researchers (e.g., personality trait theory, psychopathology research). Moreover, the question of the number of factors has had important implications especially when the factors being discussed are either a limited number of general factors or a larger number of

more specific factors. In personality, for instance, much trait theory research has focused on the "Big Five" (McCrae and Costa, 1987; 2012) personality traits, which are conceptually assumed to be the essential types of personality traits across cultures. Though differing perspectives endorsing fewer essential traits have been explored: for instance, some have suggested two factors higher-order factors (Digman, 1997), and recently a general factor of personality has been suggested (Rushton & Irwing, 2011). In psychopathology research, the organization and identification of disorders has also hinged on the application of factor analysis and interpretation of general and specific factors (for instance, the work of Brown, Chorpita, & Barlow, 1998). Results of factor analyses are sometimes taken as evidence for and against inclusion of diagnostic categories in research and clinical practice. Because this has been viewed as such an important part of psychological research, additional methods of factor analysis that directly address the general-specific factor relationships have been proposed and used over the years (some very soon after FA's introduction, others only in the past few years).

Higher-order factor analysis. One of the most basic adaptations of FA to the general-specific question is higher-order factor analysis. Depicted in Figure 3, a higher-order factor analysis is primarily useful to explain covariances between obliquely rotated (and therefore meaningfully correlated) factors. In essence, a higher-order factor is frequently conceptualized as an unobserved cause of the unobserved causes (first-order factors) of observed variables. Although this common-cause interpretation is not the only mathematical interpretation of such a factor solution (see, e.g., Adcock, 1964), it has been frequently interpreted it in this way. In higher-order models, then, a general factor is seen as an indirect cause of observed variables, acting only through the variance of the lower-order, more specific factors. While there can be any number of higher-order factors (or, conceptually, any number of orders), in most applied

models a limited number (often one or two) higher-order factors are sufficient to explain the covariances between factors. These higher-order factor models have been common in many of the fields described previously. Many contemporary studies examining the structures of intelligence, personality, and psychopathology have employed and relied on higher-order factor models or indirectly provided support for the dominant Cattell-Horn-Carroll higher-order model (e.g., Keith and Reynolds, 2010).

Bifactor analysis. Conceptually, the higher-order factor model implies that general factors are not directly but indirectly related to observed variables. This assumption is altered in the recently popular Bi-factor (or bifactor) model, first developed by Holtzinger and Swineford (1937). In this model, depicted in Figure 4, the specific factors that load only on a subset of the observed variables, and a general factor, which loads on all observed variables, are conceived on the same order. An important development of the bifactor model was made by Schmid and Leiman (1957), who showed that under the circumstances of simple structure, any single-level oblique factor solution can be transformed into an orthogonal bifactor model. To these authors, this increased the interpretability of the factor solution beyond a correlated factor solution. Other authors have also agreed that the bifactor model (or a more generalized hierarchical model) has advantages. For instance, Chen, West, and Sousa (2006) and Reise, Moore, and Haviland (2010) both suggest that the bifactor model can be used as the initial or baseline against which successively more restrictive models (including higher-order factor models) are tested. Both higher-order and bifactor models serve a shared function: explaining (or accounting for) relationships between specific factors while also providing estimates of general factors that cover multiple domains. Neither produces unequivocal results, however.

Exploratory structural equation modeling. Recently, yet another response to correlations between domain-specific factors has emerged. This method, termed exploratory structural equation modeling (ESEM) relaxes an expectation of many researchers have held even before Thurstone (1947) and Carroll (1953) formally suggested it: the utility of simple structure. In brief, simple structure is a situation in which each observed variable has one and only one meaningful factor loading, and has no factor loading on the other domain-specific factors in the model. In practice, the value of simple structure is that it renders subscale scoring much more straightforward: the alternatives include multiple cross-loading variables and/or complex weighting procedures. However, as described by Asparouhov and Muthén (2009), the insistence that variables have factor loadings equal to 0 is often not a good approximation of the data. In ESEM, instead of assuming that the non-primary factor loadings for each variable should be exactly 0, it is assumed that some small factor loadings will be acceptable. Marsh et al. (2009; 2010) have provided examples using ESEM, demonstrating that ESEM is feasible, can improve model fit over simple-structure confirmatory factor analysis models, and that one effect of doing so is a marked decrease in correlations between (first-order) factors. This is a considerably different conceptual interpretation of inter-factor correlations than in bifactor and higher-order factor models: rather than trying to explain correlations between factors using interpretable general factors, in ESEM one of the primary goals is to minimize those correlations. This is done with the assumption that inter-factor correlations are partially an artifact of model misspecification, and that small non-zero loadings are a closer approximation of the truth than fixed 0 loadings (which imply that some observed variables are not directly related to some latent variables at all).

General and specific factors in treatment outcome monitoring. Because the questions of general and specific factors relate closely to the nature of personality and psychopathology, it is not surprising that many of the measures of psychopathology that are routinely used in treatment settings have been subjected to several different factor analysis. The particular context of interest to this thesis is treatment-outcome monitoring (TOM) during psychological treatment, an increasingly popular adjunct to traditional psychological therapy. In TOM, clients are asked to complete a quantitative outcome measure at several (often each) appointments with a treatment provider. This information is frequently used in aggregate to evaluate treatment quality (e.g., Minami et al., 2009), differential effectiveness of therapists (e.g., Kraus, Castonguay, Boswell, Nordberg, and Hayes, 2011), and by individual therapist-client dyads through various feedback mechanisms (e.g., Lambert, Harmon, Slade, Whipple, & Hawkins, 2005). One of the major reasons that TOM has become popular in recent years is its potential to assess and improve routine care (Castonguay, Barkham, Lutz, & McAleavey, 2013).

Several of the most common instruments used for psychotherapy TOM – the Counseling Center Assessment of Psychological Symptoms (CCAPS; Locke et al., 2011; Locke et al., 2012), the Outcome Questionnaire-45.2 (OQ-45.2; Lambert et al., 2001), the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM; Evans et al., 2002), and the Behaviour and Symptom Identification Scale-24 (BASIS-24; Eisen, Normand, Benager, Spiro, & Esch, 2004) – all use some domain-specific scores along with one more general score. Several of these measures have received support for this general-specific factor structure in a higher-order and/or bifactor FA. For instance, Thomas (2011) illustrated the bifactor model in an analysis of the Brief Symptom Inventory (BSI; Derogatis & Melisaratos, 1993). This study not only reported an improvement in model fit when using bifactor compared to unidimensional and oblique simple

structure (a higher-order factor was not investigated), but also slightly improved diagnostic accuracy for the bifactor model. Thomas took this as preliminary evidence that bifactor factor models may provide increased clinical utility over traditional FA models, and recommended using both the general and specific factors resulting from this model. In another example, Bludworth, Tracey, and Glidden-Tracey (2010) examined the factor structure of the OQ-45.2, which is commonly used as a unidimensional measure of overall clinical dysfunction, though it was developed to capture three dimensions: symptom distress, interpersonal relations, and social role performance. The lack of support for these domain-specific subscales led Lambert et al. (2001) to caution against interpreting these scores. Bludworth et al. (2010) found that a bifactor model that accounts for these three subscales was a superior fit to the data than several alternate models (including a higher-order model). Though the authors interpreted this as support for the subscale scores of the measure, overall model fit was not objectively good even in the bifactor model (CFI = .80), potentially suggesting some remaining uncertainty.

As noted by McAleavey, Nordberg, Kraus, and Castonguay (2012), the general versus specific factor problem has important implications for TOM systems in psychotherapy. McAleavey et al. (2012) argued that the general and specific scores used in TOM, regardless of the underlying factor analytic model, should have markedly different error profiles, particularly when the clients being treated present with diverse problems and symptoms. For instance, though a client with broad distress and negative affect may report many symptoms across measures of depression, anxiety, and functional impairment, they may not report difficulties with eating disorders or alcohol abuse, necessarily. In this case, a broad distress score would most likely be an adequate marker of their distress. However, for a client with more specific concerns (e.g., someone with problematic drinking behaviors but little overall distress), a general score would

not adequately capture their experience. Nor would this client's level of general distress be conceptually comparable to his or her highly distressed but non-drinking peers, if it were measured as a simple average of diverse responses. That is, when a general measure of distress is used, two people with the same general score may have reported considerably different experiences, which may affect the quality of the measure. Additional considerations are suggested by McAleavey et al. (2012).

This situation is analogous to a clinical intelligence assessment in which two different people have demonstrated markedly divergent patterns of subtest scores while achieving similar full-scale intelligence quotient scores. Though the general intelligence factor is widely supported in theory and research, the specific factors are frequently used as well: it is not an either-or situation. In the applied practice of intelligence testing for instance, it is common to interpret the general factor less for individuals who show significant "scatter" among subtests, and some studies have suggested that this leads to improved diagnostic accuracy compared to interpreting the general score alone when subtests are divergent (e.g., Fiorello, Hale, McGrath, and Ryan, 2001; Fiorello et al., 2007), though some evidence to the contrary has also been presented (e.g., Watkins, Glutting, and Lei, 2007). That is, whether one believes that *g* underlies intelligence or multiple intelligences better describe the population, if a given person's scores do not appear to be indicative of a general underlying factor, no such factor should necessarily be assumed. Thus, there seems to be a disconnect between the widely viewed theory – that general intelligence is real psychological process, and discriminates people meaningfully – and idiographic clinical work, in which it is commonly believed that for many people general intelligence is not as important as the domain-specific subtest scores.

In many naturalistic research studies of treatments for mental health conditions, similar challenges are present. Just as with the use of *g* as an indicator of general intelligence, in many studies, psychopathology is assessed using general measures, intended to capture the sum total of all mental health symptoms or dysfunction across a broad range of clients. This is sometimes done using a general score from one of the self-report measures discussed above. However, quite unlike in the study of intelligence, there has been little theoretical suggestion that general psychopathology or distress measured in this way is an important hypothetical construct at all. Instead, overall distress of the type measured in a general score is quite frequently taken (at best) as an intervening variable which bears on the construct of interest, such as depression or anxiety symptoms (which frequently lead to clients expressing high levels of distress across numerous domains). Perhaps more commonly, however, general measures are used as a statistical convenience, which is not substantively intended to measure any particular construct at all. In the latter case, general distress measures are relied upon to provide an average of concerns; this average may be useful without providing detail about the types of concerns that are most relevant to a given person. For instance, when seeking to determine the effectiveness of an entire clinic, it is helpful to have one outcome measure (as in Minami et al., 2009) even if this is an imprecise measure. This use of a general score is most accurate for cases that conform to a general factor structure; but as in intelligence assessment, it is possible that some people in psychological treatment may not conform to this model. People with a focused concern will not be measured well by a general score, and measurement of their outcome will appear muted if not appropriately measured. Because of this, determining whether a general measure is appropriate (in what cases, or how it should be interpreted) may be very important not only in assessing an individual client, but also in aggregate uses of heterogeneous clinical data.

The preceding discussion of general-specific factor relationships has to do with the difficulty of interpreting correlations between factors. While researchers have made use of various FA methods, the substantive research questions involved in determining the presence of general and/or domain-specific factors has rarely been unequivocally resolved in psychological research. Further, in the case of psychological assessment in ongoing treatment, a theoretical understanding of any general distress factor has frequently been absent. In order to improve assessment of individuals and groups of clients in therapy, the relationship between general and specific factors of psychological distress need to be understood. I will return to this after discussing a second broad category of problems with factor analytic methods as they have been used most commonly.

Difficulties in traditional FA regarding the use of a single timepoint

Another problem with traditional FA, particularly in the study of variables that are not constant over time, is the conflation of interindividual and intraindividual variation. That is, the nomothetic may not represent the idiographic. Several persuasive cases have been made that more idiographic analysis is necessary in psychological studies, particularly of developmental processes (e.g., Barlow & Nock, 2009; Cervone, 2004; Kelderman & Molenaar, 2007; Molenaar, 2004; Shoda, Mischel, & Wright, 1994). One particularly relevant argument to psychometrics was made by Molenaar (2004) and relates to the concept of ergodicity.

The ways that a given person changes through time (day-to-day, minute-to-minute) could be affected and defined by slightly or extremely different factors than the things that make people differ from one another. A deceptively simple example might be that of genotypic variation: though every person's unique genetic makeup influences interindividual differences, as a static code it has limited ability to influence a given person over short time spans (though it

clearly does through epigenetic and developmental mechanisms, genetic changes are not rapid enough to influence many psychological processes in a direct manner). This is an explicit violation of ergodicity, a formal mathematical set of principles defining a system in which the interindividual variation and intraindividual variation are (asymptotically) equivalent (Molenaar, 2004). There are other features of ergodic systems: for instance, stationarity, which implies that a system does not change over time (most psychological features change with development, circadian rhythms, moods, and so on). Suffice it to say that psychological processes are probably never ergodic, and therefore assumptions that interindividual data will parallel intraindividual data are very weakly founded (Molenaar, 2004). That is, there is good reason to think that a result based on differences between people will have little bearing on the actual process occurring for any given person in that sample.

Within and between-person variation in factor analysis. Though the incorporation of ergodic theory into psychology is relatively recent, many researchers have made use of the difference of between-person and within-person variation in other ways. One of the first acknowledgements of the significance of this difference was made by Cattell (1943), when he made explicit the assumption that interindividual differences and intraindividual differences are conceptually and analytically distinguishable. However, the field of psychological measurement, as a whole, has continued to rely on what Cattell (1965) referred to as R-technique factor analysis, in which many individuals are sampled at a single timepoint, and factor analysis is performed on the resulting covariances among their scores. This method makes the level of analysis differences *between* people, and emergent factors representing the ways in which individuals differ from and are similar to one another (hence, interindividual variation). These factors are generally taken to be indicative of mental processes, as in a general factor of

intelligence that all people have to some degree, universal personality traits that identify individuals as more- or less-extraverted than others, and so on.

However, there is a fundamental problem in interpreting R-technique factor analysis as assessing pertinent information that might change over time: we cannot discriminate between measurement of traits (features of behavior which are thought to be more stable over time) and states (features of behavior which are by definition temporary). This difficulty was noted by Cattell (e.g., 1966a), and is logically obvious: Any person, observed at a given timepoint, will be influenced by *both* traits and states. As an example, self-reported feelings of hopelessness will be influenced by relatively stable personality traits (such as level of neuroticism and optimism) as well as recent events in their lives (such as having recently lost a job, or sleep quality). A failure to distinguish between these state and trait influences will produce an incomplete description of the individual's level of hopelessness at best, and a misleadingly inaccurate depiction of psychological functioning at worst.

Further, it is entirely possible that the variables that make persons different from one another (stable traits distinguished by interindividual variation) are not identical to the variables that make a given person different over time (unstable states distinguished by intraindividual variation). In psychometrics, Cattell's (Cattell, Cattell, & Rhymer, 1947) suggested primary alternative factor analysis method is called P-technique. Mathematically, p-technique is identical to standard R-technique factor analysis, of the sort described in the preceding sections. The difference is that instead of collecting data from many people at a single timepoint each, p-technique is entirely idiographic: a single individual is assessed at a several (often 50 to several hundred) timepoints. The resulting data is subjected to factor analysis of these intraindividual observations. P-technique factors, then, are person-specific: though multiple people could be

assessed using p-technique, the factor structure is not assumed to be the same from one person to another. Conceptual differences between P-technique and R-technique factor analysis were clearly demonstrated by Borkenau and Ostendorf (1998) and re-analysis of the same data by Hamaker, Dolan, and Molenaar (2005). These studies compared the factor solutions of a measure of the Big Five personality traits, both in interindividual variation and intraindividual, assessed in daily self-report questionnaires. Though Borkenau and Ostendorf (1998) reported substantial match between the standard Big Five and a cross-sectional (interindividual) analysis of the data, both groups of authors reported poor fit of the "standard" Big Five model to individual-specific time series data. That is, the Big Five personality traits – the most widely used personality trait theory in research – failed to adequately represent the actual functioning of many of the individual participants while appearing to adequately reflect aggregate interindividual data. The researchers concluded that p-technique factor analysis seems able to detect differences that standard factor analysis cannot.

Cattell also suggested a second method of studying time-varying factors, which is less person-specific but captures a different type of variation entirely. This method, differential-R or dR-technique, involves assessing a large number of participants as in R-technique, but then assessing the same participants at a later timepoint as well. Subsequently, the differences between time 1 and time 2 for each participant are used as the data for factor analysis. As demonstrated by Cattell and Bartlett (1971), dR-technique can be used to find common state variables, and when paired with R-technique factor analysis, may helpfully identify common traits and states. According to Cattell (1966a; Cattell and Bartlett, 1971), this is because, conceptually, a single timepoint contains both state and trait, but difference-score data, if sufficient time has elapsed for states to change, has subtracted trait influences. Both R and dR

techniques, unlike P-technique, explicitly model common factors: factors that are theoretically present in all members of a sample or population (notwithstanding the previous discussion of ergodic theory that has disproved this), and these common factors have been primarily of interest to psychological researchers. The use of dR technique seems to have largely been limited to Cattell and his research group: few studies have explicitly used this terminology. However, it is possible that the idea itself has been used under different names. One example was Minami et al. (2009), who report results of a related factor analysis as part of a larger process. In this paper, the authors did not calculate simple change scores, as Cattell and his colleagues did, but instead used residualized change scores. Moreover, the discussion of factor structure is limited because it is presented as an ancillary analysis. However, it is possible that other papers have also factor analyzed difference scores without using Cattell's terminology.

More recent analyses of within- and between-person variance. More recently, however, several developments have altered the research landscape regarding psychological measurement. One contributor has been the development and popularization of dynamic factor analysis (DFA; Molenaar, 1985) as an extension of P-technique factor analysis. Among the reasons that P-technique has not become popular is an inherent violation of one of the assumptions of factor analysis: that of independent observations. That is, factor analysis is predicated on the assumption that each observation in a data set is not more highly correlated with any other observation than it is with all other observations, except through the underlying factors. Under R-technique data, this is generally a reasonable assumption: No two participants should be especially related to each other, if all have been randomly selected. When data is collected from an individual, however, this assumption is tenuous at best: Sequential observations may be dependent on or closely related to one another. Standard p-technique

analyses do not account for this violation of assumptions. Though at least one study has clearly suggested that this does not provide any decrement to accuracy or increase in parameter bias (Molenaar & Nesselroade, 2009), the failure to account for these sequential dependencies is a weakness of this model. In contrast, DFA flexibly estimates these sequential dependencies along with permitting additional structural relationships among factors beyond p-technique. Methods based on DFA have become somewhat more common in the psychological literature (e.g., Fisher, Newman, & Molenaar, 2011; Sinclair & Molenaar, 2008), and have demonstrated value in studying multivariate time series.

A second development has been the increasing frequency of intensive longitudinal data collection. Intensive longitudinal data include dozens or hundreds of observations of a given person, often multiple times per day for weeks at a time. These longer time series permit more flexible and accurate quantitative analyses of intraindividual variation, including nonlinear change (Hayes, Laurenceau, Feldman, Strauss, and Cardaciotto, 2007; Laurenceau, Hayes, and Feldman, 2007), and an increase in interest in differences between people in the process of intraindividual change (e.g., Vittengl, Clark, Thase, and Jarrett, 2013). The increase in number of observations per person in such studies has not only allowed a more precise examination of individual subjects' mean levels on constructs of interest, but also quantitative comparisons of subjects' variability over time (often expressed as a version of an individual standard deviation (e.g., Hedeker, Mermelstein, & Demirtas, 2012; Ram & Gerstorf, 2009). This is a point worth noting, since it is conceptually and mathematically distinct from traditional FA models. Individual participants may differ in the extent to which they change over time: some people may be very stable, demonstrating little variability, others may be highly unstable, and anything in between. Note that there can be no estimate provided for this quantity in standard R-technique

FA models: the only variances that are estimated are (between-person) factor variances, interpretable for the population, not an individual.

In summary, it can be said that factor analysis, if it is based on only a single time point of multiple people, will be incomplete in its ability to model individual-specific processes. Moreover, these individual-specific processes may be important, for instance when a treatment for a given individual is under consideration. Molenaar (1987; Sinclair & Molenaar, 2008) has demonstrated for a single case of psychotherapy, a method making use of intraindividual variation to predict (and potentially improve) the effectiveness of certain therapeutic process variables. These methods are person-specific: they apply directly to a single case, and as such do not necessarily inform other cases. However, they show promise in flexibly applying DFA to psychotherapy while being sensitive to person-specific reality of psychological measurement.

Finally, a method that has become essentially standard in many longitudinal psychological studies explicitly discriminates between- and within-person variance. This is called multilevel linear modeling (MLM; also referred to as hierarchical linear modeling), in which several individuals are observed repeatedly at different times. These multilevel models essentially partition observed scores into differences within persons over time and differences between persons on average across all time points. The most common form of longitudinal MLM is probably a latent growth curve model, in which a single outcome variable is tracked over time. Instead of a single group trajectory, individual participants are each assigned – estimated – a trajectory of this variable over time, which takes into account the individual's data: their individual intercept and slope(s) can be modeled as part of determining their estimated trajectory. Such estimates from MLM have advantages, including the fact that they are not bound by the assumptions of some alternative analyses (basic regression and ANOVA without repeated

measures, for instance) that require independent observations. These advantages have led to a large number of publications using MLM as a replacement or improvement on other general linear model-based methods.

However, it should be noted that in a latent growth curve or other MLM, the individual estimates or trajectories are not person-specific as a p-technique factor analysis and related models. In p-technique, the only data that informs an individual's factor structure is that individual's data; in MLM, all individuals' data can influence all individuals' estimates. This occurs through a very specific mechanism: the use of a probability distribution (usually Gaussian) to model differences between people. This leads to "empirical Bayes" or "shrinkage" estimates, which has been viewed as a strength of the models (e.g., Lambert and Ogles, 2009): because they incorporate data from the sample as a whole, it has been suggested that empirical Bayes estimates may be more reliable than other methods, particularly in small numbers of observation per person (Diez Roux, 2002). In general, empirical Bayes estimates will tend to estimate less extreme values for people whose actual data is far away from the sample average – hence the name "shrinkage" estimates. The amount of shrinkage is determined by the reliability of the individual's data, such that a few observations deviating from the group mean would induce substantial shrinkage, while many observations which reliably seem to differ from the group mean would not shrink as much.

The methods of factor analysis and MLM have much in common. All FA models (EFA or CFA) are specific instances of SEM. Likewise, a two-level MLM involving only clients assessed over time is also an instance of SEM (Bauer, 2003), and MLM involving more levels are specific instances of the more general multilevel structural equation models (MLSEM; Mehta & Neale, 2005). However, MLSEM models, though the most general of the group of models

here considered have not been as widely used as the simpler MLM models. This is not merely because they are difficult for statistical programs to estimate (although that is somewhat true): Several modern statistical programs including Mplus (Muthén & Muthén, 1998-2012), SAS (SAS Institute, 2011), OpenMx (Boker et al., 2011), and MLwiN (Rabash et al., 2011) are capable of estimating these models. Indeed, MLSEM seems to be developing some following, particularly for multilevel mediation analysis (Preacher, Zyphur, & Zhang, 2011) and in analysis of large cross-sectional clustered datasets (such as surveys of many different countries, as in Mueleman & Billiet, 2009). However, there has not yet been a significant presence of MLSEM clinical psychology. One notable exception might be the group psychotherapy process model proposed and investigated by Johnson et al. (2005), which has been influential in group psychotherapy research. This application has shown some differences in processes between and within groups of psychotherapy. Nevertheless, few studies have used MLSEM apart from latent growth curve models in psychotherapy research. The studies proposed below aim to explore the potential of MLSEM to improve measurement of psychological constructs during the routine process of psychotherapy.

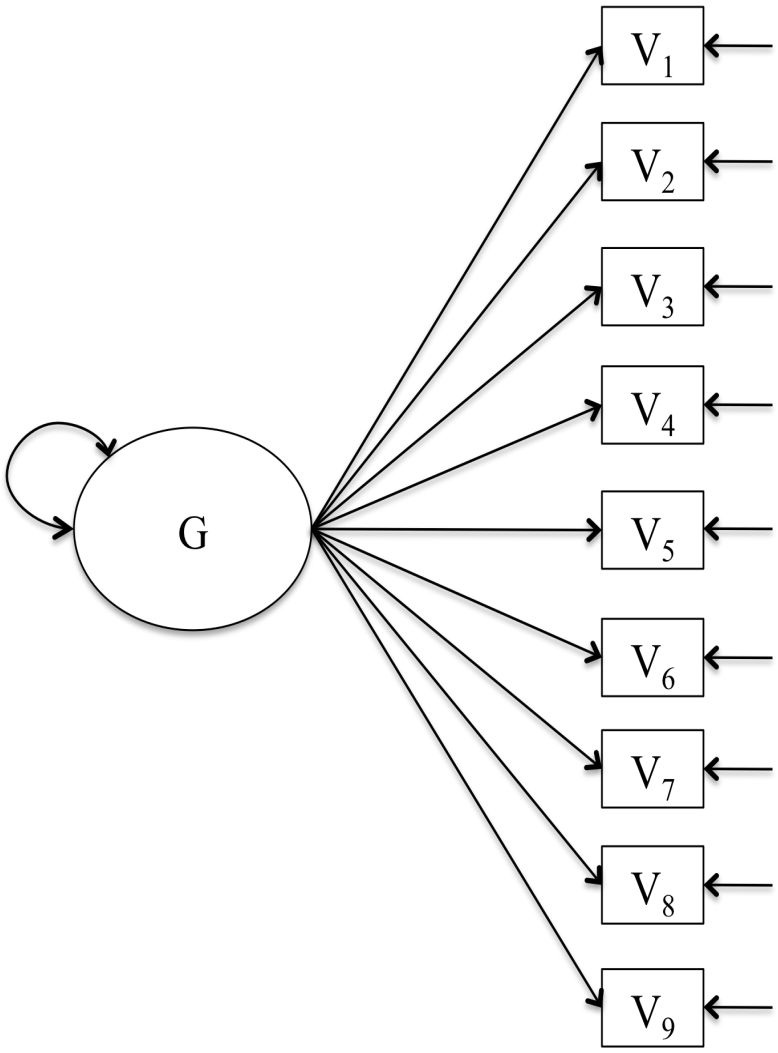

Figure 1. A one-factor model. Observed variables are in rectangles (V1-V9). A general factor, G, is in an ellipse. The double-sided arrow to and from the factor represents a factor variance. The one sided arrows leading to observed variables represent error variances in the observations. The arrows leading from the factor to the observed variables are factor loadings.

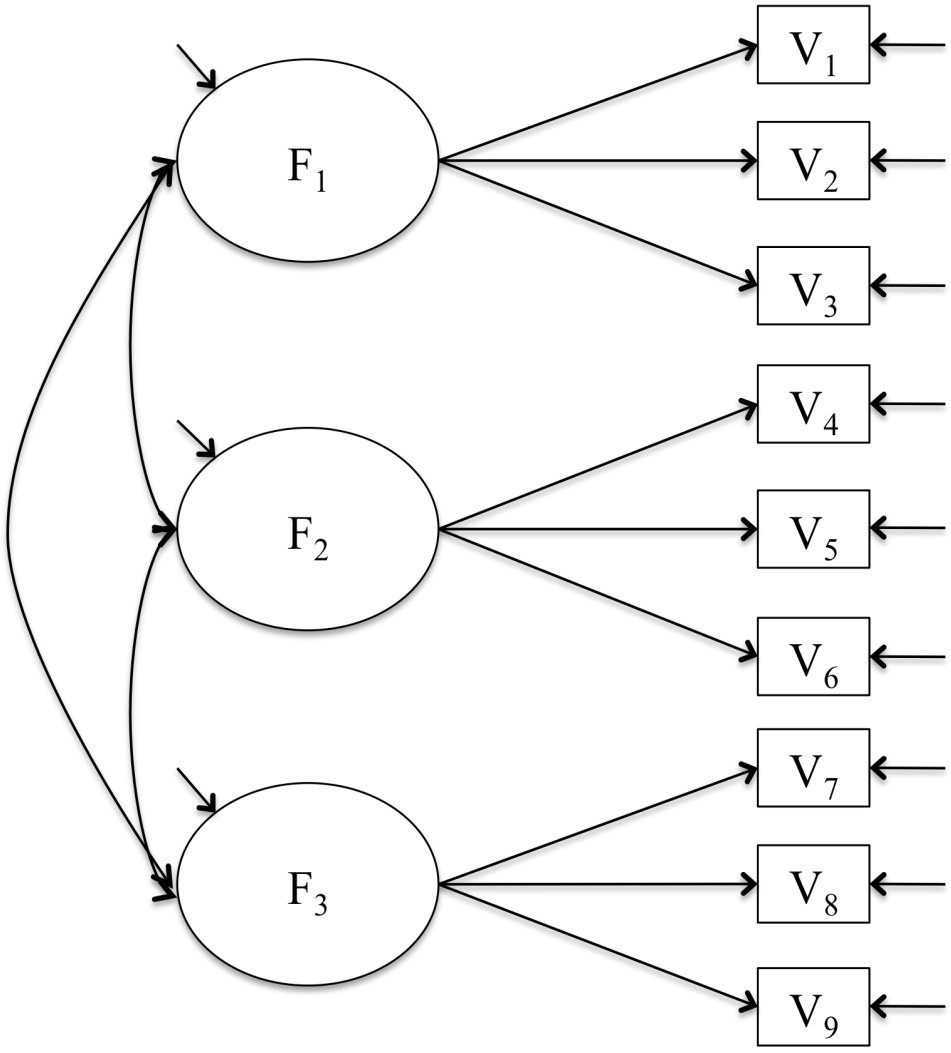

Figure 2. A three-factor model. Each specific factor (F1-F3) only loads on three observed variables. Double-sided arrows between factors represent correlations between factors.

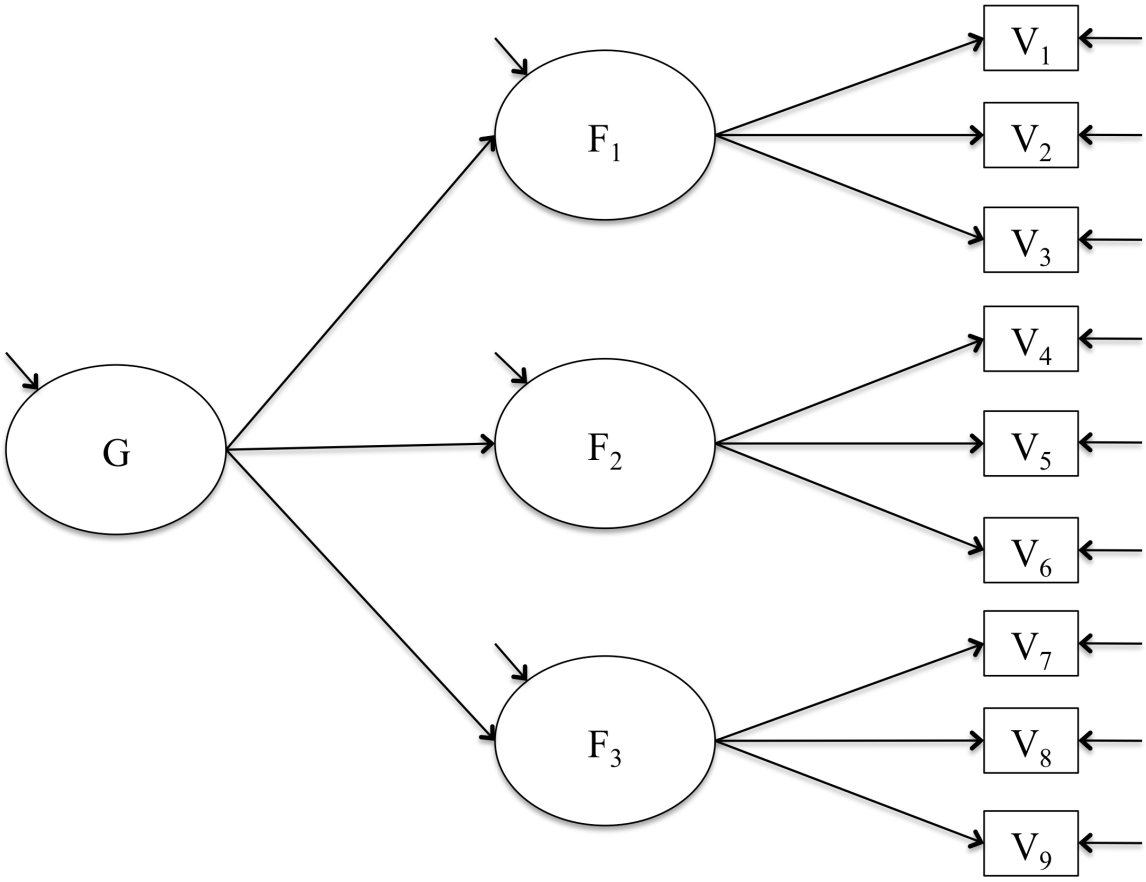

Figure 3. A higher-order factor model. The general factor, G, has factor loadings on the specific factors, F1-3, which have factor loadings on a subset of observed variables. Not pictured: correlations between F1-F3 are usually estimated.

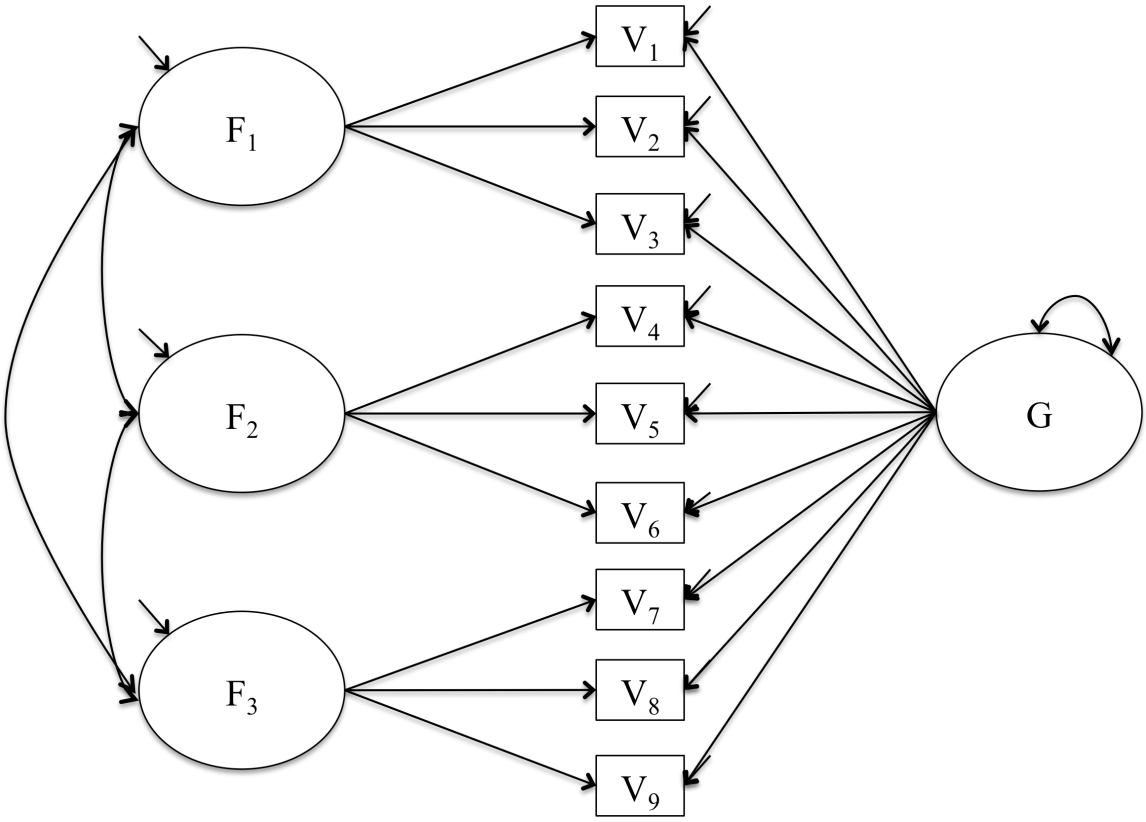

Figure 4. A bifactor model. The general factor, G, is not correlated with the specific factors F1-F3.

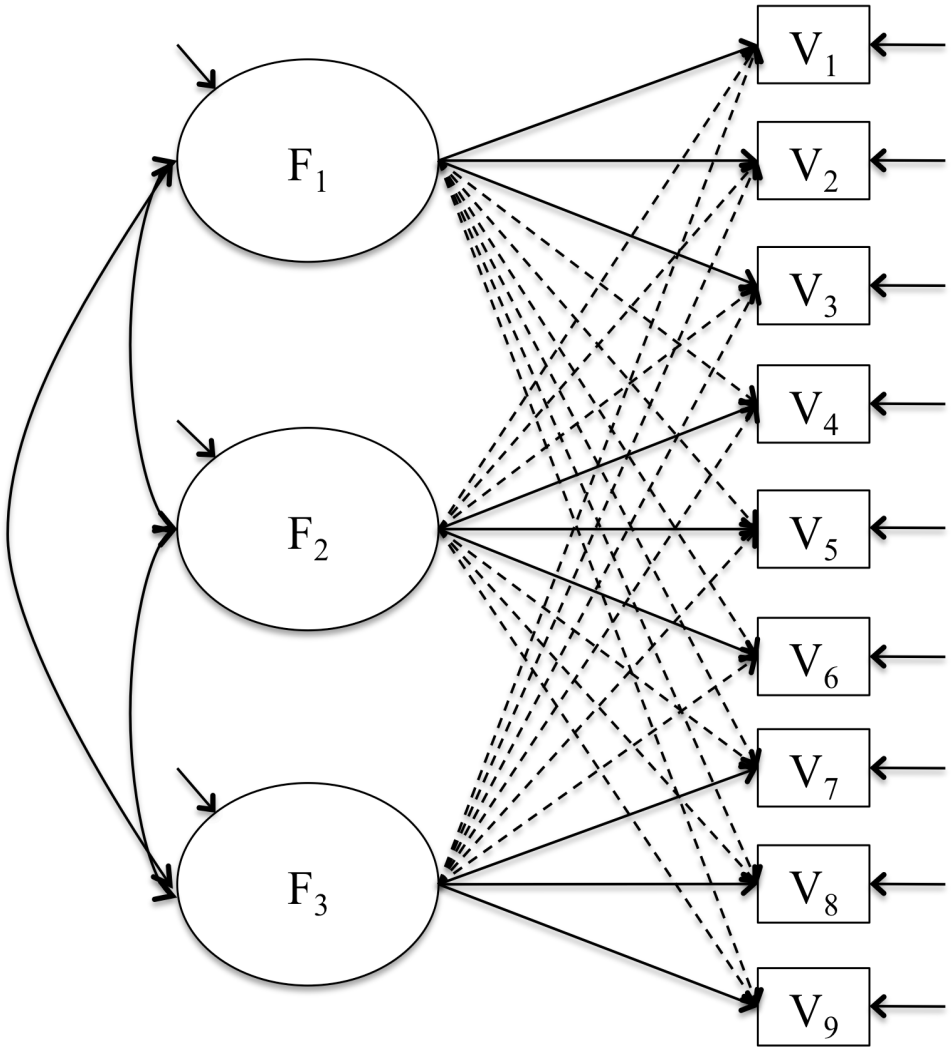

Figure 5. An Exploratory Structural Equation Model. Dashed lines represent small but nonzero estimated factor loadings.

Chapter 2: EFA and MLEFA of psychological symptom measures over time

The basic research question addressed here is one of differences between trait and state psychological symptoms. As has long been acknowledged (e.g., Cattell & Bartlett, 1971), simple R-technique factor analysis is insufficient to determine any potential differences between constant differences between people and those features of psychology that change over time. Cattell's dR technique is one method, which could inform this distinction. However, dR-technique is limited to two timepoints and is strictly an analysis of difference scores (Cattell (1966a) accounted for the possibility that more timepoints could be analyzed as a similar way, but noted that the computational and data requirements of this were not attainable at the time). The computational demands imposed by these complex data analyses can now be performed with relative ease and the type of data required for such analyses is increasingly common: in some contexts, becoming the norm rather than the exception. That is, there is no reason why we should continue to ignore the presence of multiple timepoints when examining the structure of psychological measures and variables, and MLFA may be an appropriate and beneficial method of doing so. First, let us discuss the primary conceptual questions that are unanswered at present.

As discussed above, FA is most frequently conducted on R-technique data: one observation per person, with many people included in a sample. This leaves wholly unanswered the question of what parts of the observed scores are due to trait variance and what parts are due to state variance. This has important implications for the study of psychological symptoms and psychopathology, along with treatments for mental health problems. The factors determining true and stable differences between people – traits – are likely more similar to personality than to "symptoms" in the medical sense. They are probably descriptive of an individual's average self or tendencies across situations. It is possible that they represent meaningful underlying

psychological processes, as in intelligence and personality traits, for which comparisons to other people are most appropriate. These are not symptoms except in a long-term sense, as is used to define personality disorder and other chronic illnesses. In contrast, the factors that are defined by variables changing together are temporally unstable and are only defined by differences within people. These are akin to symptoms that indicate the presence of a new or unusual systemic deviation for a person, such as sneezing or an increasingly depressed mood: factors that change over time.

It is important to emphasize that between-person factors and within-person factors are not necessarily the same. There may be the same set of factors that determine differences between people and within, but not necessarily. In terms of a broad measure of psychopathology such as the Counseling Center Assessment of Psychological Symptoms-34 (Locke et al., 2012), inter-level measurement equivalency would mean that some people are more depressed than other people, and that individuals' level of depression changes as a function of the exact same items on the scale. In contrast, a lack of equivalence between levels might appear as a clear depression factor between people (showing that some people are more depressed than others, on average) but a very broad and simpler within-person factor structure, showing that though some people are more depressed than others, when their depression lifts, so does their level of anxiety, eating concerns, and alcohol use. The opposite might also be true for other people, along with any number of other unequal factor models.

In treatment, there are two non-mutually exclusive sets of interpretations for these factors. The first set of interpretations may be called predictive interpretations. In this scheme, the between-person factors may be best conceptualized as moderators rather than treatment outcome variables, particularly in short-term treatments. Such variables might include

personality traits, cultural differences, and intelligence/reading level. Though these can be extremely important during a treatment, by definition these factors should only change over very long time scales, if at all. In contrast, variables that do seem to change over a given time scale may be more appropriate targets for intervention. That is, the within-person factors defined by symptom change may be viewed as the best outcome to monitor, focused on in case formulation, and targeted by intervention. This would be consistent with the idea that within-person symptom factors may be understood as the temporary symptoms that should be removed by the end of treatment, and if treatment has an effect on between-person factors, it would have to occur at a different time scale than the observation period. A similar observation has been proposed in the phase model of psychotherapy change (Howard, Lueger, Maling, and Martinovich, 1993), which suggests that some factors would be altered in psychological treatments more quickly (e.g., symptoms) than others (e.g., personality).

An alternative interpretation of these factors may be considered treatment descriptive, in that they would only describe the effects and limits of treatment. If different sets of factors are developed based on variation between person and within persons who are in treatment, the between-person factors can be described as those psychological constructs which were not affected by treatment: they did not change over the course of treatment. These could then be targeted in subsequent treatment development, since they can be considered treatment resistant. In contrast, a within-person factor developed on people during treatment could be considered a factor that was influenced by treatment. That is, the treatment provided actually seemed to produce some systematic change within people on a given set of factors, and can therefore be taken to be the targets of the same treatments in the future.

Multilevel factor analysis is a natural fit for determining the structure of within-person and between-person variability in factor structure of psychological symptoms (whether our concept of "symptom" will be defined between or within people). If the between- and within-person structures are identical, little conceptual change in interpretation of data will be required. If, however, the levels differ in terms of the factor structure, this will have important conclusions regarding optimal treatment monitoring and outcome assessment. Given the issue of general and specific factor solutions discussed above, for instance, it could provide justification for use of general, specific, or both factor scores in analyses and treatment outcome monitoring. By factoring out between-person variance from within-person variance, more accurate estimation of within-person factors are likely. This should allow for more specific feedback systems and more accurate assessment of therapy's effectiveness.

The presence of other methods, which have already proven their utility in studying these processes (particularly p-technique and DFA models), do not diminish the potential utility of MLFA models, and these methods may be complementary and/or integrated in the future. One potential limitation of p-technique and DFA models is that, especially for complex models with many factors and/or structural relationships, it may be necessary to have more observations per person than may be feasible to collect in certain applied situations. For example, Molenaar and Nesselroade (2009) have shown good parameter recovery of p-technique with 50 observations. While this is fewer than the number of observations typically recommended in traditional FA (the same authors also offer that 300 observations is more "respectable"), 50 observations of a given person are rarely accomplishable without EMA data collection, or at least daily recordings. Though this may be the ideal (and possibly more common future psychological research), it is not available in all contexts at present.

One context of particular interest here is weekly psychotherapy sessions. It is becoming increasingly common in this setting for clinicians to administer and monitor an outcome measures at every session, for reasons including participation in clinic-based research projects, quantitative program evaluations, and practice-based research organizations (Castonguay, Barkham, Lutz, & McAleavey, 2013). There is a building fund of data collected in this manner, with large numbers of psychotherapy clients but often in relatively short-term treatment settings comprising between 5-20 sessions. Thus, this data likely do not suit themselves to fully person-specific data analyses, because the number of observations per person is relatively small. However, these data are well-suited to MLM and multilevel analyses generally, because of the use of empirical Bayes estimates: the presence of many individuals in a data set provides the possibility that estimates of each may be more reliable (again, making the usual assumption that individuals in the sample follow a common process and distribution). In this context, p-technique may not be feasible, and dR-technique as used by Cattell does not maximally make use of more than two observations per person. But a MLFA may be perfectly suited to this data structure, and may be able to address similar conceptual questions to these techniques.

Method

Participants

Nonclinical. The nonclinical sample for this study was comprised of 1223 students recruited from the Department of Psychology subject pool at a large Mid-Atlantic university over seven consecutive semesters from Spring 2011 – Spring 2014. Each semester, between 125 and 248 participants were recruited, and received course credits. The subjects ranged in age from 18-53 (M = 19.2, SD = 2.09). The majority (899) were female, 316 were male, and one participant identified as transgender. Most students (906) were White, 150 were Asian/Asian-American, 67

were African-American/Black, 45 were Hispanic/Latino/a, 32 identified as Multi-racial, and 12 chose to self-identify race/ethnicity. Roughly half (603) of the participants were Freshmen, 350 were Sophomores, 180 were Juniors, and 73 were Seniors, with 10 participants indicating a different status. At the start of each semester, participants were asked whether they were in some form of counseling or psychotherapy, and 102 indicated that they were. Similarly, 108 participants indicated that they had a current prescription for a psychiatric medication (66 participants reported both psychological and pharmacologic treatment at the start of the study). All participants completed informed consent procedures approved by the IRB.

Clinical. Participants in CCMH data collection at member UCCs between the fall 2010 and Spring 2012 semesters were included as possible participants. However, only 757 clients, of the total 95,109 clients seen during that time had 10 or more CCAPS administrations. Of the 757, 582 also completed some demographic information. Missing data in this sample is likely due to the policies and procedures of each individual UCC, since each UCC can administer the CCAPS and demographic questionnaire on their own schedule. From the 582 providing demographics, the age ranged from 18 to 48 (M = 23.1, SD = 4.7). The majority (337) was female, 159 were male, and 3 were transgender. Most participants (402) were White, 29 were Asian American/Asian, 22 were African-American/Black, 21 were Hispanic/Latino/a, 17 were Multi-racial, and 10 chose to self-identify race/ethnicity. Sixty-eight (68) participants were Freshmen, 103 were Sophomores, 122 were Juniors, 158 were Seniors, and 120 were graduate students, with 7 participants indicating any other status.

Measures

CCAPS (Locke et al., 2011; 2012). The CCAPS instruments were designed to be multidimensional assessments of several common psychological symptoms treated in college

counseling centers. The CCAPS-34 has 34 items, reduced from the 62 of the CCAPS-62 through classical test theory and item-response theory methods (Locke et al., 2012). The 34 items are scored using seven factor-analytically derived subscales: Depression, Generalized Anxiety, Social Anxiety, Hostility, Alcohol Use, Eating Concerns, and Academic Distress. In addition, the Distress Index (DI) is scored in clinical practice. The DI is a general measure of distress, developed through bifactor modeling (CCMH, 2012; Nordberg et al., under review), and is scored by averaging 20 items from four subscales of the CCAPS-34. The seven-factor structure has received good support in a large (N=19,247) single-time point sample, each subscale appears to have good convergent validity with other measures of the construct it was designed to assess, and the DI appears to correlate highly with another measure of general distress (the OQ-45 total score; Nordberg et al., under review). In both samples, all administrations of CCAPS-62 were rescored as the CCAPS-34.

Standardized Data Set (SDS). The SDS is a variable-length questionnaire designed to facilitate and standardized intake procedures at counseling centers. It assesses basic demographic and previous service utilization history. Each counseling center may administer selected items from the SDS.

Procedure

Nonclinical. In the nonclinical sample, participants were recruited to participate in 14 weekly assessments, which encompassed most of each 16-week academic semester. Data were collected over seven semesters, each with unique participants. Each week, participants completed a CCAPS (the CCAPS-62 at week 1, the CCAPS-34 thereafter). They also completed sample SDS items for demographic information at week 1.

Clinical. The clinical data were derived from ongoing routine data collection from CCMH member UCCs. Specifically, data from the 2010-2011 and 2011-2012 academic year were aggregated and participants were included only if they completed 10 administrations of the CCAPS during this period. No exclusion criteria were applied controlling the type or amount of treatment dose provided to any client. Since each participating center in CCMH administers the CCAPS according to local clinical policies, there is likely substantial variability between participants in this sample with regard to timing and treatment. However, no restrictions were applied in order to maintain the most representative sample of participants. Most centers administer the CCAPS prior to sessions of treatment, but vary in whether they administer the CCAPS prior to each session or only some subset of sessions.

Data Analysis

Three single-level EFA models were compared for each sample prior to the MLEFA model. Model 1 is a standard (Cattel's R-technique) EFA, conducted only on the first observation of each person in each data set. This will be closest to a replication of a standard FA model with R-technique data, where each individual is only observed at a single timepoint. Model 2 is a modified R-technique EFA in which each participant's average scores for each item, across all time points, will be used as the data. This serves as a good comparison for the between-person level since both are based on time-invariant data for each participant. In Model 3, all available data for each participant will be used without accounting for the nesting of observations within individuals. Finally, in Model 4, an MLEFA was conducted, explicitly modeling both between- and within-person symptom structure. A longitudinal factor model was also attempted in the nonclinical data, in which each time point's observed indicators were modeled separately with lagged relationships between the times; however, this proved

incalculable with the large number of observed indicators, even with strict factorial invariance across times (476 observed items).

Estimation for all models was completed in Mplus v. 7.11 (Muthén & Muthén, 2012). This software permits computation of arbitrary factor models with categorical indicators. The preferred estimation method with ordered categorical indicators (as in the CCAPS' Likert-type item responses) is weighted least squares with robust corrections for mean and variance (WLSMV) as opposed to traditional maximum likelihood (ML). WLSMV estimation was used for Model 1, Model 3, and Model 4. However, in Model 2, the indicators are continuous item means rather than categorical, so ML was used. Because of the differences in estimation and data, direct model comparisons with likelihood-ratio or incremental and global fit indices are not appropriate. Instead, each model was interpreted independently. First, the optimal number of factors was selected for each model using an a priori set of fit statistics values, including root mean squared error (RMSEA) < .05, Confirmatory Fit Index (CFI) > .95, Tucker-Lewis Index (TLI) > .95, and standardized root mean square error (SRMR) < .08. For Model 4, the same fit criteria were used, except that the SRMR is calculated for both the Between and Within levels separately, so the preferred model would have both SRMRs < .08. Within each Model type, the most parsimonious factor solution that satisfied these criteria was selected as the preferred model. If a given preferred model was not interpretable (e.g., because it contained too many cross-loadings and/or at least one factor with fewer than 3 unique indicators), a more parsimonious but less well-fitting model was preferred instead.

Results

Model selection

Table 1 contains fit statistics for Models 1, 2, and 3 in the Nonclinical sample, and Table 2 has fit statistics for the same models in the Clinical sample. In all six cases, the best fit was found with 8 factors, but these solutions were not interpretable. Across these solutions, the eighth factor was measured by few (2 or 3) indicators, which were rarely unique to that factor. For instance, in Model 1 with the Nonclinical sample, the eighth factor was indicated only by items 15 ("I have spells of terror or panic") and 7 ("I am anxious that I might have a panic attack while in public"). While these items clearly represent a single construct – panic attacks and possibly Panic Disorder – two items is too few to validate a subscale, and both items also more strongly loaded onto the Generalized Anxiety factor. Though allowing the residuals of these items to correlate may result in an optimal factor model for this sample, for the purposes of this study the simpler seven-factor structure was preferred.

Table 1. Model Selection for Nonclinical data

Number of factors	RMSEA	CFI	TLI	SRMR	# Free Parameters
Model 1: First time point only, *N* = 1,215					
1	0.136	0.703	0.684	0.127	34
2	0.1	0.85	0.829	**0.079**	67
3	0.088	0.892	0.868	**0.063**	99
4	0.079	0.918	0.894	**0.052**	130
5	0.07	0.94	0.916	**0.043**	160
6	0.062	**0.956**	0.933	**0.036**	189
7*	0.052	**0.972**	**0.954**	**0.027**	217
8	**0.046**	**0.98**	**0.964**	**0.023**	244
Model 2: Average of all observations by person, *N* = 1,215					
1	0.188	0.513	0.482	0.114	102
2	0.172	0.618	0.567	0.086	135
3	0.159	0.693	0.627	**0.062**	167
4	0.14	0.78	0.713	**0.058**	198
5	0.126	0.834	0.768	**0.047**	228
6	0.115	0.872	0.806	**0.039**	257
7*	0.1	0.909	0.852	**0.029**	285
8	0.085	0.94	0.894	**0.022**	312
Model 3: All data, ignoring clustering, treating *N* = 14,376					
1	0.142	0.818	0.806	0.154	34
2	0.117	0.884	0.868	0.099	67
3	0.1	0.921	0.904	**0.063**	99
4	0.091	0.939	0.921	**0.055**	130
5	0.082	**0.954**	0.936	**0.043**	160
6	0.072	**0.967**	**0.95**	**0.032**	189
7*	0.06	**0.979**	**0.965**	**0.023**	217
8	**0.049**	**0.987**	**0.977**	**0.018**	244

Note. *: preferred model. Boldface text indicates satisfied fit criteria. In all models, the solution with 8 factors resulted in uninterpretable solution. Model 2 did not result in a model meeting fit criteria, while Models 1 and 3 both resulted in 7-factor solutions which satisfied all a priori criteria except for RMSEA < .05.

Table 2. Model Selection for Clinical data

Number of factors	RMSEA	CFI	TLI	SRMR	# Free Parameters
Model 1: First time point only, $N = 757$					
1	0.17	0.528	0.497	0.183	34
2	0.14	0.697	0.656	0.139	67
3	0.123	0.784	0.738	0.107	99
4	0.11	0.839	0.79	0.089	130
5	0.096	0.885	0.838	**0.068**	160
6	0.082	0.923	0.884	**0.046**	189
7*	0.065	**0.955**	0.926	**0.034**	217
8	0.052	**0.973**	**0.953**	**0.025**	244
Model 2: Average of all observations by person, $N = 757$					
1	0.195	0.368	0.328	0.144	102
2	0.184	0.474	0.403	0.12	135
3	0.169	0.583	0.494	0.098	167
4	0.146	0.711	0.624	**0.079**	198
5	0.127	0.796	0.714	**0.06**	228
6	0.113	0.85	0.773	**0.046**	257
7*	0.092	0.908	0.851	**0.032**	285
8	0.08	0.936	0.887	**0.022**	312
Model 3: All data, ignoring clustering, treating $N = 10{,}852$					
1	0.185	0.642	0.619	0.206	34
2	0.155	0.765	0.734	0.148	67
3	0.135	0.833	0.798	0.111	99
4	0.12	0.877	0.839	0.088	130
5	0.106	0.911	0.876	**0.069**	160
6	0.088	0.943	0.914	**0.043**	189
7*	0.068	**0.968**	0.948	**0.027**	217
8	0.057	**0.98**	**0.964**	**0.021**	244

Note. *: preferred model. Boldface text indicates satisfied fit criteria. In all models, including 8 factors resulted in uninterpretable solution. No models resulted in meeting all fit criteria.

Model 4 solutions differed between Clinical and Nonclinical samples. In the Nonclinical sample, no acceptable solution met all fit criteria (see Table 3). Though several models with 8 within-level factors met all fit statistics, these solutions were not interpretable due to having only one item load above .4 on the eighth factor (item 8, "I feel confident that I can succeed academically"). Because of this, 7 within-level factors were selected, and the most parsimonious of these that came close to meeting all fit criteria had 2 between-level factors. In this model, the TLI was the only index not to reach its criterion, but was very close to the cut point of .95 (notably close in all models with 7 within-level factors and 2-5 between-level factors). In contrast, the Clinical sample solutions did include several interpretable models reaching all fit criteria, as shown in Table 4. The most parsimonious model to meet all criteria included seven within factors and five between factors. Particular attention to the between-level SRMR values in Tables 3 and 4 shows that in the Clinical sample, between-level variability was not sufficiently accounted for by fewer than five factors, whereas two between-level factors sufficed to reduce this measure in the Nonclinical sample to an acceptably small level. It is worth noting as well that though they are less parsimonious anyway, the models with 8 within-level factors were also not interpretable in this clinical sample for the same reasons found in Models 1-3.

Table 3. Model Fit indices for Model 4, Nonclinical sample

Number of factors							
Within	Between	RMSEA	CFI	TLI	SRMR (Within)	SRMR (Between)	# Free Parameters
1	1	0.054	0.661	0.639	0.085	0.117	238
2	1	**0.043**	0.793	0.773	**0.063**	0.117	271
3	1	**0.037**	0.848	0.827	**0.051**	0.117	303
4	1	**0.034**	0.879	0.858	**0.044**	0.117	334
5	1	**0.031**	0.904	0.884	**0.039**	0.117	364
6	1	**0.027**	0.927	0.909	**0.031**	0.117	393
7	1	**0.023**	0.947	0.932	**0.023**	0.117	421
8	1	**0.022**	**0.955**	0.94	**0.02**	0.117	448
1	2	0.055	0.659	0.625	0.085	**0.075**	271
2	2	**0.043**	0.795	0.767	**0.063**	**0.075**	304
3	2	**0.037**	0.852	0.826	**0.051**	**0.075**	336
4	2	**0.034**	0.884	0.86	**0.044**	**0.075**	367
5	2	**0.03**	0.911	0.889	**0.039**	**0.075**	397
6	2	**0.026**	0.936	0.917	**0.031**	**0.075**	426
7*	2*	**0.021**	**0.958**	0.944	**0.023**	**0.075**	454
8	2	**0.019**	**0.967**	**0.954**	**0.02**	**0.075**	481
1	3	0.056	0.66	0.614	0.085	**0.057**	303
2	3	**0.044**	0.796	0.761	**0.063**	**0.057**	336
3	3	**0.038**	0.854	0.822	**0.051**	**0.057**	368
4	3	**0.034**	0.886	0.857	**0.044**	**0.057**	399
5	3	**0.03**	0.914	0.888	**0.039**	**0.057**	429
6	3	**0.026**	0.939	0.918	**0.031**	**0.057**	458
7	3	**0.021**	**0.962**	0.947	**0.023**	**0.057**	486
8	3	**0.018**	**0.971**	**0.958**	**0.02**	**0.057**	513
1	4	0.057	0.661	0.603	0.085	**0.046**	334
2	4	**0.044**	0.798	0.755	**0.063**	**0.046**	367
3	4	**0.038**	0.855	0.818	**0.051**	**0.046**	399
4	4	**0.034**	0.888	0.854	**0.044**	**0.046**	430
5	4	**0.03**	0.915	0.886	**0.039**	**0.046**	460
6	4	**0.026**	0.941	0.918	**0.031**	**0.046**	489
7	4	**0.02**	**0.964**	0.948	**0.023**	**0.046**	517
8	4	**0.018**	**0.973**	**0.96**	**0.02**	**0.046**	544
1	5	0.057	0.665	0.595	0.085	**0.037**	364
2	5	**0.045**	0.8	0.75	**0.063**	**0.037**	397
3	5	**0.039**	0.857	0.814	**0.051**	**0.037**	429
4	5	**0.035**	0.89	0.852	**0.044**	**0.037**	460
5	5	**0.031**	0.917	0.884	**0.039**	**0.037**	490

6	5	**0.026**	0.943	0.917	**0.031**	**0.037**	519
7	5	**0.02**	**0.966**	0.949	**0.023**	**0.037**	547
8	5	**0.018**	**0.975**	**0.961**	**0.02**	**0.037**	574
1	6	0.058	0.668	0.586	0.085	**0.029**	393
2	6	**0.045**	0.802	0.744	**0.063**	**0.029**	426
3	6	**0.039**	0.859	0.81	**0.051**	**0.029**	458
4	6	**0.035**	0.891	0.848	**0.044**	**0.029**	489
5	6	**0.031**	0.918	0.882	**0.039**	**0.029**	519
6	6	**0.026**	0.944	0.916	**0.031**	**0.029**	548
7	6	**0.02**	**0.967**	0.948	**0.023**	**0.029**	576
8	6	**0.018**	**0.976**	**0.961**	**0.02**	**0.029**	603
1	7	0.059	0.669	0.574	0.085	**0.021**	421
2	7	**0.046**	0.803	0.736	**0.063**	**0.021**	454
3	7	**0.04**	0.859	0.804	**0.051**	**0.021**	486
4	7	**0.036**	0.891	0.843	**0.044**	**0.021**	517
5	7	**0.031**	0.919	0.877	**0.039**	**0.021**	547
6	7	**0.027**	0.944	0.913	**0.031**	**0.021**	576
7	7	**0.021**	**0.968**	0.947	**0.023**	**0.021**	604
8	7	**0.018**	**0.977**	**0.961**	**0.02**	**0.021**	631

Note. *N*=1,078, average number of observations per person = 13.3 (all 10 or more). Models with 8 between factors did not converge. *: preferred model. Boldface indicates meeting the criteria for good fit.

Table 4. Model Fit indices for MLEFA on Clinical Sample

Number of factors							
Within	Between	RMSEA	CFI	TLI	SRMR (Within)	SRMR (Between)	# Free Parameters
1	1	0.062	0.691	0.671	0.104	0.161	238
2	1	**0.049**	0.807	0.788	0.091	0.161	271
3	1	**0.042**	0.863	0.844	**0.076**	0.161	303
4	1	**0.036**	0.903	0.887	**0.067**	0.161	334
5	1	**0.032**	0.929	0.914	**0.057**	0.161	364
6	1	**0.028**	0.947	0.934	**0.037**	0.161	393
7	1	**0.023**	**0.965**	**0.955**	**0.023**	0.161	421
8	1	**0.022**	**0.969**	**0.959**	**0.02**	0.161	448
1	2	0.064	0.671	0.639	0.104	0.121	271
2	2	0.052	0.797	0.769	0.091	0.121	304
3	2	**0.044**	0.856	0.83	**0.076**	0.121	336
4	2	**0.037**	0.901	0.879	**0.067**	0.121	367
5	2	**0.032**	0.928	0.91	**0.057**	0.121	397
6	2	**0.028**	0.949	0.934	**0.037**	0.121	426
7	2	**0.021**	**0.97**	**0.96**	**0.023**	0.121	454
8	2	**0.02**	**0.975**	**0.965**	**0.02**	0.121	481
1	3	0.066	0.668	0.624	0.104	0.108	303
2	3	0.053	0.796	0.76	0.091	0.108	336
3	3	**0.045**	0.855	0.824	**0.076**	0.108	368
4	3	**0.038**	0.901	0.876	**0.067**	0.108	399
5	3	**0.032**	0.93	0.909	**0.057**	0.108	429
6	3	**0.027**	**0.951**	0.934	**0.037**	0.108	458
7	3	**0.021**	**0.974**	**0.963**	**0.023**	0.108	486
8	3	**0.019**	**0.978**	**0.968**	**0.02**	0.108	513
1	4	0.067	0.663	0.605	0.107	0.088	334
2	4	0.054	0.793	0.748	0.091	0.088	367
3	4	**0.046**	0.853	0.816	**0.076**	0.088	399
4	4	**0.039**	0.901	0.871	**0.067**	0.088	430
5	4	**0.033**	0.93	0.906	**0.057**	0.088	460
6	4	**0.028**	**0.952**	0.933	**0.037**	0.088	489
7	4	**0.02**	**0.976**	**0.965**	**0.023**	0.088	517
8	4	**0.018**	**0.981**	**0.971**	**0.02**	0.088	544
1	5	0.069	0.66	0.588	0.104	**0.056**	364
2	5	0.055	0.791	0.738	0.091	**0.056**	397
3	5	**0.047**	0.852	0.807	**0.076**	**0.056**	429
4	5	**0.039**	0.9	0.865	**0.067**	**0.056**	460
5	5	**0.033**	0.93	0.903	**0.057**	**0.056**	490

6	5	**0.028**	**0.953**	0.932	**0.037**	**0.056**	519
7*	5*	**0.02**	**0.977**	**0.966**	**0.023**	**0.056**	**547**
8	5	**0.018**	**0.982**	**0.973**	**0.02**	**0.056**	**574**
1	6	0.07	0.662	0.578	0.104	**0.038**	393
2	6	0.056	0.793	0.732	0.091	**0.038**	426
3	6	**0.048**	0.854	0.803	**0.076**	**0.038**	458
4	6	**0.04**	0.902	0.863	**0.067**	**0.038**	489
5	6	**0.034**	0.932	0.902	**0.057**	**0.038**	519
6	6	**0.028**	**0.955**	0.932	**0.037**	**0.038**	548
7	6	**0.019**	**0.98**	**0.969**	**0.023**	**0.038**	**576**
8	6	**0.017**	**0.985**	**0.976**	**0.02**	**0.038**	**603**
1	7	0.07	0.664	0.568	0.104	**0.027**	421
2	7	0.056	0.794	0.725	0.091	**0.027**	454
3	7	**0.048**	0.855	0.798	**0.076**	**0.027**	486
4	7	**0.04**	0.903	0.86	**0.067**	**0.027**	517
5	7	**0.034**	0.933	0.9	**0.057**	**0.027**	547
6	7	**0.028**	**0.956**	0.931	**0.037**	**0.027**	576
7	7	**0.019**	**0.981**	**0.969**	**0.023**	**0.027**	**604**
8	7	**0.016**	**0.986**	**0.977**	**0.02**	**0.027**	**631**
1	8	0.071	0.667	0.557	0.104	**0.019**	448
2	8	0.057	0.796	0.717	0.091	**0.019**	481
3	8	**0.049**	0.856	0.793	**0.076**	**0.019**	513
4	8	**0.041**	0.904	0.856	**0.067**	**0.019**	544
5	8	**0.034**	0.934	0.897	**0.057**	**0.019**	574
6	8	**0.028**	**0.96**	0.934	**0.042**	**0.019**	603
7	8				DNC		
8	8				DNC		

Note. Total sample N: 757. DNC: Did not converge.

As a secondary model-selection and comparison procedure, eigenvalues are compared in Figures 6-9. All four plots demonstrate that a single very significant factor is likely present, with a varying number of additional factors possible in each sample. Two additional trends appear: First, there is a less-pronounced difference in the eigenvalues between within-person and between-person levels than appears to exist in the model fit criteria. That is, the nonclinical sample's within-person scree plot is more similar to the nonclinical sample's between-person scree plot, and the same is true of the Clinical sample. Another trend appears, corroborating part of the analysis of fit indices: the Nonclinical sample scree plot provides little evidence that many factors are necessary to explain variance on either level, while the scree plots from the Clinical sample are considerably more indicative of complex factor structures. This is consistent with the finding that the Clinical sample's between-person preferred model in Model 4 had 5 factors while that of the Nonclinical sample only had 2.

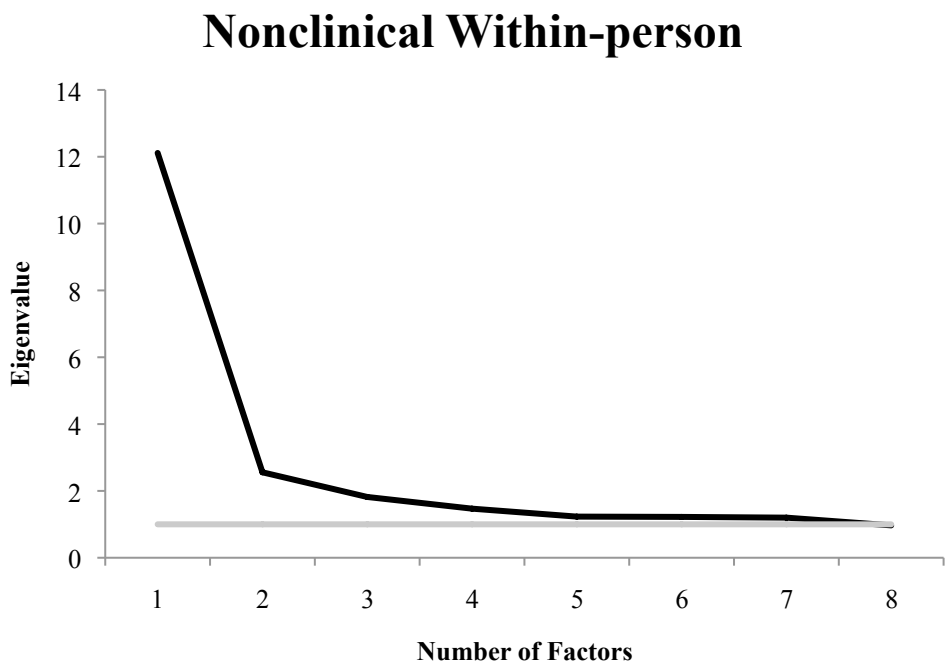

Figure 6. Scree plot for MLEFA, nonclinical sample, within-level. A line for the value of 1 is included, representing the upper limit of acceptable factors.

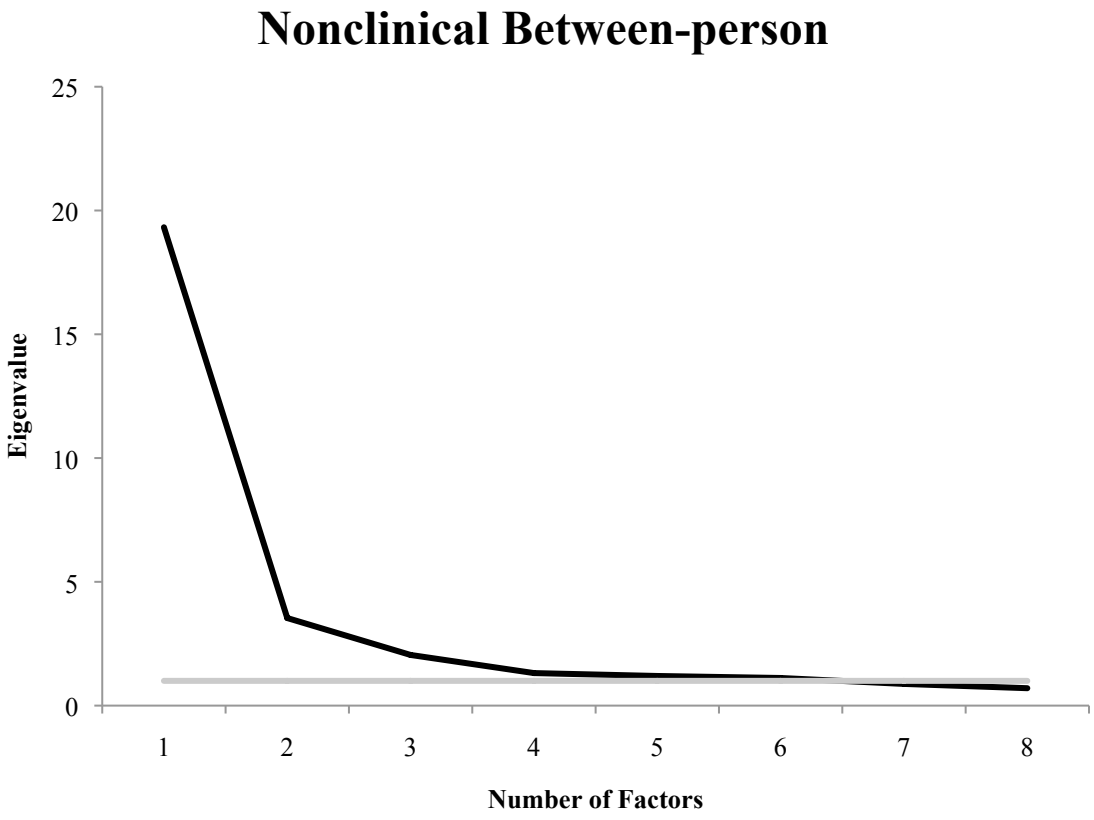

Figure 7. Scree plot for MLEFA, nonclinical sample, between-level. A line for the value of 1 is included, representing the upper limit of acceptable factors.

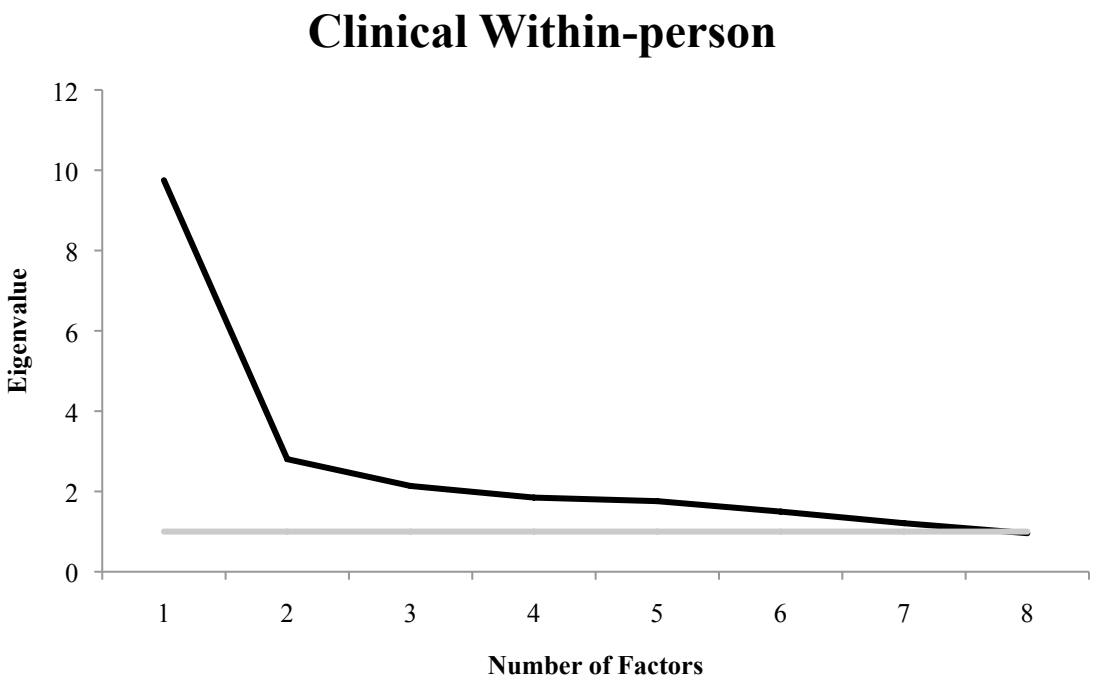

Figure 8. Scree plot for MLEFA, clinical sample, within-level. A line for the value of 1 is included, representing the upper limit of acceptable factors.

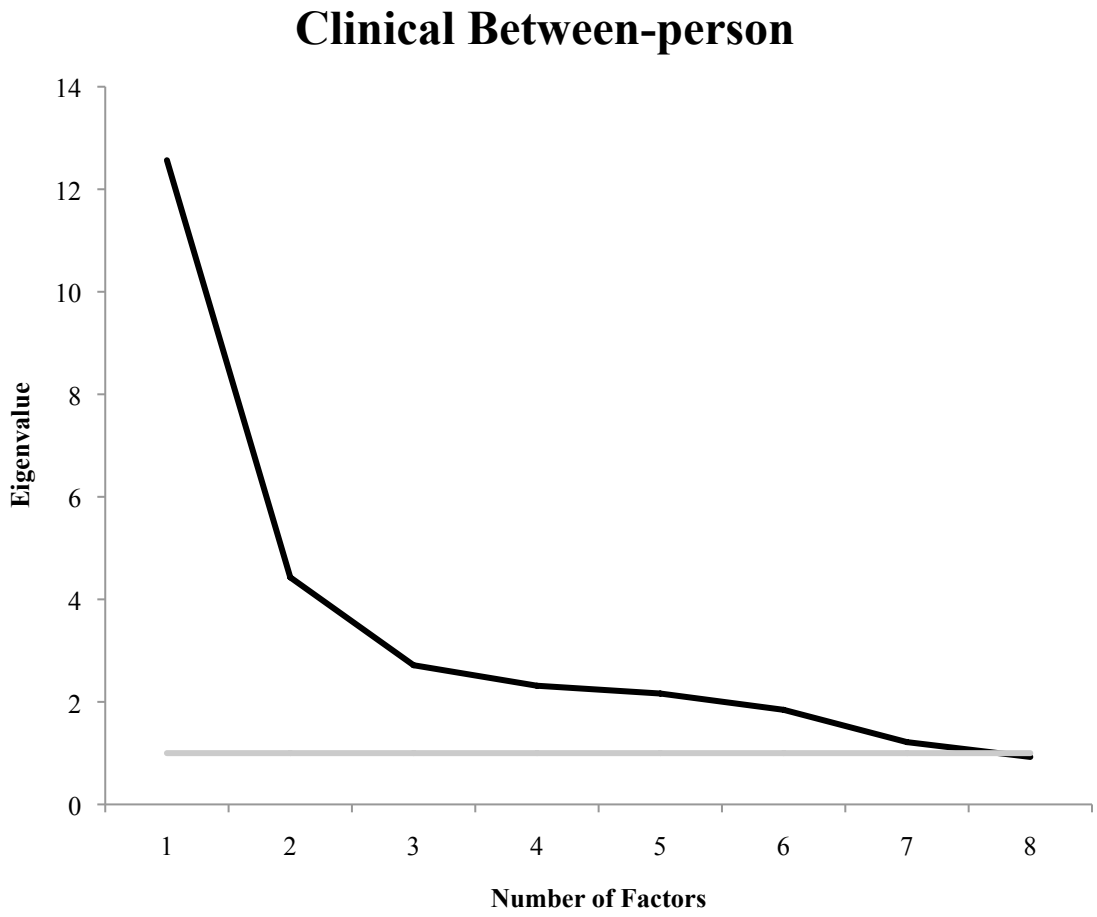

Figure 9. Scree plot for MLEFA, clinical sample, between-level. A line for the value of 1 is included, representing the upper limit of acceptable factors.

Model comparison among preferred models

Table 5 contains a summary of Models 1-3 in both samples, comparing the preferred model for each of these solutions. There is a high degree of agreement among these six factor models and with the reported factor structure of the CCAPS-34. All but three factor loadings that are used to score this instrument in clinical practice (which are based on previous factor analytic studies) were confirmed in all six models. However, there were some items that were inconsistent across models, particularly with regard to the Generalized Anxiety and Depression subscales. Item 17 was found to significantly load on Generalized Anxiety subscale in five of the six models, which could conceivably merely reflect some chance variation. Item 9, however, ("I have sleep difficulties") did not load on any factors in two models, one from the clinical and one from the nonclinical sample. While it is impossible to be sure, this raises suspicion that this item may not reflect anxiety for all college students taking the CCAPS. Item 4 ("I don't enjoy being around people as much as I used to"), which is scored as a Depression item, only loaded on the Depression subscale in two models, both in the Nonclinical sample. In other models it had no significant factor loadings. This is evidence that this item may not be useful as an assessment of depression, especially in clinical samples. Additionally two items which are face-valid measures of anger/hostility and are scored on the Hostility subscale, had significant cross-loadings onto the Depression subscale in five of the six models. These items ("I am afraid I may lose control and act violently," and "I have thoughts of hurting others") both include violence against others, and may reflect an extreme form of depression-induced irritability or a marker of extreme irritability that also impacts mood.

Table 5. Summary of preferred solutions factor loadings, Models 1-3, in Clinical and Nonclinical samples.

Item	Text	Eating Concerns	Generalized Anxiety	Social Anxiety	Depression	Hostility	Alcohol Use	Academic Distress
3	I feel out of control when I eat	ALL						
6	I think about food more than I would like to	ALL						
13	I eat too much	ALL						
2	My heart races for no good reason		ALL					
7	I am anxious that I might have a panic attack while in public		ALL					
9	I have sleep difficulties		1NC, 1C, 2C, 3NC					
10	My thoughts are racing		ALL					
15	I have spells of terror or panic		ALL					
17	I feel tense		1NC, 1C, 2NC, 3NC, 3C					
1	I am shy around others			ALL				
19	I make friends easily			ALL				
22	I am concerned that other people do not like me			ALL				
24	I feel uncomfortable around people I don't know			ALL				
26	I feel self conscious around others			ALL				
4	I don't enjoy being around people as much as I used to				1NC, 2NC			
5	I feel isolated and alone				ALL			
11	I feel worthless				ALL			
12	I feel helpless				ALL			
21	I feel sad all the time				ALL			

25	I have thoughts of ending my life	ALL			
29	I am afraid I may lose control and act violently	1NC, 1C, 2NC, 3NC, 3C, *	ALL		
34	I have thoughts of hurting others	1NC, 1C, 2NC, 3NC, 3C, *	ALL		
18	I have difficulty controlling my temper		ALL		
20	I sometimes feel like breaking or smashing things		ALL		
23	I get angry easily		ALL		
32	I frequently get into arguments		ALL		
14	I drink alcohol frequently			ALL	
16	When I drink alcohol I can't remember what happened			ALL	
27	I drink more than I should			ALL	
31	I have done something I have regretted because of drinking			ALL	
8	I feel confident that I can succeed academically				ALL
28	I am not able to concentrate as well as usual				ALL
30	It's hard to stay motivated for my classes				ALL
33	I am unable to keep up with my schoolwork				ALL

Note. Only models in which the standardized factor loading parameter was > .4 are listed. ALL: The preferred 7-factor solution for each Models 1-3 in both Clinical and Nonclinical samples contained meaningful (> .4) factor loadings; *: parameters which are not part of the established scoring structure of the CCAPS-34; 1C: Model 1, Clinical sample; 1NC: Model 1, Nonclinical sample; 2C: Model 2, Clinical sample; 2NC: Model 2, Nonclinical sample; 3C: Model 3, Clinical sample; 3NC: Model 3, Nonclinical sample.

There is also high agreement between the MLEFA models, though notable differences did emerge between clinical and nonclinical samples. Both MLEFA models were more discrepant from the standard CCAPS-34 scoring as well (based on R-technique FA). Nevertheless, both models converged on seven within-person factors that closely resembled the standard CCAPS-34 factors. Table 6 shows the factor loadings for the nonclinical sample, and the loading pattern clearly indicates that the same seven constructs emerge in this analysis as emerged in Models 1-3. However, only the Eating Concerns, Alcohol Use, and Academic Distress factor loading patterns are identical to the standard CCAPS-34 factors. While most deviations in the other subscales at the within-person level are minor, some are worth noting. First, two items did not load significantly (> .40) on any subscale: item 4 ("I don't enjoy being around people as much as I used to") and item 17 ("I feel tense"). Though each of these items had a factor loading that approached the .4 cut off in its presumed factor loading, they are not included in that factor by the a priori criteria. This reflects the fact that they were less correlated with the other items of those subscales over time. The subscale with the greatest discrepancy between this sample and the standard CCAPS-34 scoring was Depression, which had three changes. One item (item 4, already mentioned) did not load on this factor within-persons, and two other items seemed to be more closely related to depression than their target factors: item 8 ("I feel confident that I can succeed academically") and item 19 ("I make friends easily"). Though the topic of these items appear closely linked to the concepts of academic distress and social anxiety, it is interesting that one item is agentic and the other social, a split between theorized types of depression.

Table 6. Factor loadings of Model 4 (MLEFA) in the Nonclinical Sample.

		Within-person factors							Between-person factors	
Item	Text	Generalized Anxiety	Eating Concerns	Depression	Hostility	Alcohol Use	Social Anxiety	Academic Distress	General Distress	Alcohol Use
1	I am shy around others						**0.524**		**0.817**	
2	My heart races for no good reason	**0.694**							**0.769**	
3	I feel out of control when I eat		**0.829**						**0.620**	
4	I don't enjoy being around people as much as I used to[W]			0.375					**0.954**	
5	I feel isolated and alone			**0.589**					**0.959**	
6	I think about food more than I would like to		**0.783**						**0.574**	
7	I am anxious that I might have a panic attack while in public	**0.501**							**0.679**	
8	I feel confident that I can succeed academically			**0.439**					**0.581**	
9	I have sleep difficulties	**0.481**							**0.605**	
10	My thoughts are racing	**0.724**							**0.750**	
11	I feel worthless			**0.771**					**0.898**	
12	I feel helpless			**0.706**					**0.919**	
13	I eat too much		**0.765**						**0.560**	
14	I drink alcohol frequently					**0.762**				**0.800**
15	I have spells of terror or panic	**0.510**							**0.701**	
16	When I drink alcohol I can't remember what happened					**0.743**				**0.809**
17	I feel tense[W]	0.365							**0.862**	
18	I have difficulty controlling my temper				**0.807**				**0.673**	
19	I make friends easily			**0.405**					**0.734**	
20	I sometimes feel like breaking or smashing things				**0.641**				**0.588**	
21	I feel sad all the time			**0.569**					**0.911**	
22	I am concerned that other people do not like me						**0.510**		**0.880**	
23	I get angry easily				**0.762**				**0.681**	
24	I feel uncomfortable around people I don't know						**0.591**		**0.888**	
25	I have thoughts of ending my life			**0.475**	0.388				**0.665**	

26	I feel self conscious around others			**0.534**		**0.896**	
27	I drink more than I should		**0.824**				**0.861**
28	I am not able to concentrate as well as usual				**0.568**	**0.805**	
29	I am afraid I may lose control and act violently	**0.583**				**0.641**	0.361
30	It's hard to stay motivated for my classes				**0.726**	**0.730**	
31	I have done something I have regretted because of drinking		**0.600**				**0.773**
32	I frequently get into arguments	**0.624**				**0.643**	0.316
33	I am unable to keep up with my schoolwork				**0.678**	**0.728**	
34	I have thoughts of hurting others	**0.582**				**0.573**	0.376

Note. Boldface indicates standardized factor loading greater than 0.4. Cell outline indicates loadings in the scoring system of the CCAPS-34. ^{W}: item did not have any meaningful loadings at the within level. All loadings smaller than .3 are suppressed.

The between-level results, also displayed in Table 6, are notable for their greater deviation from the standard CCAPS-34 structure. The first factor seems to consist of nearly every item on the CCAPS-34, and is therefore termed General Distress. This may reflect overall negative affectivity and/or response bias of individuals to anchor their scores over time at particular levels of the Likert-type scales. The second factor most replicates the Alcohol Use subscale. This presence of this factor as a between-person variable may reflect a dichotomy between students who drink alcohol and those who never do, which would be a between-person difference unlikely to emerge on within-person factors unless a student begins to drink or abstain from drinking alcohol during the period of observation.

In the Clinical sample, the MLEFA results (displayed in Table 7) more closely represented the standard CCAPS scoring at both levels. On the within-person level, five subscales (Eating Concerns, Depression, Hostility, Alcohol Use, and Academic Distress) had identical loading patterns to the CCAPS-34 scoring. The Generalized Anxiety subscale had one item that did not significantly load, though the factor loading was close to .4 (item 9, “I have sleep difficulties”). Item 19, which nominally should load on Social Anxiety (“I make friends easily”) had no significant loadings on the within-person level.

Table 7. Factor loadings of Model 4 (MLEFA) in the Clinical Sample.

		Within-person factors						
Item	Text	Generalized Anxiety	Eating Concerns	Depression	Hostility	Alcohol Use	Social Anxiety	Academic Distress
1	I am shy around others						**0.596**	
2	My heart races for no good reason[B]	**0.677**						
3	I feel out of control when I eat		**0.809**					
4	I don't enjoy being around people as much as I used to			**0.440**				
5	I feel isolated and alone			**0.681**				
6	I think about food more than I would like to		**0.822**					
7	I am anxious that I might have a panic attack while in public[B]	**0.765**						
8	I feel confident that I can succeed academically							**0.582**
9	I have sleep difficulties[W]	0.375						
10	My thoughts are racing	**0.535**						
11	I feel worthless			**0.793**				
12	I feel helpless			**0.748**				
13	I eat too much		**0.812**					
14	I drink alcohol frequently					**0.811**		
15	I have spells of terror or panic[B]	**0.741**						
16	When I drink alcohol I can't remember what happened					**0.742**		
17	I feel tense[B]	**0.404**						
18	I have difficulty controlling my temper				**0.804**			
19	I make friends easily[W]							
20	I sometimes feel like breaking or smashing things				**0.604**			
21	I feel sad all the time			**0.687**				
22	I am concerned that other people do not like me						**0.478**	
23	I get angry easily				**0.828**			
24	I feel uncomfortable around people I don't know						**0.690**	
25	I have thoughts of ending my life			**0.580**				

26	I feel self conscious around others				**0.655**	
27	I drink more than I should			**0.900**		
28	I am not able to concentrate as well as usual					**0.530**
29	I am afraid I may lose control and act violently		**0.565**			
30	It's hard to stay motivated for my classes					**0.775**
31	I have done something I have regretted because of drinking			**0.704**		
32	I frequently get into arguments		**0.602**			
33	I am unable to keep up with my schoolwork					**0.820**
34	I have thoughts of hurting others	0.307	**0.536**			

		Between-person factors				
Item	Text	Eating Concerns	Social Anxiety/ Introversion	Hostility	Distress	Alcohol Use
1	I am shy around others		**0.933**			
2	My heart races for no good reason[B]					
3	I feel out of control when I eat	**0.978**				
4	I don't enjoy being around people as much as I used to		**0.401**		**0.477**	
5	I feel isolated and alone		**0.483**		**0.470**	
6	I think about food more than I would like to	**0.957**				
7	I am anxious that I might have a panic attack while in public[B]			0.313		
8	I feel confident that I can succeed academically				**0.708**	
9	I have sleep difficulties[W]				**0.583**	
10	My thoughts are racing				**0.434**	
11	I feel worthless		0.399		**0.478**	
12	I feel helpless		0.340		**0.578**	
13	I eat too much	**0.953**				
14	I drink alcohol frequently					**0.900**
15	I have spells of terror or panic[B]			0.307		

16	When I drink alcohol I can't remember what happened				**0.941**
17	I feel tense[B]	0.331		0.381	
18	I have difficulty controlling my temper		**0.862**		
19	I make friends easily[W]	**0.687**			
20	I sometimes feel like breaking or smashing things		**0.784**		
21	I feel sad all the time	0.369		**0.576**	
22	I am concerned that other people do not like me	**0.726**			
23	I get angry easily		**0.862**		
24	I feel uncomfortable around people I don't know	**0.912**			
25	I have thoughts of ending my life				
26	I feel self conscious around others	**0.802**			
27	I drink more than I should				**0.950**
28	I am not able to concentrate as well as usual			**0.951**	
29	I am afraid I may lose control and act violently		**0.862**		
30	It’s hard to stay motivated for my classes			**0.889**	
31	I have done something I have regretted because of drinking				**0.899**
32	I frequently get into arguments		**0.656**		
33	I am unable to keep up with my schoolwork			**0.894**	
34	I have thoughts of hurting others		**0.777**		

The five-factor solution for the between-level was a combination of some clearly recognizable factors from the CCAPS-34 (Eating Concerns, Hostility, and Alcohol Use all appeared identical to their within-level solutions), and two factors which were less so. One factor closely resembled Social Anxiety with two additional cross-loading items: item 4 ("I don't enjoy being around people as much as I used to") and item 5 ("I feel isolated and alone"). Each of these items is clearly relevant to social difficulties. Because this factor combines trait-like (time-invariant) complaints about social situations, it may be proper to consider it a partial measure of trait introversion as well as Social Anxiety. The final between-person factor combined all the items from Depression and Academic Distress subscales with two items from the Generalized Anxiety subscale. This conglomeration suggests a less broad instance of the General Distress factor seen in Table 6, in that no single clinical construct captures all of its parts. It is also worth noting that four items did not load on any between-person factors, and that all four of these items focus on physiological anxiety symptoms from the Generalized Anxiety subscale. In this sample, though those items were clearly inter-correlated on the within-person level, they did not covary between persons.

Discussion

This study represents an attempt to discriminate between time-varying and time-invariant features of psychopathology self-report using more than one time point of observation per person. The primary outcome of interest in these models was the structure of factor loadings. Models 1, 2, and 3, which are variations on traditional R-technique factor analysis, all generated factor solutions similar to the established structure of the CCAPS-34. This shows that even using multiple time points per individual does not greatly affect the observed factor structure of this instrument, though doing so violates the assumption of independent observations required of R-

technique analyses. Interestingly, two items demonstrated significant cross-loadings that are not part of the standard structure, yet were meaningful in five of the six models. These two items' content pertains more to hostility and violence than depression, yet indicated both Hostility and Depression in these models. This is a novel finding which requires future study. It is notable that the nonclinical and clinical samples did not produce meaningfully different factor structures in any of these models, meaning that these models are not sensitive to differences between samples.

In contrast, though Model 4 also generally reproduced the standard scoring of the CCAPS-34 at the within-person level, it produced notably divergent solutions at the between-person level with Clinical and Nonclinical samples. Regarding the within-person level, though each sample generally reproduced the standard scoring of the CCAPS-34, some differences did emerge. Some items (i.e., "I have sleep difficulties," "I feel tense," and "I don't enjoy being around people as much as I used to") did not load on within person factors in one of the two samples. While this did not replicate across samples, it is observed that in even in the sample in which these items did pass the .40 cut point for inclusion they were not very large loadings (ranging from .404 - .481). Thus, they should be considered weak indicators of within-person variability even if they are meaningful, and potentially removable from the within-person factor structure. These items likely did not covary with other items within-persons because they all appear to assess constructs that vary at different frequencies than other symptoms. The CCAPS-34 was designed for repeated administration at approximately weekly frequency, which is the frequency used in this study. "I have sleep difficulties" and "I don't enjoy being around people as much as I used to" may be more stable than other symptoms, while "I feel tense" may vary even more frequently, on the order of hours to minutes rather than week-to-week. Note that the fact that these items did not covary within person did not preclude their varying between-person:

the differences between people in the level of these items are completely separate from the variation over time within people.

Neither of the samples reproduced the CCAPS-34 scoring at the between-person level. This suggests that the standard structure involving seven subscales is much more a function of within-person variability than between-person differences. In a sense, this is a good sign for the interpretability of R-technique factor analyses: even though using only one timepoint the most common solution in Models 1-3 and previous research converged on the same model that described differences between persons over time. However, the MLFA also revealed different structure at the between-person level, or rather, two different structures. In the Nonclinical sample, a very simple factor solution emerged involving one Alcohol Use subscale (likely useful in discriminating those individuals who drink alcohol at all from those who do not) and one General Distress subscale, which essentially measured all other symptoms of psychological disturbance. This overall level factor is too broad to be clearly interpreted, and may represent a response bias of the participants to anchor their scores at different points on the Likert-type scale. However, it may also capture the fact that individuals who have some psychological symptoms are vulnerable to many other types of symptoms as well. In contrast, a more complex five-factor solution was required for the between-person level of the clinical sample. The more complex model here began to more closely replicate three subscales of the CCAPS-34: Eating Concerns, Alcohol Use, and Hostility all emerge as between-person factors in this sample, essentially unchanged in the factor loading patterns. Social Anxiety also emerged in a very similar form (with two additional items from the Depression subscale). The largest loading item referred to shyness explicitly, suggesting that this factor captures trait introversion in addition to or instead of strictly symptoms of social anxiety. Finally, the Distress factor combines items from

Depression, Generalized Anxiety, and Academic Distress. Notably in the Clinical sample, none of the items relating to immediate physiological symptoms of anxiety loaded meaningfully on any between-level factors, suggesting that these symptoms did not discriminate between individuals but are better considered markers of within-person distress.

The difference between the Clinical and Nonclinical samples in the structure of their between-level solution was itself interesting. The Clinical sample used in this study is a much more diverse group, representing several different schools and a wider variety of ages, races, and academic standing. In the context of this greater variability between individuals, complex between-person differences may tend to be more important. This is an important caution for factor analysis studies of psychopathology measures: using convenience samples that are not representative of the population of interest may be likely to lead to overly simplified factor structure solutions when primarily examining between-person variation.

This study demonstrates that between-person and within-person variability are not identical. Both samples converged in Model 4 on models with simpler between-person structures than within-person. Overall, this provides evidence that people's symptoms of psychopathology change in more complex ways over time than they differ between one another. Knowing this, we can suggest that psychometric assessments of time-varying constructs should maintain domain-specific subscales, while assessments meant to measure time-invariant traits may need to be broader in scope.

There are several limitations to this study. First, there is a relative lack of objective criteria for model comparison between Models 1-4. It would be preferable if objective, unbiased, and consistent criteria were available to compare these models, leading to the selection of a single optimal model. However, MLFA is an area of active research and many questions

remain about the use of fit indices and other tests. In this study, we focused on interpretation of model fit indices that are available in the multilevel context and make it is possible to compare an MLEFA model to a single-level EFA. With these criteria, it can be said that the best models tested here were the MLFA models from Model 4, based on their meeting the a priori fit index cut points, but this is only mild to moderate support. In addition, these models are completely different in their assumptions and data, and so generating a single "best" model may be misguided. Instead, given the difficulty in interpreting these values conclusively, we conclude that Models 1-3 failed to generate truly acceptable models, while Model 4 seemed to do so, at least in one sample. This leads to questions of what to do with these differing accounts of available data. Future research is also needed to determine whether the fit criteria used in this study are optimal cut points for multilevel SEM, as they are based on single level SEM studies.

In summary, this study demonstrates the feasibility and interpretive value to using MLFA in analyzing multiple timepoints of data with several persons. Though very real questions remain regarding the utility and application of these different models, there is enough evidence to suggest that these models may be considered in future research in which both time-varying and time-invariant constructs are measured.

Chapter 3: A preliminary MLCFA with randomly-varying measurement parameters

The MLEFA in Study 1 approaches a specific question in psychometrics: What is the relation between trait and state-like symptoms? And though this may prove helpful for elucidating questions of this sort, that study is limited in its ability to account for individual differences. The resulting factor solution remains a strictly nomothetic factor solution: Each participant is afforded a value on the within and between factors, but every person is assumed to have the same within-person factor structure: In terms of what changes over time, this model suggests that the same features (symptoms) change for all individuals, though some might change positively or negatively. In line with the ergodic theories described above, we can be reasonably confident that the assumptions of invariance between persons in this model are inaccurate. Specifically, two components of the measurement model may require specification per person: factor loadings and factor (co)variances. Both of these are assessed directly by p-technique factor analysis on a person-specific basis. Other components of measurement models – residual (co)variances of observed indicators, for example – conceptually may also vary by person as well, and do so in true p-technique FA. However, these parameters are less commonly of interest in applied settings, where the focus is usually on latent variables rather than observed values.

In traditional FA, each observed variable – each item in this case – has a single factor loading on each latent variable that it measures. However, it is entirely possible that each person completing a measure may interpret a given item idiosyncratically. These different interpretations across people could be due to unique experiences with certain words, reading level, or other individual differences. For instance, an item intended to measure depression using the word “worthless,” as in “I feel worthless,” may suggest to some people a sense of personal

loss of meaning or value only relative to themselves, while to others it may suggest a comparison between themselves and other people. Still other people may have a particular memory or experience associated with the word "worthless" that would make it difficult for them to identify that feeling in themselves, if, say, they were told as a child repeatedly that they should always behave as though they were "worth something." Each of these examples would alter the quality of measurement that a given observed variable could attain for a given person, and imply a different factor loading for this item on the latent variable.

The second component of measurement models that may require specification for each person is the factor variance. This is a measure of variability similar to intraindividual standard deviation when assessed in intensive longitudinal data. Many mood and personality disorders are marked by widely varying moods, for instance a euthymic person entering a depressive episode might demonstrate a high amount of variability in depressed mood compared to some other people, as their mood fluctuates from normal to extremely depressed and back again. On a different time scale, it is expected that persons with Borderline Personality Disorder will experience highly variable moods ranging from elation to anger to sadness in quick succession. In measurement terms, this implies significantly varying factor variances: People will differ from one another in terms of how variable their scores on a mood measure will be in time.

Both factor loadings and factor variance have been shown to vary across persons in empirical applications. Factor variance has perhaps been studied more frequently, for instance as has already been mentioned in intensive longitudinal data (e.g., Hedeker, Mermelstein, & Demirtas, 2012; Ram & Gerstorf, 2009). In several settings, this intraindividual variability has been shown to be predictive of outcomes of interest. Differences between individuals in factor loadings have also been studied, though not as commonly. One systematic and theory-based

suggestion that factor loadings could vary between people was made by Nesselroade, Gerstorf, Hardy, and Ram (2007) who suggested that factor loadings should be allowed to vary across people while other model parameters were held constant. These authors called their extended method the "idiographic filter," and used analyses based on this to show that certain adjectives were more emblematic of latent constructs for some participants than others. The idea of the idiographic filter has not become commonplace, although it has been extended for use in more complex DFA models (Molenaar & Nesselroade, 2012). These authors have illustrated that it is possible to measure similar latent constructs across people while allowing individual items to have varying factor loadings.

A general approach to this problem was laid out by Ansari, Jedidi, and Jube (2002), who proposed that all factor parameters, including factor loadings, could be construed as random variables within a factor analysis model (though not all simultaneously). That is, they suggested not only that factor loadings be allowed to vary between people, but that the differences between people could be modeled using probability distributions. This has been used in some theoretical and practical applications, for instance by Maydeu-Olivares and Coffman (2006), who demonstrated that including a single random parameter for factor intercept per person could be easily estimated in most SEM programs and may be a useful method in FA models. Molenaar (2004) also demonstrated that classical test theory fails to sensitively assess heterogeneity between people in factor models, when it is present. Kelderman and Molenaar (2007) further demonstrated that when factor loadings differ between participants, standard FA models simply miss this fact and can report good fit to the (R-technique) data when it is not the case – a particularly concerning finding, given that FA is a standard method of psychological analysis.

To summarize the state of affairs, this is conceptually noncontroversial and conforms to current understanding of individual differences: individuals, in all likelihood, do differ considerably in both factor loadings and factor variances. Estimating the extent of this variability, if even possible, may allow more accurate estimates of their factor scores – in the present case, their symptoms. What seems to be missing is a broad push from applied researchers to use techniques that might be able to detect this. In a potentially helpful recent development, Asparouhov and Muthén (2012) have developed and implemented what they refer to as individual-differences factor analysis (IDFA). This is an extension of MLFA, requiring multiple observations per person. The primary developments in this model that are not present in another MLFA are randomly-varying factor loadings and factor variances. This is accomplished within a general SEM modeling program and is facilitated by the use of Bayesian estimation rather than ML or WLSMV. Bayesian estimation is naturally suited to computation of random effects, since all parameters are assumed to be distributed randomly in such an analysis. Asparouhov and Muthén present an empirical example of IDFA comparing individuals with Borderline Personality Disorder to healthy controls. In this single-factor example, they found significant and non-ignorable random variance on factor loadings and factor variances. Interestingly, and counter to hypotheses, they did not find that members of the BPD group showed significantly greater mood variance than controls, once accounting for random factor loadings and factor variances between people. This demonstrates the utility of including these randomly-varying parameters in clinical assessments.

The present study is an attempt to explore the feasibility and utility of a random factor analysis model. Specifically, using a multilevel factor analysis design, random parameters for factor loadings were estimated, reflecting differences between individuals in the strength of item

loadings across time in each within-person factor. Only within-person parameters can be assessed as randomly varying across individuals, as the between-person factors already, by definition, vary across people. This is an experimental method and its feasibility in a model as complicated as the CCAPS-34 has not been explored.

For item p, individual i, and time j, the model for observed values Y_{pij} is as follows:

$$Y_{pij} = \mu_p + s_{pi}\eta_{ij} + \zeta_{pi} + \varepsilon_{pij}$$

$$\eta_{ij} = \eta_{ij}^{W} + \eta_{i}^{B} + \xi_{ij}^{W} + \xi_{i}^{B}$$

$$s_{pi} = \lambda_{p}^{W} + \lambda_{p}^{B} + \lambda_{p}^{W}\sigma_{pi} + \varepsilon_{pi}$$

$$\sigma_{pi} = \zeta_{pi}$$

Here, the μ_p represents the item intercepts (the grand mean), separate levels of variation are denoted by B (between-person) and W (within-person), latent factor scores are represented by η, factor loadings by λ, and the random factor loading is denoted by σ. Random effects ε, ξ, and ζ are all assumed normally distributed with mean = 0, and their variances are estimated. Covariances between all random effects are set to 0. This model allows separate definitions of within-person and between-person factor structure, so that λ_p^W is multiplied by η_{ij}^W only, while λ_p^B and η_i^B are separately paired. Thus, a different pattern matrix for each lambda can be separately produced, though neither differs according to person or time (only λ_p^W varies by individual as an effect of its multiplication by σ_{pi}).

The goal of this study, therefore, was to explore the impact of including random measurement parameters into an MLFA. Particular attention was paid to random factor loadings in this case, since the present study is focused on measurement of latent factors which might differ by person rather than differences between persons in how much a given type of symptom

varies over time. The rationale for this is that though change over time is very clearly of interest in studies of psychological symptoms, variation in symptoms associated with meaningful improvement and deterioration would not necessarily be separable from error variation; a more thorough analysis would account for autocorrelation and detrending. The Nonclinical sample was selected as an ideal testing ground for this method, given the comparatively simple between-person factor structure revealed in that sample, which reduces the complexity of these models.

Method

Participants

The sample in this study is the same as the nonclinical sample in Study 1. This sample was comprised of 1223 students recruited from the Department of Psychology subject pool at a large Mid-Atlantic university over seven consecutive semesters from Spring 2011 – Spring 2014. Each semester, between 125 and 248 participants were recruited, and received course credits. The subjects ranged in age from 18-53 (M = 19.2, SD = 2.09). The majority (899) were female, 316 were male, and one participant identified as transgender. Most students (906) were White, 150 were Asian/Asian-American, 67 were African-American/Black, 45 were Hispanic/Latino/a, 32 identified as Multi-racial, and 12 chose to self-identify race/ethnicity. Roughly half (603) of the participants were Freshmen, 350 were Sophomores, 180 were Juniors, and 73 were Seniors, with 10 participants indicating a different status. At the start of each semester, participants were asked whether they were in some form of counseling or psychotherapy, and 102 indicated that they were. Similarly, 108 participants indicated that they had a current prescription for a psychiatric medication (66 participants reported both psychological and pharmacologic treatment at the start of the study). All participants completed informed consent procedures approved by the IRB.

Measures

CCAPS (Locke et al., 2011; 2012). The CCAPS instruments were designed to be multidimensional assessments of several common psychological symptoms treated in college counseling centers. The CCAPS-34 has 34 items, reduced from the 62 of the CCAPS-62 through classical test theory and item-response theory methods (Locke et al., 2012). The 34 items are scored using seven factor-analytically derived subscales: Depression, Generalized Anxiety, Social Anxiety, Hostility, Alcohol Use, Eating Concerns, and Academic Distress. In addition, the recently-developed Distress Index (DI) is scored in clinical practice. The DI is a general measure of distress, developed through bifactor modeling (CCMH, 2012; Nordberg et al., under review), and is scored by averaging 20 items from four subscales of the CCAPS-34. The seven-factor structure has received good support in a large (N=19,247) single-time point sample, each subscale appears to have good convergent validity with other measures of the construct it was designed to assess, and the DI appears to correlate highly with another measure of general distress (the OQ-45 total score; Nordberg et al., under review). In both samples, all administrations of CCAPS-62 were rescored as the CCAPS-34.

Standardized Data Set (SDS). The SDS is a variable-length questionnaire designed to facilitate and standardized intake procedures at counseling centers. It assesses basic demographic and previous service utilization history. Each counseling center may administer selected items from the SDS.

Procedure

Participants were recruited to participate in 14 weekly assessments, which encompassed most of each 16-week academic semester. Data were collected over seven semesters, each with unique participants. Each week, participants completed a CCAPS (the CCAPS-62 at week 1, the

CCAPS-34 thereafter). They also completed sample SDS items for demographic information at week 1.

Data Analysis

As a baseline, a multilevel factor model incorporating within-person and between-person factors of the CCAPS-34 was estimated in a previous study in the same sample. This model included seven within-person factors and two between-person factors. The within-person factor model differed from the standard CCAPS-34 model in few ways, but largely retains its factor loading structure. The between-person factor loading structure includes one general factor, with loadings on most items, and one more specific factor with loadings on Alcohol Use and socially relevant items.

To this baseline model, random factor loadings were added for each item. Bayesian estimation was used to facilitate convergence with this large number of random parameters. Inverse-Wishart priors were included for the variance-covariance parameters. These priors will tend to push these estimates to be greater than 0. The primary method of convergence check was the Potential Scale Reduction Factor (PSR; Asparouhov & Muthén, 2010; Gelman & Rubin, 1992), which for the purposes of this study was set to 1.2. This value is more liberal than the standard of 1.1 but was deemed necessary considering the large number of iterations required to converge even to this level. This corresponds to a nontrivial amount of difference between two MCMC chains, though may be expected to be acceptable or close to it.

After completion of the random factor loadings model, secondary analyses were conducted in the form of a series of Monte Carlo simulation studies to check the viability of this solution. To simulate the CCAPS-34, a 35-item measure was simulated with 7 factors, each with 5 continuous indicators. Data from a sample of 1000 people with 10 observations each were

generated. Several different models were compared, with varying complexity, and are discussed below.

Results

The random factor loadings model converged to PSR=1.2 after 119,500 iterations of the two MCMC chains, requiring 129 hours continuous computation time on a 1.7 GHz dual-core laptop. The within-person factor correlation matrix is presented in Table 8. The correlation pattern found here is similar to correlations found with the CCAPS-34 in other contexts (e.g., McAleavey et al., 2012), in that it shows moderate correlations between all factors with highest correlations between Depression, Hostility, and Generalized Anxiety, while the lowest correlations involve Alcohol Use and Eating Concerns. The factor loadings are presented in Table 9. All factor loadings (estimated as item between-person level intercepts) were significant at both the within and between levels, though notably there were two items ("I am shy around others" and "I make friends easily") with considerably smaller loadings on the between-person Alcohol Use factor. These two factor loadings were included due to preliminary evidence that they were meaningful loadings; their direction and scale in this model suggest they are less important. Their removal would make this factor a replication of the within-level Alcohol Use scale. Based on the factor matrix and the factor loading pattern, the model appeared to largely substantiate the EFA reported in Study 1. Since this is a form of CFA, this is heartening, though less surprising because it was conducted on the same data as the EFA.

Table 8. Factor correlation matrix from CCAPS-34 random factor loading model

	Eating Concerns	Generalized Anxiety	Social Anxiety	Depression	Hostility	Alcohol Use	Academic Distress
Eating Concerns	1.000						
Generalized Anxiety	0.597	1.000					
Social Anxiety	0.585	0.679	1.000				
Depression	0.524	0.745	0.693	1.000			
Hostility	0.525	0.745	0.637	0.686	1.000		
Alcohol Use	0.515	0.523	0.420	0.385	0.554	1.000	
Academic Distress	0.495	0.636	0.602	0.674	0.547	0.337	1.000

Table 9. Factor loadings in Study 2.

		Within-person factors							Between-person factors	
Item	Text	Generalized Anxiety	Eating Concerns	Depression	Hostility	Alcohol Use	Social Anxiety	Academic Distress	General Distress	Alcohol Use
1	I am shy around others						0.804		1.796	-0.618
2	My heart races for no good reason	1.024							2.034	
3	I feel out of control when I eat		1.393						2.660	
4	I don't enjoy being around people as much as I used to[W]								0.659	
5	I feel isolated and alone			1.245					2.427	
6	I think about food more than I would like to		1.470						2.541	
7	I am anxious that I might have a panic attack while in public	1.166							2.381	
8	I feel confident that I can succeed academically			0.143					1.159	
9	I have sleep difficulties	0.769							1.545	
10	My thoughts are racing	1.209							2.074	
11	I feel worthless			2.045					3.969	
12	I feel helpless			1.772					3.481	
13	I eat too much		1.586						2.523	
14	I drink alcohol frequently					1.273				3.513
15	I have spells of terror or panic	1.182							2.411	
16	When I drink alcohol I can't remember what happened					1.280				3.548
17	I feel tense[W]								0.755	
18	I have difficulty controlling my temper				1.475				2.260	
19	I make friends easily[W]								0.582	-0.278
20	I sometimes feel like breaking or smashing things				1.221				1.855	
21	I feel sad all the time			1.262					2.662	
22	I am concerned that other people do not like me						0.953		2.175	
23	I get angry easily				1.462				2.345	
24	I feel uncomfortable around people I don't know						1.010		1.966	
25	I have thoughts of ending my life			1.081					2.212	
26	I feel self conscious around others						1.034		2.240	
27	I drink more than I should					1.388				3.747
28	I am not able to concentrate as well as usual							1.183	2.277	

29	I am afraid I may lose control and act violently	1.303			2.609	
30	It's hard to stay motivated for my classes			1.196	2.317	
31	I have done something I have regretted because of drinking		1.193			2.765
32	I frequently get into arguments	0.989			1.958	
33	I am unable to keep up with my schoolwork			1.065	2.134	
34	I have thoughts of hurting others	1.068			2.298	

Note. All loadings are significant $p < .001$. Cell outline indicates loadings in the scoring system of the CCAPS-34. [w]: item did not have any meaningful loadings at the within level.

The inclusion of random factor loadings was also apparently meaningful, since all of the factor loadings were significant ($p < .001$) with a Wald test. This suggests that the between-persons variation in factor loadings was greater than 0: People differed on all factor loadings. These factor loading variances are presented in Table 10. The variances of these distributions ranged from 0.36 to 1.2 suggesting that some items' factor loadings had notably greater between-person variability than others. Some of the most highly variable items were on the Eating Concerns and Alcohol Use subscales, but all subscales had some items with at least moderately variable loadings. Though most of the results appeared plausible, though some implausibly high residual variances for some CCAPS-34 items were found (greater than 8, twice the range of the item value). This seems to suggest some degree of model failure, possibly due to the liberal PSR convergence requirement. A PSR of 1.2 roughly translates to 30% of the variance in estimates differing between chains, which is quite high. Thus, the current estimates are taken to be less than perfectly believable.

While further iterations of these models would be advisable, they were not conducted due to concerns about the trustworthiness of the model in general. Since this is the first application of these methods, simulation studies were instead undertaken to explore their utility.

Table 10. Factor loadings' random variances in Study 2.

Item	Text	Generalized Anxiety	Eating Concerns	Depression	Hostility	Alcohol Use	Social Anxiety	Academic Distress
1	I am shy around others						0.555	
2	My heart races for no good reason	0.446						
3	I feel out of control when I eat		0.956					
4	I don't enjoy being around people as much as I used to[W]							
5	I feel isolated and alone			0.348				
6	I think about food more than I would like to		4.038					
7	I am anxious that I might have a panic attack while in public	0.711						
8	I feel confident that I can succeed academically			0.914				
9	I have sleep difficulties	0.511						
10	My thoughts are racing	0.606						
11	I feel worthless			0.981				
12	I feel helpless			0.641				
13	I eat too much		1.201					
14	I drink alcohol frequently					0.910		
15	I have spells of terror or panic	0.482						
16	When I drink alcohol I can't remember what happened					1.402		
17	I feel tense[W]							
18	I have difficulty controlling my temper				0.598			
19	I make friends easily[W]							
20	I sometimes feel like breaking or smashing things				0.477			
21	I feel sad all the time			0.433				
22	I am concerned that other people do not like me						0.549	
23	I get angry easily				0.578			
24	I feel uncomfortable around people I don't know						0.902	
25	I have thoughts of ending my life			0.360				
26	I feel self conscious around others						0.750	
27	I drink more than I should					1.110		
28	I am not able to concentrate as well as usual							0.531
29	I am afraid I may lose control and act violently				0.525			

30	It's hard to stay motivated for my classes			1.072
31	I have done something I have regretted because of drinking		0.929	
32	I frequently get into arguments	0.400		
33	I am unable to keep up with my schoolwork			0.644
34	I have thoughts of hurting others	0.520		

Note. All loading variances are significant $p < .001$. Cell outline indicates loadings in the scoring system of the CCAPS-34. [W]: item did not have any meaningful loadings at the within level.

The secondary analyses were conducted in order to determine whether random factor loading variance could be estimated with accuracy in such complex models. Briefly, a simpler model was construed in which seven factors were indicated by five continuous indicators each, and sample data were generated from 1000 clusters each sampled 10 times. Parameter values roughly similar to those seen in the CCAPS factor solution were selected: factor and item variances were set to 1.00; inter-factor correlations were set to 0.50, factor loadings were set to be 0.60 in aggregate with small variances of 0.30. The between-person model was modeled as a single-factor with uniform loadings of 0.60. For the purposes of simulation, convergence criterion of PSR = 1.1 was used.

The first simulation isolated factor loading variances as freely estimated variables, while holding the between-person measurement model null and using a constrained IDFA-type structure to model the random factor loading covariances. That is, a single general "factor loading variance factor" was estimated with fixed loadings of 0.6 on all random factor loading variances. This fixed factor thereby reduces the information required to generate and identify factor loading variances. Twenty replications were conducted because this was considered a preliminary step. Under these conditions, the random factor loadings of within-person variables and all other model parameters were well recovered. Table 11 contains a summary of these results. All parameter estimates had minimal bias (each type of parameter showed average bias < .04), with the highest bias among the variances of factor loadings (the main parameters of interest) at 0.03. The standard error biases were also encouraging, with average SE bias less than 0.07 for all parameters, and only greater than 0.05 for parameters of the observed indicators (intercepts and variances). This indicates that the parameters would not suffer loss of accuracy or power under these conditions. Finally, the coverage rates in these simulations were close to .95

for most types of parameters, providing evidence that Type I error would not be inflated either. The notable exception to this was in the group mean factor loadings, which are effectively the factor loading values in a standard FA model. For these values, though parameter bias and SE bias were both acceptably small, coverage only averaged to .79. That is, only about 80% of the 95% confidence intervals contained the true value of these parameters. This indicates potentially inflated Type I error rate (20% when it should be 5%) for this parameter type. All non-zero parameters were significantly estimated in 100% of simulations, demonstrating sufficient power. So this suggests that this model would possibly produce somewhat more inaccurate results than desired.

Table 11. Simulation one: Random factor loadings with fixed factor loading variance factor.

		Population		Estimate average				
Type of parameter	# parameters	True value	SD	Parameter estimates	SE Mean	Avg. 95% Coverage	Parameter bias	SE bias
Within-person								
Factor covariances	21	0.5	0.011	0.500	0.011	0.943	-0.001	-0.015
Factor variances	7	1.0				fixed		
Residual variances of observed indicators	35	1.0	0.021	1.001	0.020	0.939	0.001	-0.027
Between-person								
Factor loadings of random factor loadings on a general variation factor	35	0.6				fixed		
Group mean of within-person factor loadings	35	1.0	0.031	1.033	0.029	0.787	0.033	-0.038
Intercepts of observed indicators	35	0.0	0.027	-0.001	0.028	0.964	-0.001	0.064
Residual variances of observed indicators	35	0.6	0.034	0.604	0.035	0.947	0.006	0.053
Variances of random factor loadings	35	0.3	0.023	0.303	0.023	0.930	0.009	0.019

Following this, a second model including a factor loading variance factor was conducted. As in the previous model, the covariances between the random variances of factor loadings for the 35 items were modeled with a single factor, but the factor loadings were estimated rather than held fixed at the correct values. This is part of the model proposed by Asparouhov and Muthén (2012) as IDFA, designed to allow the random factor loadings to be minimized relative to variation in factor variances (which were not included in the present model). Though data could be simulated successfully and repeatedly, convergence became a computational challenge under this circumstance. Even given the proper starting values, the MCMC chains diverged too greatly from the start and exceeded 50,000 iterations without reaching PSR = 1.1. Increasing the number of observations per person to 40 while decreasing the number of people to 500, which should improve specificity of random variable estimates, did not resolve this estimation problem. This may be a case of empirical unidentifiability in the model, requiring many more observations and possibly greater variation than simulated here to ameliorate. Asparouhov and Muthén avoided this by adding constraints, such as constraining the intercept of each item's factor loading to be equal to the item's factor loading on the factor loading variance factor. While this seems to make this identifiable and estimable, since the general factor will then absorb most of the factor-specific variability from the factor loadings, this also diminishes the model's capacity to estimate item-specific variability in factor loadings. That is, the person-specific factor loadings are no longer independent of one another and instead each person has a general factor structure deviance factor, which reflects overall strength of the factor loadings for each factor. While the ultimate costs of this choice are likely minor in the estimates, this concession alters the interpretation of these parameters extensively.

Finally, a model was estimated without a general factor for random factor loadings, leaving them instead to be orthogonal, and including instead a between-person general distress factor with uniform loadings from the 35 items. This model emulates a two-level model with different between-person and within-person factor structures. Here, 20 replications were generated and estimated with relative speed: roughly 12 hours of computation were required. Results are summarized in Table 12. Most parameter types were well recovered, demonstrating little bias. Though coverage rates were all close to .95, multiple parameter types showed moderately increased Type I error, at rates of 8% rather than 5%. This would not be a major deviation, and suggests that under conditions such as this, a model like the one estimated for the CCAPS-34 could recover with general fidelity, at least under optimal conditions (including knowing the correct factor structure).

Table 12. Simulation two: Random factor loadings with different between-person and within-person factor structures.

		Population		Estimate average				
Type of parameter	# parameters	True value	SD	Parameter estimates	SE Mean	Avg. 95% Coverage	Parameter bias	SE bias
Within-person								
Factor covariances	21	0.5	0.011	0.502	0.010	0.919	0.004	-0.068
Factor variances	7	1.0				fixed		
Residual variances of observed indicators	35	1.0	0.021	1.002	0.020	0.923	0.002	-0.046
Between-person								
Factor loadings of observed indicators on between-level general factor	35	0.6	0.034	0.603	0.034	0.944	-0.005	0.001
Group mean of within-person factor loadings	35	1.0	0.022	1.002	0.023	0.946	-0.002	0.058
Intercepts of observed indicators	35	0.0	0.037	0.007	0.033	0.924	0.007	-0.102
Residual variances of observed indicators	35	0.6	0.034	0.604	0.035	0.953	0.006	0.061
Variances of random factor loadings	35	0.3	0.023	0.303	0.022	0.921	0.010	0.005

Discussion

The aim of this second study was to explore the feasibility and utility of including random factor loadings to a multidimensional model of psychopathological symptoms over time. Though it clearly is far from a comprehensive analysis of the possibilities, this study illustrates two points: First, it is possible to estimate random factor loadings between people with as few as 10 observations per person, given a large number of persons (in this case, over 1000). Second, when considering the numerous possible ways to incorporate random measurement parameters, it is obvious that some theoretically identified models will lack sufficient information in practice to produce any estimates, let alone estimates that can be trusted. These models' fickle nature became apparent after attempting to estimate similar models with either fixed or estimated between-person effects led to a failure to converge and lengthy computations. These computational failures likely relate to variance components, which, due to insufficient evidence to support them, fall to 0. This will slow the process of MCMC sampling and require more iterations to converge, to a point where the feasibility of these analyses is obviously questionable. Future analyses should investigate whether the computational and identification problems found in this context are avoidable.

In this study, only a small sample of variables that seemed most likely to create modeling problems and solutions was analyzed. Several additional avenues could be explored with Monte Carlo simulation studies to further illuminate the findings here. Some future directions include:

1. Model complexity: the factor structure used in this setting was relatively complex, requiring 35 items and 7 factors. Many instruments are simpler and have fewer within-person factors, which would be simpler to model. This might lead to better detection of random measurement parameters.

2. Model misspecification: in this study, all simulations were properly specified with proper starting values, but in applied cases (including this study), the true model is unknown. The robustness of results to misspecification – particularly the effect on estimates of random factor loadings when random factor variances are treated as fixed – is unknown.
3. Categorical indicators: In the simulations conducted, continuous factor indicators rather than categorical ones (such as Likert-type responses) were used. Though extensions are straightforward, using categorical indicators may require more observations per person and result in non-normal distributions of responses.
4. Effects of other variables: Including predictors, covariates, and outcomes of measurement parameters should be straightforward and could be of interest (for instance, perhaps a strong loading on a particular item predicts negative outcome to therapy).
5. Factor variance and covariances as random variables: Though this study primarily focused on factor loadings the covariation among latent factors could vary by person greatly as well. If there are multiple within-person factors estimated, they may correlate with one another at the between-person level. How best to estimate these values is an open question. The attempt to use IDFA was not feasible with this data structure, but other methods to model relationship of random factor loadings could be explored.
6. Power calculations for number of people, number of observations, and number of observations per person: Much like in other multilevel analyses, these are the primary (but not the only) sources of statistical power. The number of observations required is not known for these measurement parameters.
7. Magnitude of random factor loading variation relative to within-person variation in measurement and between-person variability in factor variance and item intercepts. These

are other values that determine statistical power and Type I error rates in multilevel analyses, and as such need to be explored.

8. The effect of detrending item and observed variables prior to factor analysis: A common and necessary practice for analyzing time series data is detrending, which ensures that variability is not the result of a growth or decay process. This would allow a more conceptually coherent analysis of factor variance differences between individuals, and could separate change over time from random fluctuations. However, doing so would also greatly change the interpretation of some measurement parameters, including item intercepts and factor loadings, which would then refer to deviations around an expected mean in time rather than deviations around an overall mean per person. With few measurement points per person, this could also induce more uncertainty into the model.

Given the considerable uncertainty regarding these methods raised by this study, the challenges to estimation, and the many potential confounding variables requiring future study suggested above, it is difficult to interpret the findings of the applied case of random factor loadings using the CCAPS-34. However, as preliminary analyses go, it is highly suggestive. Namely, it appears to substantiate the idea that individuals interpret, value, or rate certain items in idiosyncratic ways. Even though the factor variances were not allowed to vary between-persons, all the factor loadings varied considerably and significantly. This very clearly shows promise as a way to blend nomothetic and person-specific measurement of psychopathology. The factors being estimated are essentially unchanged from the conventional CCAPS-34 structure – complete with a very similar correlation matrix – but each person in the data set is allowed to generate a factor score in a slightly different way. This constrains the factors themselves to be similar across all participants, even if the definitions are idiographic.

One area for future exploration would be to establish a test of the clinical significance of this method: do the factor scores generated with random measurement models improve on the accuracy and predictive power of nomothetic models? A proper test could theoretically be conducted between a model with random measurement parameters and one without, for model fit, for criterion-related validity, and predictive accuracy. However, it is more complicated for comparison of models that use information from multiple time points to those that rely only on a single time point. For one thing, the mean of this series contains information from both the start and the finish of the course of observations, which effectively places this model at an advantage. More distal outcomes, such as symptom course over longer periods of time, may be required in order to detect any differences, though because these factors are so similar it is highly unlikely that factor scores will differ greatly between these models. This remains yet another open question.

Finally, though several computational challenges emerged, these are likely to be overcome by increasing computational power and other advances in the near future. Psychometric analyses should no longer rely on methods that were formulated a century ago, and should instead move towards integrating theory into quantitative measurement, expanding the range of computational methods used.

Chapter 4: General Discussion

In this thesis, two studies focused on the measurement of psychological symptoms have been presented. Each study expands on traditional factor analysis by making use of more than one observation per person, and each takes advantage of modern advances in statistical computation. The primary finding from the first study is that the structure of differences between people is not the same as the structure of differences over time, at least for psychological symptoms assessed by self-report. The degree of discrepancy between within-person and between-person variability in psychological symptoms appears to be at least partially a function of the heterogeneity within the sample, because a more heterogeneous clinical sample generated greater complexity at the between-person level. Additionally, it should be noted that the two structures are obviously not completely distinct: the between-person factor structure was essentially a simpler version of the within-person factor structure, tending towards a general factor with one or more domain-specific factors divided by symptom type.

In the second study, an initial attempt to incorporate random factor loadings was the main focus. Though some results of this analysis were definitely promising – notably, the real data analysis suggests that significant random factor loadings exist and some simulation results suggested that models of this sort should be feasible, accurate, and unbiased – other results were less promising. Numerous convergence errors were encountered, strongly suggesting that some of these models may demand too much of too little data. Further unanswered questions regarding these methods remain as well. In the end, this study demonstrates plausibility, but probably not feasibility of the method, at least using present methods. Future advances in computation and estimation of random effects may make these methods more feasible.

By accounting for differences between persons and over time as separate constructs, these studies provide a novel advancement in the field of psychometrics. This is made possible by including not only one timepoint of observation per person, but several, and then using many individuals' data to inform a somewhat person-specific result. This effectively leverages information from several people into each person's solution. This is not a true person-specific analysis, and such analyses would not be possible without many more observations per person. Indeed, fully person-specific analyses are more flexible than the current approaches. However, a true person-specific approach does not parameterize between-person differences as easily as the use of random variables. So though p-technique factor analysis, for instance, would generate ideal factor structures for within-person variability, it would not generate between-person factor structures at all.

There are several important clinical implications of the studies described here. The first is that greater precision in measurement of symptoms should allow clinicians to better understand their patients' symptoms. This could alter case formulations and targets of intervention. Especially with a multidimensional instrument like the CCAPS-34, the potential for changing case formulations is great. Bearing in mind that a particular answer to a particular item is influenced by state and trait variables, as well as person-specific idiosyncrasies, should allow clinicians to prioritize symptoms with greater certainty. If a particular symptom is notably elevated one session, following several sessions without much change, and this symptom has previously been a strong indicator of its factor for the particular client, clinicians may quickly intervene to reduce its severity. This should be expected to provide greater return on investment.

An additional clinical benefit would be if the types of analyses used in these studies could reveal factors of the CCAPS-34 which are especially prone to change in treatment, and/or factors

which are more difficult to alter in short psychosocial interventions. If certain symptom clusters tend to be malleable in short time frames of treatment, these would be excellent targets for intervention in counseling centers. On the other hand, factors which tend not to change can either be viewed as helpful information for case formulation (e.g., personality traits), or might be considered reason to refer clients for more intense treatments. Augmenting counseling with psychiatric medication, group therapy, or other services may be appropriate if clients are especially bothered by symptoms that are difficult to change during treatment.

Aside from the method, the results of these studies also bear on routine clinical practice. The major conclusion of Study 1 is that the within-person factor structure of the CCAPS-34 is very similar to the R-technique defined standard structure. So, as a first approximation, the standard factor structure is a good guess as to what items are likely to change with one another in practice. Simply phrasing this relationship in this way – change in one item is likely to co-occur with change in another item – allows one to see new avenues of intervention. If a therapist is struggling to create behavioral change because of a highly distressed client (e.g., one who reports high levels on item 14, "I drink alcohol frequently") may instead wish to change focus to other items which focus on the consequences of drinking (e.g., item 31, "I have done something I have regretted because of drinking"), with the confidence that change in one will be related to change in the other. This is clearly not a new observation in the clinical realm, but the link between such items in time has not been clear before.

A major limitation of the current approach is the lack of accounting for temporal dependencies within-person. Theoretically, a dynamic factor analysis model could be estimated as a multilevel model, with randomly varying temporal dependencies as well as other parameters. Perhaps such an analysis would yield different results for within-person factor

structure, and would clearly inform on the variability of change processes. However, these models would likely be considerably more difficult to estimate than those used here, and would likely require a greater number of observations within-persons in order to estimate the dynamic structures.

The primary remaining questions around these studies are questions of usefulness, not only feasibility. If multilevel factor analyses produce better results statistically, does it follow that better clinical understanding will follow? Not necessarily, but quite possibly. In the context of treatment outcome monitoring, two main goals are generally set: prediction of course or outcome, and interpretation of symptomatic improvement in real time. A multilevel framework could bear on both. For the question of prediction, these models generate separate factor structures for items that are more prone to change (within-person level) and those that are less prone to change (between-person), and therefore should be better predictors of outcome. However, they require knowledge of multiple time points prior to prediction, which is not always feasible. In real-time analysis of change, these models (especially those from Study 2) should provide improvements, since they would be capable of identifying any specific items that are especially significant predictors of symptoms. However, the size of any improvement cannot be known.

While past psychometric analyses have generated great insights about the structure and dynamics of unobservable psychological features, current computational advances are beginning to supersede simple R-technique factor analysis as the primary method of investigation. Psychologists and researchers should begin to use these methods and take full advantage of the opportunity to better describe quantitatively our clinical and theoretical knowledge: People are not interchangeable, and each person's symptoms may not be identical.

References

Adcock, C. J. (1964). Higher-order factors. *British Journal of Statistical Psychology, 27*, 153-160.

Ansari, A., Jedidi, K., & Dube, L. (2002). Heterogeneous factor analysis models: a Bayesian approach. *Psychometrika*, *67*, 49-77.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397-438.

Asparouhov, T., & Muthén, B. (2012, July). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. Presented at the Annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia.

Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science*, *4*, 19-21.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, *28*, 135-167.

Bludworth, J. L., Tracey, T. J., & Glidden-Tracey, C. (2010). The bilevel structure of the Outcome Questionnaire–45. *Psychological assessment*, *22*(2), 350-355.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., ... & Fox, J. (2011). OpenMx: an open source extended structural equation modeling framework. *Psychometrika*, *76*, 306-317.

Borkenau, P., & Ostendorf, F. (1998). The Big Five as states: How useful is the five-factor model to describe intraindividual variations over time? *Journal of Research in Personality*, *32*, 202-221.

Brown, T. A., Chorpita, B. F., & Barlow, D. H. (1998). Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of Abnormal Psychology*, *107*, 179-192.

Castonguay, L., Barkham, M., Lutz, W., & McAleavey, A. (2013). Practice-oriented research: Approaches and applications. To appear in M.J. Lambert (Ed.). *Bergin and Garfield's Handbook of psychotherapy and behavior change* (6th ed.), pp. 85-133. Hoboken, NJ: Wiley.

Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika, 18*, 23-38.

Cattell, R. B. (1943). The description of personality. I. Foundations of trait measurement. *Psychological Review*, *50*(6), 559-594.

Cattell, R. B. (1965). Factor analysis: An introduction to essentials I. The purpose and underlying models. *Biometrics*, *21*(1), 190-215.

Cattell, R. B. (1966a) Patterns of change: Measurement in relation to state-dimension, trait change, lability, and process concepts. In R. B. Cattell (Ed.) *Handbook of Multivariate Experimental Psychology*. Chicago: Rand McNally.

Cattell, R. B. (1966b). The scree test for the number of factors. *Multivariate Behavioral Research, 2*, 245-276.

Cattell, R. B, & Bartlett, H. W. (1971). An R-dR-technique operational distinction of the states of anxiety, stress, fear, etc. *Australian Journal of Psychology, 23*, 105-123.

Center for Collegiate Mental Health (2012). *CCAPS 2012 Technical Manual.* University Park, PA.

Cervone, D. (2004). Personality assessment: Tapping the social-cognitive structure of personality. *Behavior Therapy, 35*, 113-129.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*, 189-225.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine, 13,* 595–605. doi: 10.1017/S0033291700048017

Diez Roux, A. V. (2002). A glossary for multilevel analysis. *Journal of Epidemiology and Community Health*, *56*, 588-594.

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, *73*, 1246-1256.

Eisen, S. V., Normand, S. L. T., Belanger, A., Spiro, A., & Esch, D. (2004). The revised Behavior and Symptom Identification Scale (BASIS-24): Reliability and validity. *Medical Care, 42,* 1230–1241. doi:10.1097/ 00005650-200412000-00010

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor- Clark, J., & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry, 180,* 51–60. doi:10.1192/bjp.180.1.51

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, *4*, 272-299.

Fiorello, C. A., Hale, J. B., Holdnack, J. A., Kavanagh, J. A., Terrell, J., & Long, L. (2007). Interpreting intelligence test results for children with disabilities: is global intelligence relevant? *Applied Neuropsychology*, *14*, 2-12.

Fiorello, C. A., Hale, J. B., McGrath, M., Ryan, K., & Quinn, S. (2001). IQ interpretation for children with flat and variable test profiles. *Learning and Individual Differences, 13*, 115–125.

Fisher, A. J., Newman, M. G., & Molenaar, P. C. M. (2011). A quantitative method for the analysis of nomothetic relationships between idiographic structures: Dynamic patterns create attractor states for sustained posttreatment change. *Journal of consulting and clinical psychology*, *79*(4), 552-563.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Sciences, 7*, 457-511.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. (2005). Statistical modeling of the individual: Rationale and application of multivariate stationary time series analysis. *Multivariate Behavioral Research*, *40*, 207-233.

Hayes, A. M., Laurenceau, J. P., Feldman, G., Strauss, J. L., & Cardaciotto, L. (2007). Change is not always linear: The study of nonlinear and discontinuous patterns of change in psychotherapy. *Clinical Psychology Review*, *27*, 715-723.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between‐subject and within‐subject variances in ecological momentary assessment data using mixed‐effects location scale models. *Statistics in Medicine*, *31*, 3328-3336.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41-54.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-185. doi:10.1007/BF02289447

Howard, K. I., Lueger, R. J., Maling, M. S., & Martinovich, Z. (1993). A phase model of psychotherapy outcome: Causal mediation of change. *Journal of consulting and clinical psychology*, *61*(4), 678-685.

Johnson, J. E., Burlingame, G. M., Olsen, J. A., Davies, D. R., & Gleave, R. L. (2005). Group climate, cohesion, alliance, and empathy in group psychotherapy: Multilevel structural equation models. *Journal of Counseling Psychology*, *52*, 310-321.

Kaiser, H. T. (1960). The application of electronic computers to factor analysis. *Educational Psychology and Measurement, 20*, 141-151.

Keith, T. Z., & Reynolds, M. R. (2010). Cattell–Horn–Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools*, *47*(7), 635-650.

Kelderman, H., & Molenaar, P. C. (2007). The effect of individual differences in factor loadings on the standard factor model. *Multivariate Behavioral Research*, *42*, 435-456.

Kraus, D. R., Castonguay, L., Boswell, J. F., Nordberg, S. S., & Hayes, J. A. (2011). Therapist effectiveness: Implications for accountability and patient care. *Psychotherapy Research*, *21*, 267-276.

Lambert, M. J., Hansen, N. B., Umpress, V., Lunnen, K., Okiishi, J., Burlingame, G. M., & Reisinger, C. W. (2001). *Administration and scoring manual for the OQ-45.* Orem, UT: American Professional Credentialing Services.

Lambert, M. J., Harmon, C., Slade, K., Whipple, J. L., & Hawkins, E. J. (2005). Providing feedback to psychotherapists on their patients' progress: clinical results and practice suggestions. *Journal of Clinical Psychology*, *61*(2), 165-174.

Lambert, M. J., & Ogles, B. M. (2009). Using clinical significance in psychotherapy outcome research: The need for a common procedure and validity data. *Psychotherapy Research, 19*, 493-501.

Laurenceau, J. P., Hayes, A. M., & Feldman, G. C. (2007). Some methodological and statistical issues in the study of change processes in psychotherapy. *Clinical Psychology Review*, *27*, 682-695.

Locke, B. D., Buzolitz, J. S., Lei, P.-W., Boswell, J. F., McAleavey, A. A., Sevig, T. D., Dowis, J. D., & Hayes, J. A. (2011). Development of the Counseling Center Assessment of Psychological Symptoms-62 (CCAPS-62). *Journal of Counseling Psychology, 58*(1) 97-109. doi: 10.1037/a0021282

Locke, B. D., McAleavey, A. A., Zhao, Y., Lei, P.-W., Hayes, J. A., Castonguay, L. G., Li, H., Tate, R., Lin, Y.-C. (2012). Development and initial validation of the Counseling Center Assessment of Psychological Symptoms-34 (CCAPS-34). *Measurement and Evaluation in Counseling and Development, 45*, 151-169. doi:10.1177/0748175611432642

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., and Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22*, 471-491.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S. and Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching, *Structural Equation Modeling, 16*, 439-476.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological methods*, *11*, 344-362.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259-284.

McAleavey, A. A., Nordberg, S. S., Kraus, D., & Castonguay, L. G. (2012). Errors in treatment outcome monitoring: Implications for real-world psychotherapy. *Canadian Psychology/Psychologie canadienne*, *53*(2), 105-114.

McCrae, R. R., & Costa Jr., P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*, 81-90.

McCrae, R. R., & Costa Jr., P. T. (2012). *Personality in Adulthood: A Five-Factor Theory Perspective*. New York: Guilford Press.

Minami, T., Davies, D. R., Tierney, S. C., Bettmann, J. E., McAward, S. M., Averill, L. A., ... & Wampold, B. E. (2009). Preliminary evidence on the effectiveness of psychological treatments delivered at a university counseling center. *Journal of Counseling Psychology*, *56*(2), 309-320.

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*, 181-202.

Molenaar, P. C. (1987). Dynamic assessment and adaptive optimization of the psychotherapeutic process. *Behavioral Assessment, 9*, 389-416.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*, 201-218.

Molenaar, P. C. M., & Nesselroade, J. R. (2009). The recoverability of p-technique factor analysis. *Multivariate Behavioral Research, 44*, 130-141.

Molenaar, P. C., & Nesselroade, J. R. (2012). Merging the idiographic filter with dynamic factor analysis to model process. *Applied Developmental Science*, *16*, 210-219.

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods, 3*, No. 1, pp. 45-58).

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Neisser, U., Boodoo, G., Bouchard Jr, T. J., Boykin, A. W., Brody, N., Ceci, S. J., ... & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77-101.

Nesselroade, J. R., Gerstorf, D., Hardy, S. A., & Ram, N. (2007). Focus article: Idiographic filters for psychological constructs. *Measurement*, *5*, 217-235.

Nordberg, S., McAleavey, A. A., Locke, B. D. Castonguay, L. G., & Hayes, J. A. Developing a general measure of distress for the Counseling Center Assessment of Psychological Symptoms. *Manuscript in preparation*.

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling*, *18*, 161-182.

Ram, N., & Gerstorf, D. (2009). Time-structured and net intraindividual variability: tools for examining the development of dynamic characteristics and processes. *Psychology and aging*, *24*, 778-791.

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2011). *MLwiN Version 2.23*. Bristol: Centre for Multilevel Modelling, University of Bristol.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*, 544-559.

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment, 24*, 282-292.

Rushton, J. P., & Irwing, P. (2011). The general factor of personality. *The Wiley-Blackwell handbook of individual differences*, p.132-161.

SAS Institute. (2011). *SAS/STAT 9.3 user's guide*. SAS Institute.

Schmid, J., and Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53-61.

Shoda, Y., Mischel, W., & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology*, *67*, 674.

Sinclair, K. O., & Molenaar, P. C. M. (2008). Optimal control of psychological processes: A new computational paradigm. *Bulletin de la Societe des Sciences Medicales Luxembourg*, *1*, 13-33.

Spearman, C. (1904). “General Intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201-292.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Thomas, M. L. (2012). Rewards of bridging the divide between measurement and clinical theory: Demonstration of a bifactor model for the Brief Symptom Inventory. *Psychological assessment*, *24*(1), 101-113.

Thurstone, L. L. (1947). *Multiple-Factor Analysis*. Chicago: University of Chicago Press.

Vittengl, J. R., Clark, L. A., Thase, M. E., & Jarrett, R. B. (2013). Nomothetic and idiographic symptom change trajectories in acute-phase cognitive therapy for recurrent depression. *Journal of Consulting and Clinical Psychology, 81*, 615-626.

Watkins, M. W., Glutting, J. J., & Lei, P. W. (2007). Validity of the full-scale IQ when there is significant variability among WISC-III and WISC-IV factor scores. *Applied Neuropsychology*, *14*, 13-20.

APPENDIX: Mplus code for simulation studies

Simulation 1:

```
TITLE:      Simulation with random factor loadings,
            fixed factor loading variance factor loadings;
MONTECARLO:
            NAMES ARE y1-y35;
            NOBSERVATIONS = 10000;
            NREPS = 20;
            CSIZES = 1000(10);
            NCSIZE = 1;

ANALYSIS:  TYPE = twolevel random; ESTIMATOR=bayes;
            proc=2;

MODEL POPULATION:
            %WITHIN%
            s1-s5 | e1 by y1-y5;
            l1-l5 | e2 by y6-y10;
            b1-b5 | e3 by y11-y15;
            c1-c5 | e4 by y16-y20;
            d1-d5 | e5 by y21-y25;
            f1-f5 | e6 by y26-y30;
            h1-h5 | e7 by y31-y35;
            e1@1; e2@1; e3@1; e4@1;
            e5@1; e6@1; e7@1;
            y1-y35*1;
            e1 with e2@0.5 e3@0.5 e4@0.5 e5@0.5 e6@0.5 e7@0.5;
            e2 with e3@0.5 e4@0.5 e5@0.5 e6@0.5 e7@0.5;
            e3 with e4@0.5 e5@0.5 e6@0.5 e7@0.5;
            e4 with e5@0.5 e6@0.5 e7@0.5;
            e5 with e6@0.5 e7@0.5;
            e6 with e7@0.5;

            %BETWEEN%
            y1-y35*0.6;
            s1-s5*0.3;
            [s1-s5@1];
            l1-l5*0.3;
            [l1-l5*1];
            b1-b5*0.3;
            [b1-b5*1];
            c1-c5*0.3;
            [c1-c5*1];
            d1-d5*0.3;
            [d1-d5*1];
```

```
            f1-f5*0.3;
            [f1-f5*1];
            h1-h5*0.3;
            [h1-h5*1];
            g by s1-s5@0.6 l1-l5@0.6 b1-b5@0.6 c1-c5@0.6 d1-d5@0.6 f1-f5@0.6 h1-
h5@0.6;
            g@1;

MODEL:
            %WITHIN%
            s1-s5 | e1 by y1-y5;
            l1-l5 | e2 by y6-y10;
            b1-b5 | e3 by y11-y15;
            c1-c5 | e4 by y16-y20;
            d1-d5 | e5 by y21-y25;
            f1-f5 | e6 by y26-y30;
            h1-h5 | e7 by y31-y35;
            e1@1; e2@1; e3@1; e4@1;
            e5@1; e6@1; e7@1;
            y1-y35*1;
            e1 with e2*0.5 e3*0.5 e4*0.5 e5*0.5 e6*0.5 e7*0.5;
            e2 with e3*0.5 e4*0.5 e5*0.5 e6*0.5 e7*0.5;
            e3 with e4*0.5 e5*0.5 e6*0.5 e7*0.5;
            e4 with e5*0.5 e6*0.5 e7*0.5;
            e5 with e6*0.5 e7*0.5;
            e6 with e7*0.5;

            %BETWEEN%
            y1-y35*0.6;
            s1-s5*0.3;
            [s1-s5*1];
            l1-l5*0.3;
            [l1-l5*1];
            b1-b5*0.3;
            [b1-b5*1];
            c1-c5*0.3;
            [c1-c5*1];
            d1-d5*0.3;
            [d1-d5*1];
            f1-f5*0.3;
            [f1-f5*1];
            h1-h5*0.3;
            [h1-h5*1];
            g by s1-s5@0.6 l1-l5@0.6 b1-b5@0.6 c1-c5@0.6 d1-d5@0.6 f1-f5@0.6 h1-
h5@0.6;
            g@1;
```

OUTPUT:

tech8 tech9;

```
Simulation 2:
  TITLE:      Simulation with random factor loadings,
              Estimated factor loading variance factors;
  MONTECARLO:
              NAMES ARE y1-y35;
              NOBSERVATIONS = 20000;
              NREPS = 1;
              CSIZES = 500(40);
              NCSIZE = 1;

  ANALYSIS:  TYPE = twolevel random; ESTIMATOR=bayes;
              PROC=2;
              BITERATIONS = 10000;

  MODEL POPULATION:
              %WITHIN%
              s1-s5 | e1 by y1-y5;
              l1-l5 | e2 by y6-y10;
              b1-b5 | e3 by y11-y15;
              c1-c5 | e4 by y16-y20;
              d1-d5 | e5 by y21-y25;
              f1-f5 | e6 by y26-y30;
              h1-h5 | e7 by y31-y35;
              e1@1; e2@1; e3@1; e4@1;
              e5@1; e6@1; e7@1;
              y1-y35*1;
              e1 with e2@0.5 e3@0.5 e4@0.5 e5@0.5 e6@0.5 e7@0.5;
              e2 with e3@0.5 e4@0.5 e5@0.5 e6@0.5 e7@0.5;
              e3 with e4@0.5 e5@0.5 e6@0.5 e7@0.5;
              e4 with e5@0.5 e6@0.5 e7@0.5;
              e5 with e6@0.5 e7@0.5;
              e6 with e7@0.5;

              %BETWEEN%
              y1-y35@0.6;
              s1-s5@0.3;
              [s1-s5@1];
              l1-l5@0.3;
              [l1-l5@1];
              b1-b5@0.3;
              [b1-b5@1];
              c1-c5@0.3;
              [c1-c5@1];
              d1-d5@0.3;
              [d1-d5@1];
              f1-f5@0.3;
```

```
            [f1-f5@1];
            h1-h5@0.3;
            [h1-h5@1];
            g by s1-s5@0.6 l1-l5@0.6 b1-b5@0.6 c1-c5@0.6 d1-d5@0.6 f1-f5@0.6 h1-
h5@0.6;
            g@1;

 MODEL:
            %WITHIN%
            s1-s5 | e1 by y1-y5;
            l1-l5 | e2 by y6-y10;
            b1-b5 | e3 by y11-y15;
            c1-c5 | e4 by y16-y20;
            d1-d5 | e5 by y21-y25;
            f1-f5 | e6 by y26-y30;
            h1-h5 | e7 by y31-y35;
            e1@1; e2@1; e3@1; e4@1;
            e5@1; e6@1; e7@1;
            y1-y35*1;
            e1 with e2*0.5 e3*0.5 e4*0.5 e5*0.5 e6*0.5 e7*0.5;
            e2 with e3*0.5 e4*0.5 e5*0.5 e6*0.5 e7*0.5;
            e3 with e4*0.5 e5*0.5 e6*0.5 e7*0.5;
            e4 with e5*0.5 e6*0.5 e7*0.5;
            e5 with e6*0.5 e7*0.5;
            e6 with e7*0.5;

            %BETWEEN%
            y1-y35*0.6;
            s1-s5*0.3;
            [s1-s5*1];
            l1-l5*0.3;
            [l1-l5*1];
            b1-b5*0.3;
            [b1-b5*1];
            c1-c5*0.3;
            [c1-c5*1];
            d1-d5*0.3;
            [d1-d5*1];
            f1-f5*0.3;
            [f1-f5*1];
            h1-h5*0.3;
            [h1-h5*1];
            g by s1-s5*0.6 l1-l5*0.6 b1-b5*0.6 c1-c5*0.6 d1-d5*0.6 f1-f5*0.6 h1-h5*0.6;
            g@1;
 OUTPUT:
            tech8 tech9;
```

VITA
Andrew A. McAleavey, M.S.

Education

2014-2015	Weill-Cornell Medical School/NewYork-Presbyterian Hospital Psychology Intern
2008-2015	The Pennsylvania State University Doctor of Philosophy in Psychology
2002-2006	Brown University A.B., Psychology, with Honors

Publications

Twenty Journal articles, including

1. McAleavey, A. A., Lockard, A. J., Castonguay, L. G., Hayes, J. A., & Locke, B. D. (2015). Building a practice research network: Obstacles faced and lessons learned at the Center for Collegiate Mental Health. *Psychotherapy Research, 25*, 134-151. doi: 10.1080/10503307.2014.883652
2. McAleavey, A. A., Castonguay, L. G., & Goldfried, M. R. (2014). Clinical experiences in conducting cognitive-behavioral therapy for Social Phobia. *Behavior Therapy, 45*, 21-35. doi: 10.1016/j.beth.2013.09.008
3. McAleavey, A. A., & Castonguay, L. G. (2014). Insight as a common and specific impact of psychotherapy: Therapist-reported exploratory, directive, and common factor interventions. *Psychotherapy, 51*, 283-294. doi: 10.1037/a0032410
4. McAleavey, A. A., Nordberg, S. S., Hayes, J. A., Castonguay, L. G., Locke, B. D., & Lockard, A. J. (2012). Clinical validity of the Counseling Center Assessment of Psychological Symptoms-62 (CCAPS-62): Further evaluation and clinical applications. *Journal of Counseling Psychology, 59*, 575-590. doi: 10.1037/a0029855
5. McAleavey, A. A., Nordberg, S. S., Kraus, D., & Castonguay, L. G. (2012). Errors in treatment outcome monitoring: Implications for real-world psychotherapy. *Canadian Psychology, 53*, 105-114.

Six Book chapters, including:

1. McAleavey, A. A., & Castonguay, L. G. (2015). The process of change: Common and unique factors. In O. Gelo, A. Pritz, and B. Rieken (Eds.) *Psychotherapy research: General issues, process, and outcome*. London: Springer.

Scholarly Presentations

Thirty-one Presentations at Refereed Scientific Conferences

Research reviewing experience

Ad-hoc reviewer: *Assessment, Clinical Psychology & Psychotherapy, Clinical Psychology: Science and Practice, Journal of Clinical Psychology, Journal of Counseling Psychology, Measurement and Evaluation in Counseling and Development, Psychotherapy, Psychotherapy Research*