CRITERION DYNAMISM AND GROWTH MIXTURE MODELING: EXPLORING SELECTION ASSESSMENT UTILITY BY IDENTIFYING LATENT CLASSES OF PERFORMANCE CHANGE OVER TIME

by

DARRIN GRELLE

(Under the Direction of Gary Lautenschlager)

ABSTRACT

There is growing consensus among organizational researchers that job performance changes over time, especially in jobs requiring skill acquisition. The nearly ubiquitous finding of declining correlations between selection assessments and performance as time increases calls into question the utility of selection tools. Studying performance cross-sectionally or at the mean level cannot address why predictive validity decreases over time. Latent growth models assume that growth trajectories come from a single population with normal variance around the parameters, but this is likely a naïve assumption. Growth mixture modeling (GMM) does not make this assumption and was used to identify classes of performance change that would be considered desirable by organizations' stakeholders. Objective job performance data for three metrics were collected over nine months for a sample of 203 call center agents. Class membership probabilities were calculated using Mplus (Muthén & Muthén, 2007) for a model in which individual performance trajectories were estimated with assessments of cognitive ability, emotional resilience, sales ability, and conscientiousness as covariates. Across the three metrics, GMM identified three to four classes of growth. Though the assessments did not predict membership in "desirable" trajectories, hypothesized predictor-criterion relationships were supported in some classes and not others. This suggests that GMM is appropriate for identifying subpopulations in which the predictor-criterion relationships do not behave as expected.

INDEX WORDS: Growth mixture modeling, Job performance, Selection, Latent growth modeling

CRITERION DYNAMISM AND GROWTH MIXTURE MODELING: EXPLORING SELECTION ASSESSMENT UTILITY BY IDENTIFYING LATENT CLASSES OF PERFORMANCE CHANGE OVER TIME

by

DARRIN GRELLE

B.A., Saint Louis University, 1997

M.A., American University, 2000

M.S., The University of Georgia, 2006

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2008

CRITERION DYNAMISM AND GROWTH MIXTURE MODELING: EXPLORING SELECTION ASSESSMENT UTILITY BY IDENTIFYING LATENT CLASSES OF PERFORMANCE CHANGE OVER TIME

by

DARRIN GRELLE

Major Professor:	Gary Lautenschlager
Committee:	Charles Lance Robert Vandenberg

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2008

DEDICATION

I would like to dedicate this dissertation to Greg for being patient during this seemingly endless process and giving me love and support when I need it. “I’m almost done!”

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Purpose of the Study

Industrial/Organizational psychologists have been tackling the "criterion problem" since the discipline came into being. From selection, to compensation, to performance appraisal, to training and development, every area of industrial psychology requires the measurement of job performance at some stage. Experts in the measurement of job performance have long debated how to define, measure, and conceptualize performance. One of the largest points of contention among job performance researchers is the stability of job performance over time. The debate between advocates of static versus dynamic criteria was carried out over several decades of research in the twentieth century (Austin & Vilanova, 1992). It has generally been accepted by most researchers in the organizational sciences that job performance is, for the most part, dynamic. Though we may have accepted that performance changes over time, the way in which change is defined is debated and the way it is measured may be lacking.

One of the reasons that criterion dynamism was regarded as a problem is the almost ubiquitous pattern of decreasing correlations between predictors and criteria as the time between the measurements increase. If correlations are decreasing, what does this mean for the utility of predictors used for selection? After years of debate, this question was never really addressed. It is my argument that the statistical procedures used to measure performance change over time were not equipped to answer the question. The purpose of this study was to use growth mixture modeling (GMM), an advanced statistical modeling technique, to model important intraindividual differences in performance change over time.

In addition to achieving a greater understanding of performance change over time in an entry-level job, GMM can be used to provide insight into individual differences in predictor-criterion relationships. Assessments that are designed to select job candidates that are more likely to succeed on the job are based on competencies that are known to be important to the job in question. Despite the effort to use highly predictive assessments, meta-analytic validities for the most common assessment types range from about 0.3 – 0.5 even after correcting for criterion unreliability and range restriction (Schmidt & Hunter, 1998). Our use of selection assessments assumes that they are unidirectionally informative for all job candidates, but the use of GMM might indicate that unobserved individual differences might moderate the relationship between predictors and the growth and change of job performance.

This study seeks to advance the understanding of performance dynamism by uncovering the source of within and between individual differences in performance that earlier studies of job performance using less sophisticated statistical techniques were unable to identify. This study will focus on the measurement of performance for the purpose of validating measures designed for selecting entry-level customer service representatives.

CHAPTER 2

CRITERION DYNAMISM AND GROWTH MIXTURE MODELING: EXPLORING SELECTION ASSESSMENT UTILITY BY IDENTIFYING LATENT CLASSES OF PERFORMANCE CHANGE OVER TIME

The Criterion Problem

Bellows (1941), Thorndike (1949), and others indicated in the earliest years of industrial/organizational psychology that more effort is put into the study of the prediction of performance than the understanding of performance itself. Performance on the job is difficult to capture in its entirety, performance can be viewed through different lenses (Smith, 1976), and the measurement of performance is subject to an array of contaminants resulting from factors beyond an individual's control (Binning & Barrett, 1989; e.g. poor hotel reservation sales in New Orleans after Hurricane Katrina).

As the field of industrial psychology grows, statistical techniques advance, and corporations become increasingly interested in the measurement of job performance, the understanding of job performance has grown (see Austin & Vilanova, 1992 and Austin & Crespin, 2006 for a review). One statistical advance that aided in the understanding of job performance was factor analysis. Bolanovich (1946), for example, used factor analytic methods to determine that six factors described communality among items on an employee rating scale for field engineers. As these methods were applied to the criterion domain, researchers came to understand that job performance was complex and multidimensional and therefore difficult to unify into a composite that made sense rationally and empirically (Toops, 1944).

Ghiselli (1956a) succinctly described three types of dimensionality that affect our ability to understand job performance and measure it accurately. Criteria are multidimensional in that one indicator is typically insufficient to describe job performance. Ghiselli refers to this as “static” dimensionality because at a snapshot in time, an individual can fall somewhere in a multidimensional criteria space with no principle job performance component that underlies all criterion factors. This is contrasted with “dynamic” dimensionality. Criteria are dynamic because an individual can move in the multidimensional criteria space over time for various reasons that will be discussed extensively in the next section. Criterion dynamism was only vaguely defined at this point, but the notion of job performance changing over time sparked a debate among organizational scientists that would last over fifty years. The third type of dimensionality Ghiselli described was “individual” dimensionality. Criteria are dependent on the individual in that two individuals in the same job can be considered “great” for different reasons. Ghiselli believed that in 1956 it was “embarrassing” that satisfactory answers to the problems the three types of dimensionality that he described posed had not been found. This three-part classification of dimensionality was supported by factor analytic work done by Inn, Hulin, and Tucker (1972).

Criterion Dynamism

Early research in the predictor-criterion relationship did not regularly gather criteria data across time (Ronan & Prien, 1966), but as researchers began to understand the importance of longitudinal data, numerous studies faced the issue of rapidly declining validities (e.g. Ghiselli, 1966). In his article that introduced the concept of dynamic criteria, Ghiselli (1956a) defined criterion dynamism as a change in predictor validity over time. Ghiselli along with Haire (1960) later expanded the definition to include a change in the rank order of criteria by strength of relationship with predictors over time and a change in average group performance over time. Researchers that collected performance data over time noticed that for most tasks, not only do predictive validities tend to decrease, but the correlations between measures of performance decrease as the time between the measurements increases (e.g. Fleishman & Hempel, 1954). This phenomenon was first identified by Perl (1934) and has been referred to as the “simplex” pattern.

Two models compete to understand the simplex phenomenon. The changing task model (Fleishman & Hempel, 1954) proposes that an individual’s abilities remain constant, but the structure of the task changes and ability requirements shift over time. The changing subjects model (Alvares & Hulin, 1973) proposes that task requirements remain constant, and the individual gains and develops abilities over time as he or she gains experience and learns. Early tests of these models were largely inconclusive due to lack of sophisticated statistical methodologies and adequate experimental control. A purely task model was ruled out by identifying changes in the rank order of criterion validities over time, but task changes were found to influence temporal change in performance (Dunham, 1974).

The debate over whether criteria are truly dynamic ensued throughout the seventies, eighties, and nineties. Data from one set of researchers would be used as evidence of criteria dynamism (e.g. Fleishman & Hempel), subsequent researchers would reanalyze the data to demonstrate that the criteria are stable after correcting for range restriction or employing some form of statistical control (Barrett, Caldwell, & Alexander, 1985). Fleishman and Mumford (1989) and Austin, Humphreys, and Hulin (1989) in turn commented on Barrett et al’s approach to Fleishman’s data and their conceptual framework.

Much of the debate around whether or not criteria were truly dynamic centered around the way in which dynamism was defined and the statistical methods employed to test theories. Barrett, Caldwell, and Alexander (1985) systematically reviewed the three definitions of criteria dynamism listed above. They felt (as did several others) that defining dynamism in terms of changes in group means was conceptually weak. This method did not consider individual differences in change trajectories or random influences. The second two definitions are tied in that change in individual performance without change in the rank ordering of individuals would not affect validity coefficients. Barrett and colleagues (1985) reanalyzed several studies that found evidence of the simplex pattern and concluded that change in validity coefficient over time could be explained by range restriction and what they refer to as "temporal unreliability" in criteria. Austin and colleagues (1989) claim that temporal unreliability is just another way of saying criteria are dynamic.

Deadrick and Madigan (1990) provided definitions and guidelines for the assessment of criteria dynamism that helped end the debate. They distinguished actual change in individual performance over time ("performance consistency") from changes in evaluation over time and measurement reliability. They suggested that it is impossible to separate issues of reliability from performance consistency in subjective ratings and performance metrics highly influenced by external factors. They suggest the use of non-global, reliable measures of performance in studies of criteria dynamism. Using their strict definitions of criteria dynamism and measurement guidelines, they observed a simplex pattern in the performance of sewing machine operators.

With an increase in better overall research design, measurement, and statistical analysis, evidence has mounted that criteria are dynamic. Most researchers agree that, in general, performance changes over time. The critical next steps in the understanding of job performance involved understanding the nature of intraindividual change, moderators of change, and how performance prediction is affected by criteria dynamism.

Criteria will not always be dynamic. For simpler jobs with a high degree of day-to-day consistency that do not require a broad range of physical and cognitive abilities, variance in performance over time can be mostly explained by situational factors (Rothe & Nye, 1958). Whereas, in most research on the predictive validity of the SAT for college GPA, the SAT demonstrates a decrease in predictive validity over four years of college, Butler and McCauley (1987) found that this was not the case in military academies where courses were regulated for consistency.

Criteria dynamism also appears to be related to job complexity. Rambo, Chomiak, and Price (1983) found that increased task complexity is related to less stability in criteria. Job complexity is also a well known moderator of the relationship between cognitive ability and job performance (Hunter & Hunter, 1984). Murphy (1989) sought to explain this and other key findings in the research on criteria dynamism and performance prediction with a dual-phase model. In his model, which is consistent with aspects of both the changing tasks and changing subjects models, jobs have two main phases: a transition phase in which the job is new or key components have recently changed and a maintenance phase in which new tasks have been learned and key job components are not changing. During transition phases, ability factors are most important for predicting performance. Dispositional factors like motivation and personality are most important for predicting performance during maintenance phases. More complex jobs will have more periods of transition, and therefore will require greater cognitive ability for successful job performance (Hunter, 1986).

Advances in statistical techniques have added new dimensions to the understanding of criterion dynamism. Rather than focusing on the simplex pattern in correlation matrices or looking at changes in the rank order of criterion validities, we can look at the degree to which performance changes over time with autoregressive error and measurement error removed. The most important advance for the understanding of dynamic criteria is the estimation of change over time within each individual and the amount of variance around initial status and change over time across individuals. This is important because "changing rank order over time must, by definition, be the result of different patterns of intraindividual change (Hofmann, Jacobs, & Gerras, 1992, p. 186)." Hofmann and colleagues estimated individual growth trajectories and then identified different classes of growth trajectories for professional baseball players using median splits and comparisons to randomly generated data. Hofmann, Jacobs, and Baratta (1993) advanced the idea of modeling interindividual differences in change functions using Hierarchical Linear Modeling (HLM; Bryk & Raudenbush, 1987) and cluster analysis on insurance salespeople. This methodology identified three clusters of change patterns. Simulated data were used to validate the three clusters identified (Jain & Dubes, 1988). These analyses indicated that three distinct patterns of growth were present.

Deadrick and Madigan (1990) also used HLM to explore individual change patterns in job performance. They determined that for sewing machine operators, psychomotor ability was important for job performance early on and cognitive ability was important for performance improvement over time. Though the measured individual differences in this study explained a significant amount of interindividual variability, 95% of the variance was due to unidentified causes.

Ployhart and Hakel (1998) used latent growth modeling (LGM; Meredith & Tisak, 1990; Willet & Sayer, 1994) on sales figures for securities analysts. Using the LGM framework, Ployhart and Hakel were able to test nested models of intraindividual change over time (e.g. linear, quadratic). They were also able to test the extent to which biodata measures aimed at predicting sales performance and measures of personality explained interindividual variance in latent growth parameters. They identified that biodata and personality explained interindividual variability in the linear and quadratic growth parameters. Despite explaining a great deal of interindividual variance, Ployhart and Hakel found a significant amount of unexplained variance in the final model.

In the studies above, advanced statistical techniques allowed for more robust measurement of performance change over time, and each method had distinct advantages and disadvantages (Bliese & Ployhart, 2002). Muthén and Curran (1997) demonstrated that though the procedures originated in different disciplines, random coefficient modeling and LGM can be shown to be conceptually equivalent and mathematically very similar. Also, growth mixture modeling (GMM), which will be discussed in greater detail in a later section, is equally applicable to both modeling procedures. The primary advantage to using random coefficient modeling was that growth trajectory estimation for each individual did not require that the entire sample have the same number of measurement points or the same time interval between each time point. LGM has become more flexible and no longer requires a rigid time structure that would make it difficult to compare individuals that stay within an organization for different lengths of time (Ho, O'Farrell, Hong, & You, 2006). Sturman and Trevor (2001) found that individuals that stay within an organization have distinctly different growth trajectories from individuals that leave. For this reason, time flexibility in the modeling technique is critical. The advantage to LGM over RCM is the flexibility with which LGM can model non-linear change. LGM also allows for residual variance to be estimated at each time point rather than just around the growth parameters. Because the sample used for this study is in an entry-level call center position for which high turnover is a major problem and change is not expected to be linear (Townsend, 2007), LGM is a more appropriate basis upon which to build GMMs.

The Call Center and Prediction of Call Center Performance

The assessment of performance change over time has been examined in a number of jobs and industries. This study will be focusing on an entry-level customer service position at a call center. Though many of the features of a call center job are very specific to that industry, several key knowledge, skills, and abilities (e.g. customer service, computer knowledge, clerical ability, time management, and multitasking ability; O*NET, 2008) are important for most jobs making the analysis of job performance for this position applicable to other domains.

Four assessments were used to select these individuals: A measure of conscientiousness, a measure of cognitive ability, a measure of customer service skills that includes an emotional resilience component, and a biodata measure focusing on selling skill. The call center position and the assessments used to predict performance in that role provide unique opportunities to study criteria dynamism and its relation to selection assessments. First, most of the performance metrics used to evaluate call center performance are unique to this job (e.g. average handle time, call conversion, and call quality; Levin, 2007). Most call center metrics focus on specific job attributes with strict measurement guidelines. Thus, they are highly reliable and non-global as recommended by Deadrick and Madigan (1990). Sales metrics and measures of average call length are obviously objective metrics, but even the measurement of call quality is mostly objective. CSRs are required to adhere to well defined scripts when dealing with customers. Though general disposition is evaluated, the measurement of call quality is more of a procedural checklist than a subjective measure of overall quality. Every call center differs in how they rank the importance of each metric for evaluating overall employee performance, but what is consistent across call centers is that all collect several (as many as fifteen) highly reliable, objective measures of performance from the moment the CSR logs in to his or her terminal to the time the CSR logs out. For this reason, call center CSRs are ideal for the measurement of performance over time. Three metrics plus tenure data will be examined in this study: average handle time, call quality, and average revenue generated per hour. These metrics will be described in greater detail in the method section.

Second, though call center performance depends on several general abilities (e.g. general keyboard skills, communication skills; Townsend, 2006), the tenacity to endure emotional labor is regarded by many call centers as the only competency of interest (Mulholland, 2002). Call center representatives in all domains (e.g. collections, customer service) must endure long hours of positive emotion maintenance despite the high likelihood of feeling negative emotions when dealing with a hostile customer (Lewig & Dollard, 2003). This separation between felt and displayed emotions is called “emotional dissonance” (Ashforth & Humphrey, 1993). The experience of emotional dissonance is higher among call center representatives than among other entry-level jobs (e.g. manufacturing, administrative, bank teller, and sales; Zapf, Isic, Bechtold, & Blau, 2003) emotional Labor, “the effort, planning and control needed to express organizationally desired emotions during interpersonal transactions," (Hoschschild, 1983) is required to maintain a positive and consistent experience for the customer. Emotional labor is psychologically taxing, however. Emotional labor can lead to job burnout which in turn can lead to decreased job satisfaction and increased turnover intentions (Lee & Ashforth, 1996) and decreased job performance (e.g. decreased call quality, Goldberg & Grandey, 2007; self-reported job performance, Totterdell & Holman, 2003).

An important moderator of the relationship between emotional labor and individual and organizational outcomes is the "depth" of one's emotional response (Grandey, 2003). "Deep-level" responses occur when an individual puts forth cognitive and emotional effort to reframe a negative situation and change the felt emotional response so that the displayed positive response is more genuine and requires less emotional effort. "Surface-level" responses occur when an individual only modifies his or her behavior and not the felt emotion. Surface acting requires more attentional resources to maintain and results in poor job performance in call center customer service jobs (Goldberg & Grandey, 2007). Surface acting also leads to greater emotional dissonance (Totterdell & Holman, 2003). Thus, individuals that are able to respond in a deeper way will have higher job performance over time and experience less emotional dissonance leading to a decreased likelihood of turnover.

The customer service predictor used in this study was designed to measure emotional resilience with items that target the depth of one's emotional response when encountering stressful customer interactions and the tendency to behave in a patient, helpful, and engaging manner when dealing with everyday customer interactions. Though the day-to-day interaction component of the customer service predictor ought to be related to a customer's affective response to a call, the assessment was not meant to predict call handling skills or revenue generated. Therefore it will not predict initial performance based on the metrics collected in this study. Because of the emotional resilience aspect of the measure, however, the customer service assessment ought to predict job performance later in the CSRs tenure. Based on the evidence presented above, if the CSR is unable to handle the stress of emotional dissonance, the CSRs job performance will decline and they will eventually leave the organization.

Hypothesis 1: The customer service assessment will not be significantly related to initial status or linear growth parameters in a one class latent growth model for all criterion variables.

Hypothesis 2: The customer service assessment will be significantly positively related to the quadratic growth parameter (negatively for AHT) in a one class latent growth model for all criterion variables.

When individuals are faced with stressful work situations, the stress can be alleviated through the use of effective coping techniques. Conscientiousness has been shown to be positively related to the use of task-focused coping mechanisms (Matthews et al, 2006). By focusing one's attention on one's work goals, stress can be alleviated by simultaneously getting more work done and eliminating the possibility of non-productive worry. Therefore, individuals high in conscientiousness should be less susceptible to performance decline due to stress.

Hypothesis 3: Conscientiousness will be significantly positively (negatively for AHT) related to the quadratic growth parameter in a one class model for all criterion variables.

Conscientiousness has also been shown to be consistently predictive of job performance across job types and levels (Barrick & Mount, 1991). Also, many of the job performance metrics used in the call center depend on the CSRs' ability to adhere to strict schedule and performance guidelines. Conscientiousness is correlated with dependability and rule adherence (Barrick & Mount), therefore, conscientiousness should predict average handle time and call quality. Murphy (1989) suggests that conscientiousness, though important to job performance in general, is less critical than ability during transition phases. For this reason, conscientiousness is not expected to predict linear rate of change.

Hypothesis 4: Conscientiousness will be significantly related to the initial status parameter of a one class model, but will not be related to the linear growth parameter for all criterion variables.

Third, the ability to multitask is a critical competency (Townsend, 2006). CSRs are required to speak to customers, log information in a computer, look up information on a computer, and send and receive emails at the same time. The ability to multitask is highly correlated with cognitive ability (König, Bühner, & Mürling, 2005). Though Deadrick, Bennet, and Russell (1997) found that cognitive ability did not predict initial status in job performance for sewing machine operators (a job with lower cognitive demand), the cognitive demand of multitasking will require higher cognitive ability early on. The CSRs better able to multitask will be able to handle calls faster (Kantrowitz, Beaty , Grelle, & Wolf, 2008). Therefore, a CSR higher in cognitive ability will be able to handle the cognitive demand of multitasking early on, and have shorter average call lengths than CSRs less able to multitask (shorter calls are desired as the CSR can assist more customers overall if they speak to each for less time). The relationship between cognitive ability and the rate of increase in performance found by Deadrick and colleagues (1997) is expected to hold here.

Hypothesis 5: Cognitive ability will be significantly negatively correlated with the initial status parameter of average handle time (AHT) in a one class model.

Hypothesis 6: Cognitive ability will be significantly negatively correlated with the linear growth parameter of AHT in a one class model.

The relationship between multitasking ability and personality characteristics is not well explored. No relationship has been found between multitasking and extroversion (König, et al, 2005), but the other four Big Five factors have not been linked to multitasking directly. Big Five traits have been linked to polychronicity, the preference for multitasking (Conte & Jacobs, 2003), but a link between multitasking ability and polychronicity has not been found (König, et al, 2005).

No evidence exists in the literature that would suggest that sales ability would be related to average handle time. It is reasonable to predict, however, that a plausible relationship between the two exists. A CSR that is more focused on closing a sale ought to have lower AHT. Because there is no precedent in the literature for making a hypothesis about the relationship between sales ability and AHT, sales ability will be included in an exploratory fashion in a single class model to test an alternative model to determine if sales ability helps to explain changes in AHT over time.

The only relationship not yet explored is that between the biodata measure of sales skill and average revenue generated per hour. The biographical data model is based on the notion that past performance is a good predictor of future performance (Owens, 1976).

Biodata can consist of items referring to events that occurred in the past that can be verified and items that refer to subjective attitudes and beliefs. Biodata as a general construct has been shown through meta-analysis to be correlated with supervisor ratings at $r = .37$ (Hunter & Hunter, 1984). The validity of biodata measures hinges on the fact that if an individual has demonstrated that they have engaged in a desired behavior in the past, they will be likely to engage in that behavior on the job. A biodata measure designed to predict desired behaviors across sales jobs was shown to have high validity based on correlations with supervisor ratings (Stokes, Toth, Searcy, Stroupe, & Carter, 1999). Ployhart and Hakel (1998) determined that a similar sales focused biodata measure predicted securities dealers' initial performance, but not the rate at which they improved. For call center representatives, total revenue generated is a function of both sales ability and the number of customers they can help an hour. Therefore, CSRs with lower average handle time will be able to help the most people. Selling skill is generally considered to be a stable trait. For this reason, the sales skill biodata measure used in this study is expected to predict CSRs sales at the beginning of their tenure, but the improvement in sales will be a function of cognitive ability as it is expected to predict average handle time.

Hypothesis 7: The biodata measure of selling skill will be significantly positively related to the initial status parameter of revenue generated per hour in a one class model.

Hypothesis 8: Cognitive ability will be significantly positively correlated to the linear growth parameter of revenue generated per hour in a one class model.

Growth Mixture Modeling

As measurement specificity increases, unexplained variance will decrease. Assessing job performance cross-sectionally at a specific time point is very non-specific and can be subject to many sources of error. One can account for some of this error by taking a broader cross section of data and averaging it together, but this method does not indicate the extent to which an individual changed (for better or worse) over time (Barrett, Caldwell, & Alexander, 1985). A best fitting line can account for this, but significant intraindividual variability in rates of change can mask this effect. Creating a best fitting line also assumes that the specified change function (linear to cubic) is the same for everyone. Incorporating LGM or RCM into a larger model with predictors can help explain some of the individual differences in performance change, but they cannot account for unmeasured sources of intraindividual variation (Deadrick, Bennett, & Russell, 1997; Ployhart & Hakel, 1998). Ghiselli (1956b) demonstrated that in a cross sectional study of taxi driver performance, two predictors that had little predictive utility alone or when combined could still show utility. If one predictor was used to screen out some individuals, the correlation between the other predictor and job performance for the remaining population was much higher than it was for the whole population. A similar situation might occur when predicting performance growth. Some predictors may only be valid for certain subpopulations that cannot be identified with LGM or RCM alone.

To date, the focus of the dynamic aspect of the criterion problem is the concern that prediction for selection purposes may lack validity if changes in rank order are occurring. Although the debate had raged for several decades now and research has demonstrated that traditional predictors vary in their ability to predict initial performance versus performance change (e.g. Ployhart & Hakel, 1998; Sturman & Trevor, 2001; Thoresen, Bradley, Bliese, & Thoresen, 2004), no work has been done on how the utility of selection tools would be affected by dynamic criteria. The utility of selection measures is affected by the variability in job performance of the applicants as compared to the variability in performance of those hired for the job. A valid predictor should restrict the range in job performance as compared to randomly selecting from the applicant population (Schmidt & Hunter, 1998). In practical terms, a predictor should select individuals that would be consistently high performers over time as compared to inconsistent performers or individuals that are consistently poor performers.

Using Murphy's (1989) theory of maintenance versus transition, Kanfer and Ackerman's (1989) theory of job performance change, and Schmidt and Hunter's (1998) meta-analysis of selection method utility, let us consider an example to illustrate how not considering criteria dynamism might collude estimations of utility. Three individuals take three pre-employment assessments. When their scores are summed, their scores meet the minimum required by the organization. In this hypothetical organization, revenue generated is the only performance metric. Individual A has high selling skill, high cognitive ability, and is high in conscientiousness and emotional resilience. Individual A measures consistently high in job performance over time but does not increase much. Individual B has low selling skill, is low in cognitive ability, and is high in conscientiousness and emotional resilience. Individual B performs poorly early on. Although he improves slightly over time, he lacks the ability to become a good performer. Individual C has low selling skill, is high in cognitive ability, and is low in conscientiousness and emotional resilience. Individual C is a poor performer early on, but improves dramatically as she learns to multitask. Once she has learned the job, however, she becomes unable to handle the emotional dissonance due to low stress tolerance and emotional resilience and leaves the organization. If these three individuals are taken together, an inaccurate picture of predictor validity and performance consistency might be drawn. Ability would have medium validity looking just at initial status, high validity at the midpoint, medium validity at a later time point, and high validity if performance were averaged across all time points. Job experience would have high validity at initial status and later time points, but medium validity at middle time points and averaged over time. Personality would have low to medium validity across all time points and averaged over time.

Based on the hypothetical situation above, an organization may choose to use ability and job experience to hire employees. This would be problematic because there would only be a small window in which two of the employees were performing at a high level and no time when all three were satisfactory. If LGM were used to analyze the data, one would see significant variability around the initial status parameter that could be explained by previous job experience. The mean slope across the three individuals would be positive with variability mostly explained by cognitive ability. Because only one individual had a curved growth trajectory, an overall quadratic effect would probably not be seen. Again, even using an advanced analysis technique, conscientiousness and motivation appear to lack predictive validity.

One of the assumptions of standard LGMs or RCMs is that the population's growth parameters are homogenous. Both models allow for the inclusion of exogenous variables that can explain variability around the mean growth parameters. Multiple group growth modeling (Muthén, 1989) allows for the inclusion of categorical grouping variables. In these models, the growth parameters and all other model parameters are permitted to differ among groups. One can hold parameters constant across groups and test nested models with progressively rigid equality restrictions. These models, however, require an observable grouping variable that is known prior to data analysis.

In the present study, for reasons described above, different patterns of performance change over time were expected that could not be completely explained by the exogenous predictor variables or an observed categorical variable. In this case, the grouping variable is unobserved. Growth mixture modeling (GMM; Muthén, 2001) allows one to estimate model parameters for a number of groups or "classes" for which the grouping variable is unobserved or not recorded for a portion of the sample. In growth mixture modeling, a categorical grouping variable is replaced by an estimation of the likelihood a case is present in each class. Each case's contribution to the estimation of model parameters for a given class is dependent upon the estimated likelihood that the case is in that class. In GMM, variability around parameter estimates within each class is permitted allowing the inclusion of exogenous predictor variables in the model. Therefore, it is possible to identify "desirable" growth trajectories for CSRs and assess the differences in the strength of the relationship among the exogenous variables and the growth parameters in the desirable and undesirable classes. Comparing desirable to undesirable classes, if the difference between the group means for a predictor is not statistically different, then that predictor lacks utility. In the above example, previous job experience predicted initial status for job performance. It does not, however, predict how well a CSR will handle the stress of the job. Therefore, two classes could be seen: one with consistently high job performance and another with high performance that quickly drops. The former is obviously preferable, but if one were to compare the mean scores on the job experience measure between the two classes, no difference would be seen.

In the measurement of performance change over time, I have demonstrated that several growth trajectories are possible for entry level CSRs based upon different competencies and reactions to job stressors. GMM has not been applied to data in the organizational sciences, and this is an optimal opportunity to demonstrate the power of GMM for better understanding performance change over time, achieve a greater understanding of the validity and utility of predictors used for selection, and present a new technique to organizational researchers. Due to the multidimensional nature of the criteria in performance in this study and performance in general, it is difficult if not impossible to predict how many different classes of growth and change might be revealed when using GMM on these data. For this reason, no specific hypotheses are proposed, but multiple classes are expected.

CHAPTER 2

METHOD

Participants

The participants in this study are call center customer service representatives (CSRs) at a large telecommunications company. The CSRs work in an "in-bound" call center in which customers call the company to set up service. The primary goals of the CSRs are to process calls in an efficient manner and encourage customers to register for more products and services. The data were collected as part of a predictive validity study conducted by a selection system design company. Selection assessment data were available for 4,870 job applicants from across the United States of America. Of these, 1,459 were hired and began work. Demographic information for the applicant and matched samples are in Table 1. Though 1,459 individuals were hired and began work, a multitude of factors affected their inclusion in this study. Some of the sites were inconsistent in their record keeping and did not provide enough data for individuals at those sites to be included. Many of the CSRs in the study were hired using the predictors in question, but were hired too long before the beginning of the criterion-related validity study to be included. They had been on the job too long before performance data collection began to assess their growth trajectories. Finally some cases were excluded due to improbable or impossible values on some of the performance criteria (e.g. negative values for AHT).

Sixty-six percent was selected as the minimum proportion of usable data present for inclusion in data analysis. This value was chosen because full information maximum likelihood imputation (FIML) was used and FIML requires a minimum covariance coverage of 10% valid cases per cell. For a CSR that worked for three months, at least two months of performance data had to be available in order to be included in data analysis. Because of considerable variability in the length of time each CSR remained on the job, the number of missing values determined to be acceptable varied from person to person. After all unusable cases were excluded, the final number of CSRs used in each analysis ranged from 173 to 201 people.

Table 1. Sample Demographics for Applicant and Matched Samples

ETHNICITY	N	%	GENDER	N	%	AGE GROUP	N	%
Applicant Sample								
Asian/Pacific Islander	212	4.8%	Female	2563	58.1%	< 18	6	0.1%
African-American	1220	27.7%	Male	1776	40.3%	18-20	623	14.1%
Hispanic	634	14.4%	Null	69	1.6%	21-39	2987	67.8%
Native American	27	0.6%				40-50	497	11.3%
Caucasian	2128	48.3%				> 50	200	4.5%
Other	81	1.8%				Null	95	2.2%
Null	106	2.4%						
Total	4408		Total	4408		Total	4408	
Matched Sample								
Asian/Pacific Islander	60	4.1%	Female	831	57.0%	18-20	239	16.4%
African-American	348	23.9%	Male	602	41.3%	21-39	973	66.7%
Hispanic	155	10.6%	Null	26	1.8%	40-50	153	10.5%
Native American	7	0.5%				> 50	56	3.8%
Caucasian	824	56.5%				Null	38	2.6%
Other	23	1.6%						
Null	42	2.9%						
Total	1459		Total	1459		Total	1459	
Usable Sample								
Asian/Pacific Islander	14	6.9%	Female	124	61.1%	18-20	33	16.3%
African-American	42	20.7%	Male	77	37.9%	21-39	138	68.0%
Hispanic	41	20.2%	Null	2	1.0%	40-50	16	7.9%
Native American	0	0.0%				> 50	11	5.4%
Caucasian	98	48.3%				Null	5	2.5%
Other	3	1.5%						
Null	5	2.5%						
Total	203		Total	203		Total	203	

Materials

Predictor Variables. Four assessments were administered to applicants for the CSR position. The four assessments were chosen based on a job analysis. Selection decisions were based on a weighted score composite of the four assessments. Due to the proprietary nature of the assessments used, some detail has been omitted.

Cognitive ability was assessed using a 20-item timed test in which applicants are asked to view a figure containing different pieces of information. Applicants were then asked questions that required them to process the information they were provided to determine which of five options was correct. An example of the type of figure an applicant might see is the call record of a fictional customer of a credit card company that uses a fictional coding scheme. Questions would then refer to specific dates or events that occurred, and the applicant would be required to find the event and use the key provided to decode what occurred (e.g. “How many times did the customer call to check her balance?”). The item design is intended to be face-valid for individuals applying for entry-level customer service positions.

Conscientiousness was assessed using a 30-item Likert-type scale measure. Most of the items are face-valid items that focus on work situations an entry-level call center representative would encounter (e.g. returning from breaks on time, following instructions, complying with company policy). This measure has been shown to have convergent validity with global measures of conscientiousness, and has been correlated with measures of schedule adherence and supervisor ratings of dependability in entry-level call center positions.

The customer service skill measure is a 60-item true/false assessment that determines the extent to which an applicant will show persistent enthusiasm when interacting with customers, will apologize sincerely for inconveniences, be patient, tolerate rude customers calmly, and search for information or products for customers. This measure was selected because it assesses the tendency to respond to negative customer situations in a positive way. It has been shown to predict supervisor ratings of customer service in entry-level customer service positions.

Sales skill is assessed in a true/false biodata format. Applicants are asked to respond true or false to statements about selling situations that have occurred in the past or hypothetical situations that could occur in the future. The 60-item measure is intended to predict the likelihood that applicants will suggest or show alternative solutions based on customer needs, direct conversation toward a commitment/order/sale, show confidence even after a hard refusal/rejection, and strive to close a transaction every time. This assessment has been shown to predict revenue generated by entry-level call center agents and supervisor ratings of selling ability, confidence, and persistence.

Criterion Variables. The organization in which this study was conducted collects numerous performance metrics. Not all were used in this study, however. First, because each of the performance metrics is going to be assessed individually, it would become overwhelming to examine all of them. Also, many of the performance metrics are not independent of one another. For example, the organization collects data on revenue generated per hour, revenue generated per call, and revenue generated per order. Three metrics averaged over the course of each full month on the job and tenure were selected for analysis. Data were collected over the course of eleven months, but only 21% of the CSR population remained employed for longer than nine months. For this reason, only the first nine months of employment were analyzed.

Average handle time (AHT) is the most common performance metric collected by call centers (Levin, 2007). AHT is the amount of time a CSR spends on the phone for each call. AHT should be low so the CSR can handle more customers per hour. AHT can be influenced by policy compliance (moving away from the script), ability to enter customer information into the computer quickly and accurately, and assertiveness with the customer (Levin, 2007). AHT is reported in seconds.

Revenue per call (RPC), as the name suggests, is a function of the number of services the CSR is able to get the customer to sign up for during the call. Customers have a wide variety of services from which they can choose. An effective CSR will be able to guide the customer through these options quickly and efficiently while simultaneously encouraging the customer to add services. RPC is reported in dollars/per call.

The third criterion variable is call quality. This organization collects call quality ratings at random intervals throughout the day. CSRs are rated by supervisors on the extent to which they adhere to an established script laid out by the organization, provide accurate information to the customer, collect and code information accurately, and engage the customer in a courteous manner. Call quality is reported as a percent score determined by the number of “quality points” earned divided by the number of quality points possible. Call quality data are not typically gathered during a CSRs first month on the job. Call quality analyses will begin at month two.

Tenure is the fourth criterion variable. Each CSR began work at different times and stayed for different lengths of time. Data for all CSRs was tenure equated such that “Time 0” was the first day on the job regardless of actual start date. Due to the nature of the industry, no seasonal effects with respect to the criteria were expected. This was empirically verified. Many of the CSRs hired during the duration of the study were still employed when data collection ended. Therefore, the accuracy of a tenure variable measured as days on the job was confounded by the number of days each CSR had the opportunity to turnover. A "days of employment" variable would be inaccurate. For this reason, tenure was measured as a series of dichotomous variables. At the end of each month on the job, the CSR was coded as 0 if they turned over or 1 if they stayed. Therefore, tenure was measured by a total of 9 dichotomous variables. If a CSR was still employed at the end of data collection, they received a code of 1 for each month they were on the job and no code for all remaining months of the study. If they were terminated or left the job during the duration of data collection, they received a 1 for every full month they worked and a 0 for all remaining months of the study. Those CSRs that stayed and those that left were treated differently because if a CSR was still employed at the end of data collection, it is unknown how long they would remain employed after data collection was over. If a CSR left during the course of the study it was known that they would remain "turned over" for the rest of the nine-month period studied (rehires were excluded from analysis to avoid potential confounds). Tenure was not modeled using the GMM technique due to the measurement complexity involved. The method used to include tenure in these analyses is described in the data analysis portion of this document.

Procedure

The organization that developed the assessments used in this study made arrangements with the telecommunications organization for which the CSRs worked to collect predictor and criterion data to accumulate validity evidence on the assessments used. Applicants interested in the CSR position were asked to go to the call center to take the online assessments in a proctored environment. The assessments were administered on the same computers used at call terminals via the web. The cognitive ability portion of the assessment battery was timed, but all other portions had no time limit. Although a top down selection method was recommended, the organization chose to interview several individuals with lower predictor scores. The final selection decision for each individual was based on a structured interview. The organization did not provide the data from the structured interview.

After selection, CSRs participated in two weeks of training. Tenure was measured from the moment they started on the job post training. All performance metrics were collected daily at all times the CSR was logged on to his or her terminal and were averaged at the end of each month.

Data Analysis

Mplus 5 (Muthén & Muthén, 2007) was used to conduct all analyses. This program allows for the estimation of LGM and GMM models and provides several fit indices appropriate for each type of analysis. The three criteria in this study were analyzed separately rather than creating a latent composite criterion variable.

A simple to complex model building process was used to first test all hypotheses and then address GMM research questions. The simplest models were linear growth models with no covariates. This was done to assess the extent to which the performance metric changes over time and measure the magnitude of the variability around the initial status and linear growth parameters. The next model was a quadratic growth model. As the linear model is nested in the quadratic model, a chi-square difference test was used to determine which function better fit the data.

Once the general shape of growth models for each variable was determined, the predictor variables were added to the model. The initial status and each growth parameter were regressed on each predictor variable to test the hypothesized relationships. These analyses were used to determine the extent to which the predictor variables explain variability around each growth parameter and to measure the magnitude of the remaining variability around each parameter. The appropriateness of these model modifications were determined by conducting a chi-square difference test and examining the parameter estimates.

Growth mixture models were estimated next. For all the GMMs estimated in this study, all variance components were estimated, but set to be equal across the specified number of classes. This was done to reduce the computational complexity of estimating multiple variance components per class. Also, given the sample size, allowing the variance components to be freely estimated would drastically reduce the degrees of freedom available.

The first GMM in the series was a model with no covariates. Nylund, Asparouhov, and Muthén (2007) suggest identifying the number of classes in a GMM with no covariates before continuing to a model with covariates added. For AHT, however, the GMM analyses with no covariates indicated that only one-class was present in the data, but conditional analyses were conducted anyway to determine if the regression estimates differed across classes. Ideally, a holdout sample would be used to determine the number of classes, and the remainder of the sample would be used to cross validate the results. In this study, the samples would be too small to generate estimates in which one could be confident. The primary complication in GMM is that, though likelihood of class membership is estimated, the number of classes is not and must be specified. In most cases, researchers will estimate several models with different numbers of classes. The standard fit indices used to judge fit in structural equation modeling are inappropriate in GMM as they assume that the data are drawn from a single population. The Akaike's Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwartz, 1979) are commonly used to compare different GMM models because they are based on likelihood ratios and do not require that the models being compared to be nested. Both the AIC and the BIC reward models that more accurately reproduce the observed data, but punish for lack of parsimony. For this reason, Bauer and Curran (2003) determined that these criteria often guided the selection of models specifying multiple classes over those with only one true class (based on simulated data) when the sample data were not normally distributed. The nesting of GMM models does not follow traditional SEM format (see Lo, Mendell, & Rubin, 2001 for a review), and the likelihood ratios do not follow a chi-square distribution. Therefore, the Lo, Mendell, and Rubin likelihood ratio test (LMR-LRT) was developed to compare a model with k classes to one with k-1 classes. The problem with the LMR-LRT test is that it assumes a within-class normal distribution. Therefore, it is impossible to compare a one versus a two class model in which the one class model has a nonnormal distribution. Muthén (2003) suggests the use of within class skewness and kurtosis tests (Nylund et al, 2007) that compare the data generated by the model to the sample data in addition to the LMR-LRT to determine the appropriate number of classes. Unfortunately, the skewness and kurtosis tests are not available for analyses with missing data. As an alternative, Mplus can generate a bootstrap likelihood ratio test (BLRT) to generate a confidence interval around the likelihood ratio and that can be used to compare a model with k classes to one with k-1 classes. A significant p-value (less than 0.05) in this test indicates that the null hypothesis (the k-1 classes is more appropriate) is rejected.

Nylund and colleagues found that the BLRT identified the correct number of classes in a simulation study more often than any other technique. They found that the BIC was the next most successful indicator. The data in that study were simulated, however, and therefore had no true meaning and the true number of classes was known. Based on the analyses of Nylund and colleagues and the work of Burnham and Anderson (2004), when making decisions on the number of classes, the BLRT and AIC were weighted heaviest followed by the LMR and the BIC. In these analyses, clear "winners" were not always clear and the use of professional judgment and the evaluation of the plausibility of outcomes were necessary. The AIC, BIC, LMR, and BLRT statistics are all provided by the Mplus program and were used in conjunction with professional judgment to determine the appropriate number of classes.

In addition to the challenge of determining the number of classes, GMM is a computationally complex iterative process that can often converge on local solutions. When this occurs, the optimal solution is not identified and faulty conclusions can be drawn from the results. As a way to remedy this situation, the Mplus program generates random starting values and identifies a model multiple times. If the best log likelihood generated repeats multiple times, one can have confidence that the ideal solution has been reached. For each analysis, 200 random starts with 40 iterations each were used to derive a solution for a specified number of classes. As the number of classes being specified increased, the frequency of the optimal log likelihood went down. In these situations, different starting values were specified for the model growth parameters. The solutions were compared. If the number of individuals in each class changed after specifying different starting values, this was considered a local solution and a different set of starting values was used. If, after four iterations of random and specified starting values a consistent model could not be identified, the random starts were increased to 500 with 80 iterations each.

After identifying the number of classes for each criterion, GMMs were specified that included the predictor variables. The inclusions of the predictors in the GMM yields estimates of the influence of each predictor on the likelihood of membership in a particular class as well as providing independent assessments of the influence of each predictor on the initial status and change vectors within each class. All hypothesized predictor to growth parameter relationships were included in these models even if no support for a hypothesized relationship was found in the single group LGMs. This was done to test the possibility that the hypothesized relationship only holds for some classes. The default option in Mplus is to restrict the beta estimates to be equivalent across all classes to reduce computational demand, but this option was overridden and the regression coefficients were estimated for each class independently. This, however, prevented the ability to regress class membership on the predictor variables in a multinomial regression because the models became underidentified when these regressions were added to the model.

To explore the possibility that class membership was a function of one's mean level on the predictors, simple one way ANOVAs were conducted to test the mean difference on each predictor by class membership.

The nature of the tenure variable makes it difficult to incorporate into any of the above models. Survival analysis is a method of estimating the likelihood of a non-repeatable event occurring within a specific time period. In this case, the survival function would reflect the probability that a CSR would remain employed at the end of the study. Because the event is (in theory) non-repeatable, the probability can only increase or remain the same. Some studies have explored the option of integrating survival analysis into GMM (Muthén & Masyn, 2005), but the survival function can be independently estimated for each empirically derived class. These analyses were attempted with this data, but because the sample sizes were not very large and a fair amount of data was missing, no models converged to a permissible solution.

CHAPTER 3

RESULTS

Initial Analyses

Before any calculations could be done, the data had to be tenure equated. The eleven months of data that were available included individuals with a wide range of time with the organization. Several individuals had been with the organization for several months before data collection began and were therefore excluded. Though the organization was very

Table 2. Intercorrelations among Predictors and Average Handle Time

Variable Name	1	2	3	4	5	6	7
1. Cognitive Ability	1.000						
2. Conscientiousness	-0.047	1.000					
3. Customer Service	-0.021	0.265**	1.000				
4. Sales Ability	0.046	-0.136**	0.303**	1.000			
5. AHT Month 1	-0.300**	-0.081	0.065	-0.120	1.000		
6. AHT Month 2	-0.207**	0.119	0.043	-0.218**	0.752**	1.000	
7. AHT Month 3	-0.133*	0.105	0.059	-0.090	0.562**	0.845**	1.000
8. AHT Month 4	-0.157*	0.186**	-0.001	-0.039	0.477**	0.767**	0.890**
9. AHT Month 5	-0.120	0.228**	-0.034	-0.063	0.335*	0.660**	0.805**
10. AHT Month 6	-0.181*	0.143	-0.057	-0.030	0.142	0.571**	0.705**
11. AHT Month 7	-0.225**	0.148	-0.076	0.000	0.255	0.578**	0.652**
12. AHT Month 8	-0.096	0.121	-0.038	-0.079	0.306	0.531**	0.597**
13. AHT Month 9	0.009	0.130	-0.046	-0.120	0.556*	0.542**	0.529**

Variable Name	8	9	10	11	12	13
1. Cognitive Ability						
2. Conscientiousness						
3. Customer Service						
4. Sales Ability						
5. AHT Month 1						
6. AHT Month 2						
7. AHT Month 3						
8. AHT Month 4	1.000					
9. AHT Month 5	0.910**	1.000				
10. AHT Month 6	0.831**	0.890**	1.000			
11. AHT Month 7	0.756**	0.762**	0.881**	1.000		
12. AHT Month 8	0.675**	0.640**	0.745**	0.873**	1.000	
13. AHT Month 9	0.586**	0.533**	0.634**	0.730**	0.902**	1.000

*p < .05; **p < .01

concerned with collecting a broad range of performance data on the CSRs, the reporting of the data was flawed. Many individuals had missing data points. Considering tenure and the available data range, individuals with more than one third of their data missing were excluded. Missing data was imputed using full information maximum likelihood (FIML), but having more than a third of an individual's data imputed was undesirable. FIML uses all of the available observed data including means, variances, and covariances to estimate values for the missing information. FIML assumes that the data are missing at random (Little & Rubin, 2002). Data are considered missing at random if the pattern of missingness is not dependent on the values of the missing data. Data are missing because the organization did not keep complete records. Missingness had nothing to do with any criterion values that were provided. After excluding unusable cases, between 128 and 201 individuals were included in these analyses. Smaller ranges of data in terms of the number of months analyzed were considered, but almost all of the individuals that completed nine months on the job had data for months eight and nine. This meant that reducing the date range did not increase the numbers of usable cases, so nine months was considered the best balance between sample size and opportunity to observe change.

It was mentioned above that the debate of the criterion dynamism began in part by the nearly ubiquitous simplex pattern. To determine if the simplex pattern appeared in these data as well, a simple bivariate correlation matrix with pairwise deletion was generated for each criterion. These correlations are shown in tables 2 - 4. Generally, as the time between measurement occasions increases, the correlations decrease. The pattern of decreasing validities, however, is more complex. For Call Quality, the predictors in this study do not predict performance at any point in the nine months of data studied, so no discernable pattern

Table 3. Intercorrelations Among Predictors and Call Quality

	Variable Name	1	2	3	4	5	6	7
1.	Cognitive Ability	1.000						
2.	Conscientiousness	-0.047	1.000					
3.	Customer Service	-0.021	0.265**	1.000				
4.	Sales Ability	0.046	-0.136**	0.303**	1.000			
5.	Quality Month 2	0.039	0.165	0.114	-0.060	1.000		
6.	Quality Month 3	0.005	-0.021	0.046	-0.007	0.423**	1.000	
7.	Quality Month 4	0.030	0.011	-0.025	-0.056	0.567**	0.577**	1.000
8.	Quality Month 5	0.064	0.127	-0.091	-0.079	0.464**	0.464**	0.541**
9.	Quality Month 6	0.024	0.018	-0.127	-0.029	0.285	0.294**	0.411**
10.	Quality Month 7	-0.031	0.075	-0.091	-0.041	0.214	0.377**	0.424**
11.	Quality Month 8	-0.134	-0.042	-0.097	-0.148	0.308	0.204	0.294*
12.	Quality Month 9	-0.108	0.086	0.042	-0.021	0.459	0.297	0.603**

	Variable Name	8	9	10	11	12
1.	Cognitive Ability					
2.	Conscientiousness					
3.	Customer Service					
4.	Sales Ability					
5.	Quality Month 2					
6.	Quality Month 3					
7.	Quality Month 4					
8.	Quality Month 5	1.000				
9.	Quality Month 6	0.404**	1.000			
10.	Quality Month 7	0.415**	0.404**	1.000		
11.	Quality Month 8	0.389**	0.458**	0.315*	1.000	
12.	Quality Month 9	0.512**	0.523**	0.555**	0.309*	1.000

*p < .05; **p < .01

should be present. There is a general decline in correlations between AHT and cognitive ability, conscientiousness, and sales ability. There is no relationship between AHT and customer service skill. There is no relationship between Revenue per Call and cognitive ability and conscientiousness, but the correlations between Revenue per Call and customer service skill and sales ability appear to increase rather than decrease.

Average Handle Time

Average handle time was examined first. In LGM, growth is modeled as a function of a latent initial status variable on which all criterion variables load with factor loadings fixed at one and a latent slope variable on which all criterion variables load with factor loadings fixed at zero and one for times one and two. For the latent slope variable, one has the option

Table 4. Intercorrelations Among Predictors and Revenue per Call

	Variable Name	1	2	3	4	5	6	7
1.	Cognitive Ability	1.000						
2.	Conscientiousness	-0.047	1.000					
3.	Customer Service	-0.021	0.265**	1.000				
4.	Sales Ability	0.046	-0.136**	0.303*	1.000			
5.	RPC Month 1	0.091	0.036	0.014	0.073	1.000		
6.	RPC Month 2	0.111	-0.041	-0.015	0.097	0.965**	1.000	
7.	RPC Month 3	0.044	-0.010	0.017	0.090	0.884**	0.828**	1.000
8.	RPC Month 4	0.039	-0.044	0.066	0.084	0.728**	0.768**	0.916**
9.	RPC Month 5	0.008	-0.103	0.070	0.074	-0.011	0.860**	0.909**
10.	RPC Month 6	0.085	-0.171*	0.112	0.125	-0.168	0.689**	0.834**
11.	RPC Month 7	0.121	-0.084	0.151	0.078	-0.125	0.614**	0.732**
12.	RPC Month 8	0.062	-0.144	0.171	0.215*	-0.123	0.576**	0.734**
13.	RPC Month 9	0.053	-0.091	0.164	0.227*	-0.028	0.428**	0.576**

	Variable Name	8	9	10	11	12	13
1.	Cognitive Ability						
2.	Conscientiousness						
3.	Customer Service						
4.	Sales Ability						
5.	RPC Month 1						
6.	RPC Month 2						
7.	RPC Month 3						
8.	RPC Month 4	1.000					
9.	RPC Month 5	0.933**	1.000				
10.	RPC Month 6	0.883**	0.943**	1.000			
11.	RPC Month 7	0.781**	0.817**	0.819**	1.000		
12.	RPC Month 8	0.782**	0.824**	0.828**	0.957**	1.000	
13.	RPC Month 9	0.679**	0.730**	0.742**	0.936**	0.944**	1.000

*p < .05; **p < .01

of estimating the shape of the curve by estimating the factor loadings of the criteria on the slope factor, or one can fix the factor loadings to equally spaced intervals that represent the number of data collection points. Due to the complexity of the analyses, AHT was modeled using the latter format such that the factor loadings of the criterion variables on the slope factor were fixed to equally spaced intervals going from zero to eight. The overall fit of the model was assessed using the χ^2 statistic, the Tucker-Lewis Index (TLI), the Root Mean Squared Error of Approximation (RMSEA), and the standardized root mean square residual (SRMR). In the linear LGM, the initial status parameter significantly differed from zero, but the linear slope did not. Both residual variance components, however, indicated that a significant amount of unexplained variance in the parameters was present. Fit for this model was poor $\chi^2(76, N = 201) = 430.931$, $p < .001$ (TLI = 0.776, RMSEA = 0.152, SRMR = 0.260). The predictors were included in the estimation of this model, but were only allowed to covary with each other. These variables were included in the model so that nested comparisons of the simple to complex LGMs could be made using chi-square difference tests. The inclusion of a quadratic growth term resulted in a dramatic improvement in fit $\Delta\chi^2(4, N = 201) = 165.786$, $p < .001$, but the overall fit of the model was still poor (TLI = 0.871, RMSEA = 0.116, SRMR = 0.129). Also, the initial status, linear, and quadratic means all significantly differed from zero. According to the growth parameter estimates, CSRs have a fairly even AHT until their fifth month when their calls start to get longer. A significant amount of unexplained variance surrounded the mean of all three growth parameters in the model.

Because of the significant improvement in fit with the addition of the quadratic component, the quadratic model served as the baseline when including the predictor variables in the model. The initial status, linear, and quadratic parameters were each regressed on conscientiousness, customer service skill, and cognitive ability. Customer service skill did not predict initial status ($\beta_{IS.CS} = 1.496$, $SE_{IS.CS} = 2.451$, $p = .476$) or linear change ($\beta_{CH.CS} = -0.345$, $SE_{CH.CS} = 1.219$, $p = .390$) as was predicted in hypothesis 1. Customer service also did not predict quadratic change ($\beta_{Q.CS} = -.098$, $SE_{Q.CS} = 0.156$, $p = .132$). Hypothesis 2 was not supported. Conscientiousness had a significant negative effect on the quadratic growth parameter ($\beta_{Q.Cons} = -0.566$, $SE_{Q.Cons} = 0.207$, $p < .01$) as was predicted by hypothesis 3. Hypothesis 4 proposed a negative relationship between conscientiousness and the initial status and linear growth parameters. Conscientiousness was unrelated to initial status ($\beta_{IS.Cons}$

$= 2.807$, $SE_{IS.Cons} = 3.279$, $p = .195$) but had a significant relationship in the opposite direction for the linear growth parameter ($\beta_{CH.Cons} = 3.640$, $SE_{CH.Cons} = 1.626$, $p = .013$). Hypothesis 5 was supported in that cognitive significantly predicted a lower initial status on AHT ($\beta_{IS.CA} = -3.678$, $SE_{IS.Cons} = 1.764$, $p = .018$). Cognitive ability was not, however, related linear or quadratic growth ($\beta_{CH.CA} = 0.023$, $SE_{CH.CA} = 0.881$, $p = .492$; $\beta_{Q.CA} = 0.036$, $SE_{IS.Cons} = 0.113$, $p = .484$). Despite the inclusion of several nonsignificant paths, this model was a significant improvement in fit over the quadratic growth model with no paths between the predictors and growth parameters $\Delta\chi^2(18, N = 201) = 48.983$, $p < .001$, but the overall fit of the model was still not good (TLI = 0.881, RMSEA = 0.122, SRMR = 0.118).

Though no hypothesis was proposed, it was postulated that a relationship between sales ability and AHT might exist. To test this, a model was generated from the previous model by replacing customer service skill with selling ability in the three regression equations. Customer service skill was removed because it was shown to have no relationship with the growth parameters. The relationships between conscientiousness and cognitive ability and the three growth parameters remained relatively the same ($\beta_{IS.Cons} = 3.285$, $SE_{IS.Cons} = 3.088$, $p = .145$; $\beta_{CH.Cons} = 3.428$, $SE_{CH.Cons} = 1.545$, $p = .013$; $\beta_{Q.Cons} = -0.594$, $SE_{Q.Cons} = 0.198$, $p = .013$; $\beta_{IS.CA} = -3.380$, $SE_{IS.CA} = 1.740$, $p = .014$; $\beta_{CH.CA} = 0.084$, $SE_{CH.CA} = 0.872$, $p = .460$; $\beta_{Q.CA} = 0.026$, $SE_{Q.CA} = 0.112$, $p = .488$). Sales ability was significantly related to initial status ($\beta_{IS.SA} = -5.524$, $SE_{IS.SA} = 2.518$, $p = .014$). Though sales ability was expected to negatively relate to the linear parameter, sales ability was significantly positively related to linear growth ($\beta_{CH.SA} = 2.113$, $SE_{CH.SA} = 1.261$, $p = .047$). Sales ability did not predict quadratic growth ($\beta_{Q.SA} = -0.256$, $SE_{CH.SA} = 0.163$, $p = .018$). This model (conscientiousness, cognitive ability, and sales ability) and the one previous (conscientiousness, cognitive ability, and customer service skill) are not nested and could not be compared directly. This model could be compared to the baseline quadratic model, however. This comparison indicated that this model, too, was a significant improvement in fit $\Delta\chi^2(18, N = 201) = 35.952$, $p < .01$, but the general fit of this model was also still not good (TLI = 0.872, RMSEA = 0.127, SRMR = 0.123).

The first GMM estimated was a one-class model to serve as a baseline against which the two-class model could be compared. The parameter estimates are equal to those of the standard LGM above with linear and quadratic terms and no covariates. These estimates are included in table 5. The LMR and BLRT are not reported for one-class models as there is not k – 1 class against which to compare it. Entropy, a measure of the quality of the separation of the classes (Celeux & Soromenho, 1996), is also not estimated for one-class models. The next model estimated was a two-class model. Estimation of the two-class model terminated normally with the optimal loglikelihood repeating multiple times. Although there was a slight decrease in the AIC (ΔAIC = 13.346) and loglikelihood (-2LL = 21.346), the BIC (ΔBIC = 0.133), LMR (p = 0.368), and BLRT (p = 0.364) all indicated that a two-class model was not appropriate. Because the data did not suggest that the growth parameters were drawn from a heterogeneous population, further investigation into a three or more class was not needed.

Despite the lack of evidence to suggest that these data were drawn from a heterogeneous population with regard to growth and change in AHT, it is still possible that the influence of individual differences on those parameters might differ for unobserved subgroups of CSRs. For this reason, a one-class model including cognitive ability, conscientiousness, customer service skill, and sales ability was estimated to serve as a baseline for an investigation into the possibility of multiple classes of predictor to criterion growth relationships in the data. All hypothesized relationships and the significant relationships identified in the LGMs were included in this model. Initial status was regressed on cognitive ability and sales ability. Linear change was regressed on cognitive ability, conscientiousness, and sales ability. Quadratic change was regressed on cognitive ability, conscientiousness, and customer service skill.

Again, a one-class model was estimated to serve as the baseline of comparison between the one and two-class models. Two of the four indicators derived in the estimation of the two-class model indicated that a two-class model was more appropriate for these data (ΔAIC = 27.888, ΔBIC = -11.752, LMR = 51.085, $p = 0.168$, BLRT = 51.888, $p < .05$). Examination of the output revealed that the relationship among the predictors and growth parameters differed between the two classes. Two relationships (out of the eight estimated) were significant in both classes, but in opposite directions. Two relationships were significant in one class but not the other. Also, in real data situations, the AIC is preferred over the BIC in that the assumption in the AIC is that a best fitting approximation is among the set of competing models, whereas the BIC assumes that a true model exists (Buhrman & Anderson, 2004). A "true" model is less likely to exist in the social sciences. Therefore, though the values of the AIC and BIC offered opposing recommendations, the AIC was preferred. Based on these factors, it was determined that a two-class model was more appropriate.

The three-class model was estimated next. The results of the three-class model estimation did not clearly indicate if a two or three-class model was more appropriate for these data (ΔAIC = 13.572, ΔBIC = -26.067, LMR = 36.991, $p = 0.015$, BLRT = 37.572, $p = 0.136$). The output was again examined to see how the relationships among the predictor variables and the growth parameters differed across classes. Three relationships out of eight

Table 5. Three-Class Model Estimates for Average Handle Time with Predictors

Class	Estimates	SE	Variance	SE
Class One (n=10)				
Initial Status Factor	151.066	530.796	9615.438**	1878.849
Initial Status on Cognitive Ability	41.782**	11.997		
Initial Status on Sales Ability	-6.023	6.420		
Linear Slope Factor	-556.260*	262.991	1234.882**	287.981
Linear Factor on Cognitive Ability	-14.510*	6.910		
Linear Factor on Conscientiousness	24.633**	3.322		
Linear Factor on Sales Ability	-0.967	2.413		
Quadratic Slope Factor	222.460	59.190	17.664**	4.467
Quadratic Factor on Cognitive Ability	-0.320	1.089		
Quadratic Factor on Conscientiousness	-4.350**	0.621		
Quadratic Factor on Customer Service Skill	-0.979	0.692		
Class Two (n=36)				
Initial Status Factor	638.419**	184.437	9615.438**	1878.849
Initial Status on Cognitive Ability	-10.911**	2.866		
Initial Status on Sales Ability	3.202	3.301		
Linear Slope Factor	17.442	88.052	1234.882**	287.981
Linear Factor on Cognitive Ability	7.808**	1.505		
Linear Factor on Conscientiousness	-0.068	2.057		
Linear Factor on Sales Ability	-1.922*	0.802		
Quadratic Slope Factor	20.839*	8.833	17.664**	4.467
Quadratic Factor on Cognitive Ability	-0.911**	0.242		
Quadratic Factor on Conscientiousness	-0.443*	0.285		
Quadratic Factor on Customer Service Skill	0.091	0.100		
Class Three (n=155)				
Initial Status Factor	681.807**	120.795	9615.438**	1878.849
Initial Status on Cognitive Ability	-3.184*	1.945		
Initial Status on Sales Ability	-2.525	2.134		
Linear Slope Factor	-169.543**	47.328	1234.882**	287.981
Linear Factor on Cognitive Ability	-1.252	0.810		
Linear Factor on Conscientiousness	3.336**	0.823		
Linear Factor on Sales Ability	0.855	0.567		
Quadratic Slope Factor	24.936**	6.115	17.664**	4.467
Quadratic Factor on Cognitive Ability	0.240*	0.110		
Quadratic Factor on Conscientiousness	-0.418**	0.130		
Quadratic Factor on Customer Service Skill	-0.182**	0.064		

*p < .05; **p < .01

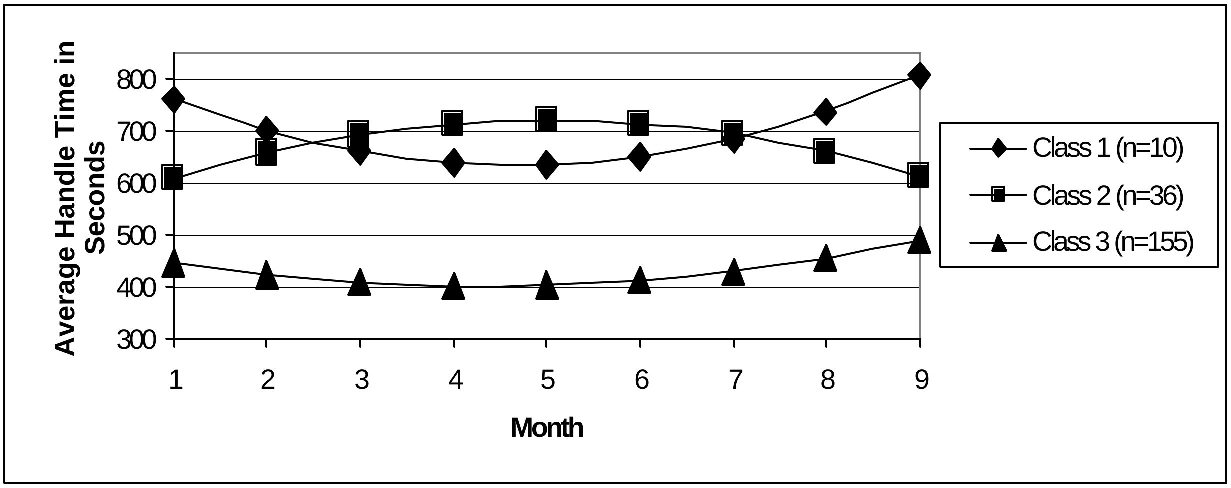
were significant in at least two of the classes but differed in direction in at least two of the classes. Only one relationship was nonsignificant in all three classes, and no relationship was significant and in the same direction in all three classes. Again, though the AIC and BIC offered opposing recommendations, the AIC was preferred. Based on these decision criteria, a three-class model was considered more appropriate than a two-class model.

A four-class model was estimated for comparison to the three-class model. Entropy was low (.718) suggesting that there was poor distinction among the classes. All indicators suggested that a four-class model was not superior to a three-class model (ΔAIC = -6.371, ΔBIC = -46.010, LMR = 17.356, p = 0.953, BLRT = 17.629, p = 1.000). Based on the series of analyses, the three-class model was considered the most appropriate model for these data. The parameter estimates and regression coefficients differed by class, and are presented in Table 5. Using within class means for the predictors, the three-class function was plotted in Figure 1 to demonstrate how each class's growth differed.

The first class only consisted of ten individuals. They demonstrated a high initial AHT, but started to rapidly speed up in their handle time until about month five when they slowed down. CSRs in this class were separate from the rest of the sample in that they demonstrated an unusual pattern of relationships between cognitive ability, conscientiousness, and the criterion. Individuals high in cognitive ability in this class were more likely to have slower AHT ($\beta_{IS.CA}$ = 41.782, $SE_{IS.CA}$ = 11.997, $p < .001$) and individuals high in conscientiousness were more likely to have steeper curves ($\beta_{CH.Cons}$ = 24.633, $SE_{CH.Cons}$ = 3.322, $p < .001$).

Cognitive ability played a large roll in the second class (n = 36). CSRs high in cognitive ability are likely to have lower initial status ($\beta_{IS.CA}$ = -10.911, $SE_{IS.CA}$ = 2.866, $p < .001$) and a steeply increasing initial rate of change ($\beta_{CH.CA}$ = 7.808, $SE_{CH.CA}$ = 1.505, $p < .001$) that quickly turned around and decreased ($\beta_{Q.CA}$ = -0.911, $SE_{Q.CA}$ = 0.242, $p < .001$). All other relationships were consistent with the one-class LGM.

Figure 1. Estimated Growth Curves by Class for Average Handle Time



The largest class (n = 144) had the most ideal growth trajectory. Though these individuals started to slow down in their handle time after month five, they stay within the ideal 400 – 500 second per call range throughout the nine months studied. Also, within this class, hypothesis 2 was supported in that there was a significant relationship between customer service skill ($\beta_{Q.CS}$ = -0.182, $SE_{Q.CS}$ = 0.064, p = .002). For class 3 individuals, the upward turn would be flatter for individuals with higher customer service skill. All of the other hypothesized relationships supported in the one-class LGM held for this class.

To determine if mean levels on the predictor variables differed across the three classes, an ANOVA was conducted for each predictor separately. Post-hoc comparisons were done to determine if specific classes differed from one another. There were no significant differences among the classes on the mean levels of all four predictor variables.

Call Quality

The second variable examined was call quality. The analyses for this variable were very similar to those of AHT with some exceptions. Call quality is typically not measured during a CSRs first month on the job. For this reason, only months two through nine were examined for this variable. In the linear growth model, and all subsequent models, the data were structured such that month one represented initial status even though month one was not included in the analyses.

The fit of the linear model was not good $\chi^2(63, N = 128) = 96.215$, $p < .01$ (TLI = 0.878, RMSEA = 0.064, SRMR = 0.128). The inclusion of the quadratic term resulted in a problematic solution because the variance estimate for the linear slope term was negative. The slope term was determined as not significantly different from zero (Estimate/SE = -.458, p = .323) and was fixed to zero. This model was a significant improvement in fit over the linear growth model $\Delta\chi^2(1, N = 128) = 9.010$, $p < .01$, and the overall fit of the model was good (TLI = 0.906, RMSEA = 0.056, SRMR = 0.133).

Because the linear growth parameter for call quality was fixed to zero, it could not be included in any regressions with the covariates. Sales focus was not expected to relate to call quality and was not included in these analyses. Though no hypotheses were made about cognitive ability and call quality, cognitive ability is typically the best predictor of performance (Hunter, 1986). Therefore it was included. Initial status and quadratic growth were each regressed on cognitive ability, customer service skill, and conscientiousness. The inclusion of the predictors did not significantly improve fit $\Delta\chi^2(14, N = 128) = 20.508$, $p > .05$, but the overall fit of the model was still good (TLI = 0.927, RMSEA = 0.055, SRMR = 0.137). Hypothesis 1 was supported in that customer service skill did not predict initial status in call quality ($\beta_{IS.CS} = -0.038$, $SE_{IS.CS} = 0.255$, p = .440). Customer service did not predict the quadratic term either ($\beta_{Q.CS} = -0.002$, $SE_{Q.CS} = 0.006$, p = .382). Conscientiousness did not significantly predict either intial status ($\beta_{IS.Cons} = 0.148$, $SE_{IS.Cons} = 0.326$, p = .326) or quadratic growth ($\beta_{Q.Cons} = 0.008$, $SE_{IS.Cons} = 0.007$, p = .127). Neither Hypothesis 3 nor 4 was supported. Cognitive ability did not predict initial status ($\beta_{IS.CA}$ = -0.002, $SE_{IS.CA}$ = 0.006, $p = .382$), but did predict quadratic change ($\beta_{Q.CA}$ = -0.015, $SE_{Q.CA}$ = 0.004, $p < .001$).

A one-class model with a quadratic term and the variance of the linear term fixed to zero was estimated as the baseline against which to compare a two-class model with similar constraints. Estimation of the two-class model terminated normally with the optimal loglikelihood repeating multiple times. Though entropy was low (0.681), all but one of the four indicators suggested that the two-class model was more appropriate than a one-class model for these data. (ΔAIC = 11.696, ΔBIC = 0.287, LMR = 18.731, $p = 0.136$, BLRT = 19.696, $p = .013$).

A three-class model was compared to the two-class model and though two of four indicators suggested that the three-class model was more appropriate (ΔAIC = 8.519, ΔBIC = -2.888, LMR = 15.710, $p = 0.223$, BLRT = 16.520, $p = 0.020$), the three-class solution was not permissible due to a negative variance estimate for the quadratic growth estimate in all three classes (all variance components were set to be equivalent across all specified classes). Examination of the variance components indicated that the variance estimates did not significantly differ from zero (variance/SE = -0.212, $p = -0.417$). Because these estimates did not significantly differ from zero, a new model was estimated in which the quadratic variance component was set to zero. As this model is nested, a chi-square test of twice the loglikelihood difference can be computed to determine if the new constraint reduced the overall fit of the model. The chi-square was nonsignificant (-2LL(2, N = 128) = 2.266, $p > .05$) and the AIC an BIC were lower. The LMR and BLRT tests still indicated that a three-class model was better than a two-class model (LMR = 14.766, $p = 0.182$, BLRT = 15.527, $p < .001$).

The four-class model against which the three-class model was compared maintained the quadratic variance constraint imposed on the last model. All four indicators suggested that the four-class model was not better than the three-class model (ΔAIC = 0.206, ΔBIC = -11.203, LMR = 7.804, p = 0.273, BLRT = 8.206, p = .128). Based on the series of analyses, it was determined that a three-class model with variance constraints on the linear and quadratic growth components best fit the data.

Because the variance components for both the linear and quadratic growth components were fixed to zero, they could not be included in any regressions. Initial status was regressed on conscientiousness and cognitive ability as these were the only variables expected to relate to initial status. Though the information criteria increased (ΔAIC = -4.531, ΔBIC = -21.644), there were observable differences in the relationships between the predictors and initial status across the three classes. The model estimates for this three-class model are presented in Table 6.

Table 6. Three-Class Model Estimates for Call Quality with Predictors

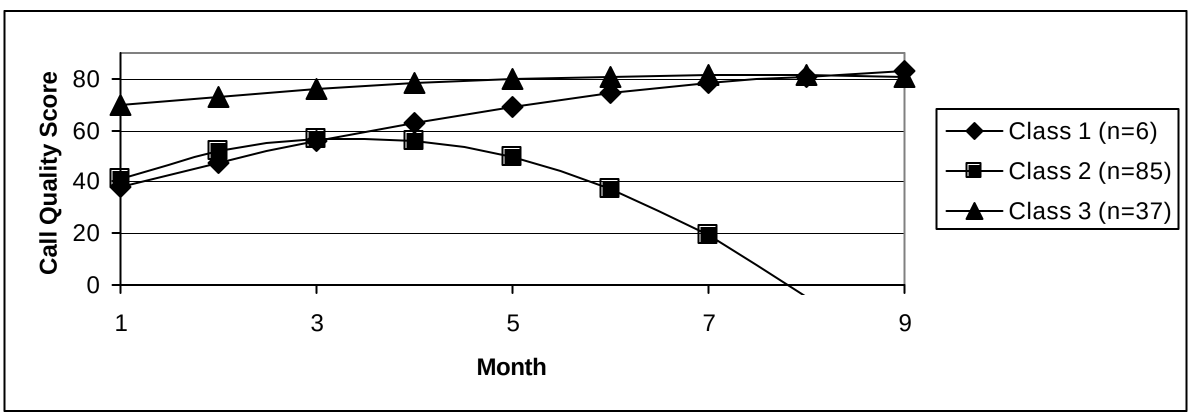
Class	Estimates	SE	Variance	SE
Class One (n=6)				
Initial Status Factor	64.658**	17.241	56.895**	9.093
Initial Status on Cognitive Ability	-0.272	0.436		
Initial Status on Conscientiousness	-0.397	0.462		
Linear Growth Factor	10.091**	1.714		
Quadratic Growth Factor	-0.559**	0.184		
Class Two (n=85)				
Initial Status Factor	210.406**	40.825	56.895**	9.093
Initial Status on Cognitive Ability	-1.868**	0.457		
Initial Status on Conscientiousness	-3.663**	1.039		
Linear Growth Factor	13.35**	2.397		
Quadratic Growth Factor	-2.837**	0.422		
Class Three (n=37)				
Initial Status Factor	53.296**	9.279	56.895**	9.093
Initial Status on Cognitive Ability	-0.055	0.148		
Initial Status on Conscientiousness	0.48*	0.252		
Linear Growth Factor	3.474**	0.780		
Quadratic Growth Factor	-0.264**	0.098		

*p < .05; **p < .01

The growth curve plots for the three classes (Figure 2) identified clearly indicate three separate growth trajectories. Class one starts lower but rapidly catches up to the CSRs in class three, which start high and increase steadily. Class two starts with class one, but hits a peak at month three and rapidly declines. The model estimates projected folks in class two having negative scores months eight and nine, but this is likely due to a high amount of missing data among individuals in class three that turned over because of performance issues.

The regressions for the call quality variable did not reveal much because only initial status could be examined. According to the analyses, in class two, individuals with higher cognitive ability and conscientiousness have lower initial status. Also, according to the analysis of variance, individuals in class three have a significantly higher mean cognitive ability than classes one and two $F(2,125) = 3.589$, $p = .031$. There were no other significant differences in the means of the predictor variables across the three classes.

Figure 2. Estimated Growth Curves by Class for Call Quality



Revenue per Call

The analysis of RPC was carried out in a manner very similar to that of the other two criteria. A linear growth model was estimated for all nine months of data with the growth trajectory fixed. The model fit was very poor $\chi^2(76, N = 173) = 564.227$, $p < .001$ (TLI = 0.666, RMSEA = 0.193, SRMR = 0.823). Addition of a quadratic term resulted in an unacceptable model due to negative residual variance around the month one and month nine mean estimates. The variances for RPC at months one and nine were fixed to 0. This model was compared to the linear growth model. The fit improved significantly $\Delta\chi^2(3, N = 173) = 229.305$, $p < .001$. Though the overall fit was not excellent, it was comparatively much better (TLI = 0.817, RMSEA = 0.143, SRMR = 0.155).

The predictor variables were all added to the model. Initial status, linear growth, and quadratic growth were each regressed on conscientiousness, cognitive ability, customer service skill, and sales ability. The inclusion of the covariates did not statistically improve fit $\Delta\chi^2(12, N = 173) = 16.637$, $p > .05$, and the overall fit of the model was poorer (TLI = 0.798, RMSEA = 0.155, SRMR = 0.192). Hypothesis 1 was supported in that no relationship was found between customer service skill and initial status or linear growth ($\beta_{IS.CS} = 0.050$, $SE_{IS.CA} = 0.050$, $p = .159$; $\beta_{CH.CS} = 0.038$, $SE_{CH.CS} = 0.037$, $p = .160$). No relationship was found between customer service skill and quadratic growth ($\beta_{Q.CS} = -0.002$, $SE_{Q.CS} = 0.005$, $p = .375$). No support was found for Hypotheses 3 and partial support was found for Hypothesis 4. Conscientiousness did not predict initial status, linear growth, or quadratic growth ($\beta_{IS.Cons} = -0.082$, $SE_{IS.Cons} = 0.062$, $p = .095$; $\beta_{CH.Cons} = -0.063$, $SE_{CH.Cons} = 0.046$, $p = .084$; $\beta_{Q.Cons} = 0.004$, $SE_{Q.Cons} = 0.005$, $p = .268$). Hypotheses 7 and 8 were not supported. No relationship was found between sales ability and initial status, linear growth, or quadratic growth ($\beta_{IS.SA} = 0.017$, $SE_{IS.SA} = 0.048$, $p = .359$; $\beta_{CH.SA} = 0.015$, $SE_{CH.SA} = 0.036$, $p = .341$; $\beta_{Q.SA} = 0.000$, $SE_{Q.SA} = 0.006$, $p = .492$). Though no hypotheses guided its inclusion, cognitive ability was related to initial status and linear growth ($\beta_{IS.CA}$ = 0.082, $SE_{IS.CA}$ = 0.037, p = .012; $\beta_{CH.CA}$ = -0.045, $SE_{CH.CA}$ = 0.026, p = .044).

A one-class model with the variances for months one and nine fixed to zero was estimated as a baseline against which to compare a two-class model. The two-class model had high entropy and most indicators suggested that a two-class model fit the data (ΔAIC = 201.174, ΔBIC = 188.561, LMR = 199.496, p = 0.266, BLRT = 209.174, p < .001). The two-class model was compared to a three-class model. The three-class model also had high entropy and most indicators suggested that a three-class model was more appropriate than a two-class model (ΔAIC = 192.802, ΔBIC = 180.189, LMR = 191.511, p = 0.097, BLRT = 200.802, p < .001). To further explore the number of classes found in the data, the three-class model was compared to a four-class model. Most of the indicators suggested that a four-class model was more appropriate than a three-class model (ΔAIC = 79.563, ΔBIC = 66.950, LMR = 83.512, p = 0.273, BLRT = 87.563, p < .001). The four-class model was compared to a five-class model. Some of thhe indicators suggested that a five-class model was more appropriate (ΔAIC = 35.626, ΔBIC = 23.013) whereas others did not LMR = 199.496, p = 0.266). Also, the new class consisted of only one individual. Based on the smaller changes in the AIC and BIC, the nonsignificant LMR test, and the new class with one individual, it was determined that a four-class model was most appropriate for these data.

The predictor variables were added to the four-class model. Initial status was regressed on cognitive ability, conscientiousness, and sales ability. Linear growth was regressed on cognitive ability and sales ability. Quadratic growth was regressed on customer service skill and conscientiousness. The four-class model estimates are provided in Table 7. Examination of the output revealed large differences in the relationships with the criteria across classes. Certain relationships that were predicted that did not hold when the sample was analyzed as one class held for some of the classes identified with GMM. In class one, sales ability had a positive effect on linear growth (0.518, SE = 0.203, $p = .005$). In class two, sales ability had a positive effect on linear growth (0.093, SE = 0.028, $p < .001$) and customer service skill had a positive effect on quadratic growth (0.005, SE = 0.003, $p = .023$). In class three, cognitive ability had a positive effect on linear growth (0.024, SE = 0.012, $p = .021$) and customer service skill had a positive effect on quadratic growth (0.006, SE = 0.02, $p < .001$). There were no significant hypothesized relationships in class four. Using within-class predictor means, the four-classes are plotted in Figure 3.

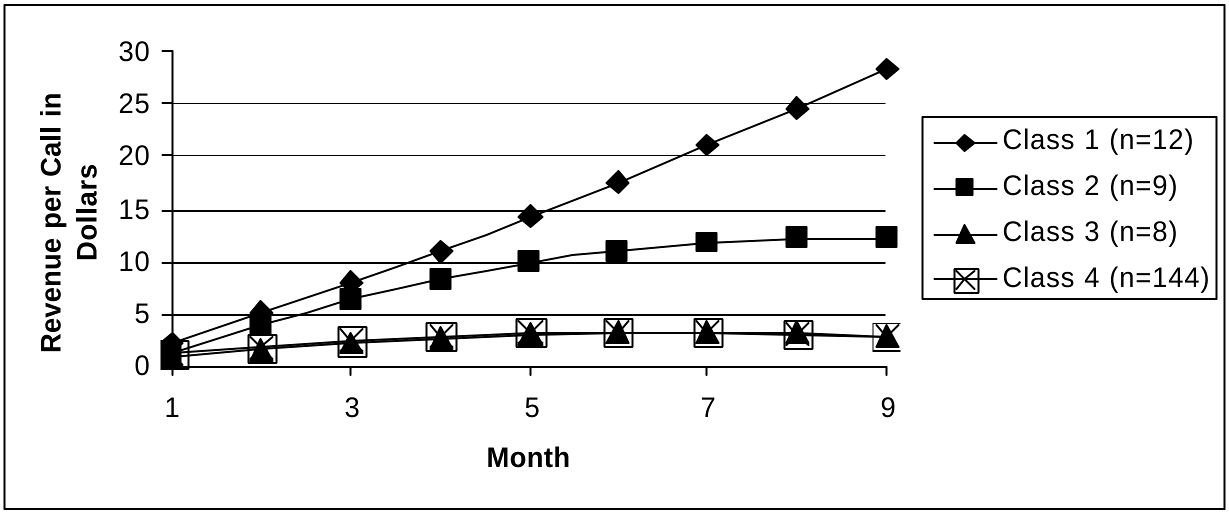
Analysis of variance was conducted on the mean levels of the predictors across the four classes. No significant differences were found.

Table 7. Four-Class Model Estimates for Revenue per Call with Predictors

Class	Estimates	SE	Variance	SE
Class One (n=12)				
Initial Status Factor	-33.885	42.318	1.901	2.334
Initial Status on Conscientiousness	-0.528	0.694		
Initial Status on Cognitive Ability	0.426	0.520		
Initial Status on Sales Ability	0.857	1.053		
Linear Growth Factor	-26.434*	12.764	2.580**	0.881
Linear Growth on Cognitive Ability	0.054	0.090		
Linear Growth on Sales Ability	.518**	0.203		
Quadratic Growth Factor	1.635*	0.790	.066**	0.022
Quadratic Growth on Customer Service Skill	-.012*	0.005		
Quadratic Growth on Conscientiousness	-0.025	0.019		
Class Two (n=9)				
Initial Status Factor	88.663**	5.127	1.901	2.334
Initial Status on Conscientiousness	-.820**	0.174		
Initial Status on Cognitive Ability	0.273	0.247		
Initial Status on Sales Ability	-1.211**	0.123		
Linear Growth Factor	0.000	0.000	2.580**	0.881
Linear Growth on Cognitive Ability	-.123**	0.047		
Linear Growth on Sales Ability	.093**	0.028		
Quadratic Growth Factor	0.000	0.000	.066**	0.022
Quadratic Growth on Customer Service Skill	.005*	0.003		
Quadratic Growth on Conscientiousness	-.014*	0.007		
Class Three (n=8)				
Initial Status Factor	3.751	13.356	1.901	2.334
Initial Status on Conscientiousness	0.031	0.094		
Initial Status on Cognitive Ability	-0.006	0.134		
Initial Status on Sales Ability	-0.072	0.147		
Linear Growth Factor	0.000	0.000	2.580**	0.881
Linear Growth on Cognitive Ability	.024*	0.012		
Linear Growth on Sales Ability	0.005	0.009		
Quadratic Growth Factor	-.531**	0.205	.066**	0.022
Quadratic Growth on Customer Service Skill	.006**	0.002		
Quadratic Growth on Conscientiousness	0.003	0.005		
Class Four (n=144)				
Initial Status Factor	-0.682	2.325	1.901	2.334
Initial Status on Conscientiousness	-0.033	0.035		
Initial Status on Cognitive Ability	0.055	0.037		
Initial Status on Sales Ability	0.040	0.031		
Linear Growth Factor	1.119*	0.477	2.580**	0.881
Linear Growth on Cognitive Ability	-.015*	0.007		
Linear Growth on Sales Ability	-0.001	0.008		
Quadratic Growth Factor	-0.083	0.078	.066**	0.022
Quadratic Growth on Customer Service Skill	0.000	0.001		
Quadratic Growth on Conscientiousness	0.000	0.001		

*p < .05; **p < .01

Figure 3. Estimated Growth Curves by Class for Revenue per Call



CHAPTER 4

DISCUSSION

The purpose of this study was to introduce growth mixture modeling to the job performance arena. Before this could be done, change in performance was evaluated using latent growth modeling. This was done so that the hypotheses about the relationships between the predictors and the change functions of the criteria could be tested using a more well-known method to serve as a basis of comparison when the hypotheses were tested using the more advanced GMM techniques. Support was found for about half of the hypotheses using LGM. For the relationships that were not supported, conditional GMMs indicated that there is population heterogeneity with regard to how the measured individual differences predict future job performance initial status and change.

Average handle time was the first variable examined. Customer service skill did not predict initial status or linear growth as proposed in hypothesis 1. It also did not predict quadratic change as was expected when the relationship was evaluated for the sample population as a whole in a one-class model. The three-class conditional GMM, however, indicated that this relationship holds for a large portion of the sample (77%). It was predicted that customer service skill would relate to the extent to which performance would start to decline for CSRs that handled a large amount of emotional dissonance. Numerous individual differences could be affecting the two classes of individuals' ability to cope with the stress of working in a call center environment. It is also possible that these classes of individuals happened to face more challenging customers on the job.

Another relationship that was hypothesized that only held for one class was the relationship between cognitive ability and the linear slope of AHT. Individuals higher in cognitive ability ought to be able to learn to multitask better and therefore should handle calls faster, but that appears to be the case in class one. No explicit measure of previous job experience was included in this study. According to Voelkle, Wittmann, and Ackerman (2006) the role of cognitive ability in skill acquisition decreases as the task becomes automated. It is likely that the ten individuals in class one had lower previous job experience in a call center role and therefore needed to learn the job. For this reason, cognitive ability would play a larger role. Also, in this class, AHT decreased steadily until about month five when call times began increasing again. In this class, though the difference was not statistically significant, the mean cognitive ability was higher than in the other two classes. The role of call center agent is not very complex from a cognitive standpoint. CSRs higher in cognitive ability likely became bored with the task and lost interest.

In class two, individuals higher in cognitive ability actually have a significant positive linear growth factor. The linear growth factor is difficult to interpret when a quadratic term is present. The quadratic term determines the turn of the curve. A positive quadratic term will be a U shape. A negative quadratic term will be an upside-down U shape. In the presence of a quadratic term, the linear term will determine where the peak (or nadir) of a quadratic function will appear. If the quadratic term is negative, a higher linear value will move the peak to the right. If the quadratic term is positive, a higher linear value will move the nadir to the left. For this reason, terms must be viewed in conjunction to understand what they mean. In class two, as cognitive ability increases, the curve will become increasingly negatively quadratic and the peak will move to the right. This means that for very high cognitive ability values, one would see a rapidly increasing AHT that does not peak within the parameters of this study. Again, this seems counterintuitive, but in a lower-level job, an individual with higher cognitive ability is more likely to lose interest in the task and turnover (Townsend, 2007).

Conscientiousness has the same effect in all three classes. As conscientiousness increases, the quadratic functions become more negative and the peaks move further to the right. Therefore, as conscientiousness increases, the curve of AHT becomes increasingly steeper with its peak further to the edge of the study parameters. Though it was predicted that individuals high in conscientiousness would adhere to their scripts better and would handle calls faster for that reason, the opposite appears to be true. This is may be due to individuals high in conscientiousness lacking multitasking ability. In a study of polychronicity, individuals high in conscientiousness tended to be very single task focused based on self report (Kantrowitz et al, under review). Therefore, highly conscientious CSRs may focus on one customer call related activity at a time.

If one were to show the three growth trajectories estimated in Figure 1 to a human resources professional at any organization with a call center, they would say that class three is the desirable class. The objective of this study was to find ways to not only identify desirable and undesirable classes, but to predict membership in them as well. There were no significant differences in the mean levels of any of the predictors across the three classes. Therefore, class membership could not be predicted using the predictors included in this study, but this is not to say that other predictors might not possess the ability to distinguish among candidates. One drawback to these analyses is sample size. It is likely that if the sample were larger, the difference between class one and the other two classes in cognitive ability would be significant. The only problem with this is that class one is probably the least desirable trajectory. Should candidates that are too high in cognitive ability be turned down for a job in a call center? Overall prediction aside, it is important to identify the individual differences that predict membership in the three classes due to the large differences in the relationships among the predictors and growth parameters across classes. With the exception of conscientiousness, the predictors in class three are related to the criteria as they would be expected. Prediction would be improved by understanding what it is about the individuals in classes one and two that obfuscates these relationships.

Call quality was examined next. The analyses of call quality suffered from extensive missing data. The measurement of call quality, though wrought with rigor, is time consuming for the supervisors that need to do the evaluation and is the metric most likely to be ignored (Levin, 2007). Despite this, some interesting relationships were found. Based purely on a visual analysis of the estimated trajectories in Figure 2, class three is clearly the most desirable trajectory. Class three has a statistically significantly lower cognitive ability mean than the other two classes. The individuals in class one eventually catch up to the CSRs in class three, but the CSRs in class two start to improve before a drastic downward curve. None of the other predictors differentiated among the three classes. Therefore, important unmeasured individual differences must be differentiating these classes.

Job experience is likely the factor influencing the separation between class one and three. Despite having lower cognitive ability overall, the CSRs in class three have a fairly high and stable pattern across the nine months of the study. The individuals in class one catch up at about month 7 indicating that higher levels of performance can be achieved. No variance was left in the linear or quadratic functions after accounting for class membership, so the increase in performance in class one could only be cognitive ability and unmeasured causes.

Call quality is an elective metric (Levin, 2007). Though there is likely some protocol, it is up to the supervisor who will be evaluated and when quality assessments will be done. Individuals that have had performance problems are more likely to be evaluated than any other group. This is likely the reason the poorest performing class is the largest one. A personality prone to counterproductive work behaviors may be the unmeasured difference between class one and two.

Class differentiation for revenue per call appears to mainly be a function of some drastically different linear and quadratic growth terms more so than large differences in the relationships among the predictors and the growth terms. The primary limitation of the analysis of revenue per call is that selling is encouraged but not required of CSRs. Therefore, it should not be surprising that the class with the lowest performance across time, is the largest class. Sales ability predicts a steeper increase in revenue per call over time in the two classes with steep increases in revenue per call. There were no significant differences found in the predictors across the four classes, but power is severely limited by the fact that classes one, two, and three have twelve, nine, and eight CSRs each. There is not enough information to know if the job demands of the individuals whose revenue per call increased steadily over the nine months of the study differed from the individuals in class three and four.

Limitations and Future Directions

The largest limitation in this study is that the data were archival. Though it is easier to generalize data that come from a real world setting and the data were collected outside of researcher bias, using data collected by individuals more concerned with the bottom line than science posed many challenges. Though call center data are typically captured with high rigor, data points were missing for many individuals within the window of data provided. Also, although data are always being collected for performance management reasons, only eleven months of performance data were provided. Therefore, only individuals that started the first, second, or third months of the study could be included because of the importance of capturing the early data points of the learning curve. Though power for some analyses may have suffered, the sample was sufficient to uncover distinct classes of growth and change and observe some interesting relationships. Unfortunately, it limited the ability to study tenure, the criterion that organizations probably care the most about.

Another major limitation to the study is the nature of the job studied. The metrics captured in the call center are ideal for a longitudinal study, but they may lack meaning in any other job domain. Future studies should use subjective ratings of performance, preferably from multiple raters for estimates of reliability and inter-rater agreement. Criterion unreliability can be corrected for in structural equation modeling. Supervisor ratings of job performance have meaning across all jobs.

Future evaluations of the utility of growth mixture modeling for the understanding of job performance would require increased sample size, one or more different jobs, and subjective performance ratings made by trained raters. Jobs can be categorized by type and level, and these variables can be integrated into a GMM in which variance attributable to job type and job level can be controlled for. This will allow the results to generalize to a much greater extent.

Conclusions

The purpose of this study was to demonstrate that growth mixture modeling ought to have the ability to differentiate ideal from undesirable growth trajectories in job performance and that the predictors frequently used by industrial/organizational psychologists could be used to predict likelihood of membership in these classes. The sum of the results is much more complicated than that. What is clear is that there are multiple classes of performance change over time within a nine month span of performance of call center customer service representatives. It is also clear that some of these classes are ideal and others are not. What is not clear is why the relationships among the predictors and the growth parameters differ across classes. Some classes support expected relationships between predictors and criteria, whereas others do not. For the most part, these classes cannot be predicted using the predictors that were available.

Another purpose of this study was to use GMM to better understand the intricacies of the dynamic criteria debate. Barrett and collegues (1985) suggested that in order for the validities to go down, there must be a change in the rank order of individuals. Declining validities were observed for AHT and call quality, and there is clearly a crossing of paths among the classes identified. For sales ability and revenue per call, however, the validities actually increased. According to Figure 3, individuals fan out rather than cross multiple times. What is very clear from these analyses is that people are changing their performance over time and that different predictors are important for the prediction of different metrics at different times during an employee’s tenure.

Billions of dollars are spent on selecting the best employees, but despite over a century of I/O work, there are still weaknesses in our methodology. GMM has the potential to examine job performance at a global level that cannot be achieved with other methods. Rather than just looking at means, snapshot time points, or even latent growth parameters, GMM allows us to examine change over time in its entirety. This allows a hiring manager to select the growth trajectories that best meet the needs of his or her business. In a job with high turnover, the hiring manager many not want to deal with people that start with low job performance no matter how quickly they improve. Another hiring manager may want to eliminate any individuals that show signs of plateauing or declining performance. The next task is identifying the individual differences that can predict the likelihood of membership in the desired classes. If we can show that some predictors in the I/O arsenal have the ability to reliably distinguish individuals in these classes, GMM can revolutionize the staffing industry.

REFERENCES

Akaike, H. 1973. Information Theory and an Extension of the Maximum Likelihood Principle. pp. 267-281 in *Second International Symposium on Information Theory*, edited by B.N. Petrov and F. Csaki. Budapest: Akademiai Kaido. (Reprinted, with an introduction by J. deLeeuw, in *Breathroughs in Statistics*, Volume I, edited by Samuel Kotz and Norman L. Johnson. New York: Springer, 1992, pp. 599-624.)

Alvares, K.M. & Hulin, C.L. (1973). An experimental evaluation of a temporal decay in the prediction of performance. *Organizational Behavior and Human Performance, 9,* 169-185.

Ashforth, B.E. & Humphrey, R.H. (1993). Emotional labour in service roles: The influence of identity. *Academy of Management Review, 18,* 88-115.

Austin, J.R. & Crespin, T.R. (2006). From "Criterion Problem" to Problems of Criteria in Industrial and Organizational Psychology: Progress, Pitfalls, and Prospects. In, Bennett, W. Jr., Lance, C.E., & Woehr, D.J. (Eds) *Performance measurement: Current perspectives and future challenges* (pp 9-48). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Austin, J.R., Humphreys, L.G. & Hulin, C.L. (1989). Another view of dynamic criteria: A critical reanalysis of Barrett, Caldwell, and Alexander, *Personnel Psychology, 42,* 583-596.

Austin, J.T. & Vilanova, P (1992). The criterion problem: 1917 – 1992. *Journal of Applied Psychology, 77,* 836-874.

Barrett, G.V., Caldwell, M.S. & Alexander, R.A. (1985). The concept of dynamic criteria: A critical reanalysis. *Personnel Psychology, 38,* 41-56.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-26.

Bauer, D. J. & Curran, P.J. (2003). Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological Methods, 8,* 338-363.

Bellows, R.M. (1941). Procedures for evaluating vocational criteria. *Journal of Applied Psychology, 25,* 499-513.

Binning, J. & Barrett. G.V. (1989). Validity of personnel decisions: A review of the inferential and evidential bases. *Journal of Applied Psychology, 74,* 478-494.

Bolanovich, D.J. (1946). Statistical analysis of an industrial rating chart. *Journal of Applied Psychology, 30,* 23-31.

Bliese, P.D. & Ployhart, R.E. (2002). Growth modeling using random coefficient models: model building, testing and illustrations. *Organizational Research Methods, 5,* 362-387.

Bryk, A.S. & Raudenbush, S.W. (1987). Application of hierarchical linear models to assessing change, *Psychological Bulletin, 101,* 147-158.

Burnham, K. P., & Anderson, D.R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research, 33*, 261-304.

Butler, R.P., & McCauley, C. (1987). Extraordinary stability and ordinary predictability of academic success at the United States Military Academy. *Journal of Educational Psychology, 79,* 87.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13,* 195-212.

Conte, J.M., & Jacobs, R.R. (2003). Validity evidence linking polychronicity, and big five personality dimensions to absence, lateness, and supervisory performance ratings. *Human Performance, 16,* 1078-129.

Deadrick, D.L., Bennett, N. & Russell, C.J. (1997). Using hierarchical linear modeling to examine dynamic criteria over time. *Journal of Management, 23,* 745-757.

Deadrick, D.L., & Madigan, R.J. (1990). Dynamic criteria revisited: A longitudinal study of performance stability. *Personnel Psychology, 44,* 717-744.

Dunham, R.B. (1974). Ability-skill relationships: An empirical explanation of change over time. *Organizational Behavior and Human Performance, 12,* 373-382.

Fleishman, E.A. & Hempel, W.E. (1954). Changes in factor structure of a complex psycho-motor test as a function of practice. *Psychometrika, 18,* 239-252.

Fleishman, E.A. & Mumford, M. (1989). Abilities as causes of individual differences in skill acquisition. *Human Performance, 2,* 201-223.

Ghiselli, E.E. (1956a). Dimensional problems of criteria. *Journal of Applied Psychology, 40,* 1-4.

Ghiselli, E.E. (1956b). Differentiation of individuals in terms of their predicatability. *Journal of Applied Psychology, 40,* 374-377.

Ghiselli, E.E. (1966). *The Validity of Occupational Aptitude Tests,* New York: Wiley.

Ghiselli, E.E. & Haire, M. (1960). The validation of selection tests in light of the dynamic nature of criteria. *Personnel Psychology, 13,* 225-231.

Goldberg, L.S. & Grandey, A.A. (2007). Display rules versus display autonomy: emotional regulation, emotional exhaustion, and task performance in a call center simulation. *Journal of Occupational Health Psychology, 12,* 301-318.

Grandey, A.A. (2003). When "the show must go on": Surface and deep acting as predictors of emotional exhaustion and service delivery. *Academy of Management Journal, 46,* 86-96.

Ho, H-Z., O'Farrell, S.L., Hong, S., & You, S. (2006) Developmental research: Theory, method, design, and statistical analysis. In J.L. Green, G. Camilli, & D. Elmore (Eds.), *Complementary methods for research in education* (3rd ed. pp. 207-225). Washington, DC: American Education Research Association.

Hochschild, A.R. (1983). *The managed heart.* Berkley, CA: University of California Press.

Hofmann, D.A., Jacobs, R., & Baratta, J.E. (1993). Dynamic criteria and the measurement of change. *Journal of Applied Psychology, 78,* 194-204.

Hofmann, D.A., Jacobs, R., & Gerras, S.J. (1992). Mapping individual performance over time. *Journal of Applied Psychology, 77,* 185-195.

Hunter, J.E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340-362.

Hunter, J.E., & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72-98.

Inn, A., Hulin, C.L., & Tucker, L.R. (1972). Three sources of criterion variance: Static dimensionality, dynamic dimensionality, and individual dimensionality. *Organizational Behavior and Human Performance, 8,* 58-83.

Jain, A.K. & Dubes, R.C. (1988). *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice-Hall.

Kanfer, R. & Ackerman, P.L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74,* 657-690.

Kantrowitz, T., Beaty, J., Grelle, D., & Wolf, M (2008). Investigating preferences for multitasking: connecting polychronicity to job performance across job levels and criteria. Manuscript submitted for publication.

Konig, C.J., Buhner, M. & Murling, G. (2005). Working memory, fluid intelligence, and attention are predictors of multitasking performance, but polychronicity and extraversion are not. *Human Performance, 18,* 243-266.

Lee, R.T. & Ashforth, B.E. (1996). A Meta-analytic examination of the correlates of the three dimensions of job burnout. *Journal of Applied Psychology, 81,* 123-133.

Levin, G. (2007). Measuring the things that matter; A deep dive into seven key metrics that are most critical in gauging and securing customer satisfaction, loyalty, and contact center effectiveness. *Call Center Magazine, 2007,* 24-32.

Lewig, K.A. & Dollard, M.F. (2003). Emotional dissonance, emotional exhaustion and job satisfaction in call centre workers. *European Journal of Work and Organizational Psychology, 12,* 366-392.

Little, R.J. & Rubin, D.B. (2002). *Statistical Analysis With Missing Data.* (2nd ed.). Hoboken, NJ: John Wiley.

Lo, Y., mendell, N.R., & Rubin, D.B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88,* 767-778.

Matthews, G., Emo, A., Funke, G. Zeidner, M., Roberts, R.D., Costa Jr., P.T. & Schulze, R. (2006). Emotional intelligence, personality, and task-induced stress. *Journal of Experimental Psychology: Applied, 2006,* 96-107.

Meredith, W. & Tisak, J. (1990). Latent curve analysis, *Psychometrika, 55,* 107-122.

Mulholland, K. (2002). Gender, emotional labour and teamworking in a call centre. *Personnel Review, 31,* 283-303.

Murphy, K.R. (1989). Is the relationship between cognitive ability and job performance stable over time? *Human Performance, 2,* 183-200.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557-585.

Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: new opportunities for latent class/latent growth modeling. In A. Sayer & L. Collins (Eds.), *New methods for the analysis of change* (pp. 291-322). Washington, DC: American Psychological Association.

Muthén, B. (2003). Statistical and substantive checking in growth mixture modeling: comment on Bauer and Curran (2003). *Psychological Methods, 8*, 369-377.

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345-368). Newbury Park, CA: Sage Publications.

Muthén, B. & Curran, P. (1997). General longitudinal modeling of individual difference in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2,* 3731-402.

Muthén, L. & Muthén, B. (2007). Mplus (Version 5.0) [Computer software].

Nylund, K.L., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling, 14,* 535-569.

Owen, W.A. (1976). Background data. In M.D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 609-644). Chicago: Rand-McNally.

Perl, R.E. (1934). An application of Thurstone's method of factor analysis to practice series. *Journal of General Psychology, 11,* 209-212.

Ployhart, R.E. & Hakel, M.D. (1998). The substantive nature of performance variability: Predicting interindividual differences in intraindividual performance. *Personnel Psychology, 51,* 859-901.

Rambo, W.W., Chomiak, A., & Price, J.M. (1983). Consistency of performance under stable conditions of work. *Journal of Applied Psychology, 68,* 78-87.

Ronan, W.W., & Prien, E. (1966). *Toward a criterion theory: A review and analysis of research and opinion.* Greensboro, NC: Smith Richardson Foundation.

Rothe, H.F. & Nye, C.T. (1958). Output rates among coil winders. *Journal of Applied Psychology, 42,* 182-186.

Schmidt, F.L. & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics, 6,* 461-464.

Singer, J.D. & Willett, J.B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.

Smith, P.C. (1976). Behavior, results and organizational effectiveness: The problem of criteria. In M.D. Dunnette (Ed.). *Handbook of industrial and organizational psychology* (pp. 745-775). Chicago: Rand McNally.

Stokes, G.S., Toth, C.S., Searcy, C.A., Stroupe, J.P., & Carter, G.W. (1999). construct/rational biodata dimensions to predict salesperson performance: Report on the U.S. Department of Labor sales study. *Human Resources Management Review, 9,* 185-218.

Sturman, M.C. & Trevor, C.O. (2001). The implications of linking the dynamic performance and turnover literatures. *Journal of Applied Psychology, 86,* 684-696.

Thoresen, C.J., Bradley, J.C., Bliese, P.D., & Thoresen, J.D. (2004). The big five personality traits and individual job performance growth trajectories in maintenance and transitional job stages. *Journal of Applied Psychology, 89,* 835-853.

Thorndike, R.L. (1949). *Personnel selection.* NY: Wiley.

Toops, H.A. (1944). The criterion. *Educational and Psychological Measurement, 4,* 271-297.

Totterdell, P. & Holmann, D. (2003). Emotional regulation in customer service roles: Testing a model of emotional labour. *Journal of Occupational Health Psychology, 8,* 55-73.

Townsend, K. (2007). Recruitment, training and turnover: another call centre paradox. *Personnel Review, 3,* 476-490.

U.S. Department of Labor (2008). *O*NET 10.0 Database.* Retrieved April 14, 2008, from http://online.onetcenter.org/

Voelkle, M.C., Wittmann, W.W., Ackerman, P.L. (2006). Abilities and skill acquisition: A latent growth curve approach. *Learning and Individual Differences, 16,* 303-319.

Willet, J.B. & Sayer, A.G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116,* 363-381.

Zapf, D., Isic, A., Bechtold, M., & Blau, P. (2003). What is typical for call centre jobs? Job characteristics, and service interactions in different call centres. *European Journal of Work and Organizational Psychology, 12,* 311-340.