THE IMPACT OF TENURE, DEMOGRAPHICS AND LEADERSHIP ON PERFORMANCE AND QUALITY IN A FIELD SETTING

A Thesis

Presented to the

Faculty of the College of Graduate Studies

Of Angelo State University

In Partial Fulfilment of the

Requirements for the Degree

MASTER OF SCIENCE

by

JOHN RICHARD STARNE

May 2015

Major: I/O Psychology

THE IMPACT OF TENURE, DEMOGRAPHICS AND LEADERSHIP

THE IMPACT OF TENURE, DEMOGRAPHICS AND LEADERSHIP

ON PERFORMANCE AND QUALITY IN A FIELD SETTING

by

JOHN RICHARD STARNE

APPROVED:

Dr. Kraig L. Schell

Dr. Cheryl K. Stenmark

Dr. Ismael Diaz

Dr. Jeffrey D. Dailey

May 2015

APPROVED:

Dr. Susan E. Keith Date
Dean of the College of Graduate Studies

Abstract

This study examined the applicability of skill development and performance evaluation theories in a real-world context. Studies of expertise generally show that skills develop according to a power law function, while the performance appraisal literature suggests that rater behavior can be constrained by the quality of the available data and social/organizational forces. A database donated by a public opinion polling company as part of an ongoing consultation, which contains both objective employee performance and subjective ratings of that performance, was used to examine the validity of the literature in these two areas. In general, the data revealed that performance followed a power law curve asymptoting after about 12-18 working days, as predicted by current theory. Subjective ratings showed less variation with practice and differences in score patterns depending on the concreteness of the dimension evaluated, in line with current theory. Recommendations for the client organization are also included.

Table of Contents

List of Tables

List of Figures

Introduction

Much existing psychology was initially generated in laboratory settings where high internal validity is maintained in an artificial environment. While organizational theory is often tested within organizations the results are not subject to general publication due to proprietary and competitive issues typical of business. This thesis provides an opportunity to apply task performance and performance appraisal theory to data provided by an organization that is willing to allow results to contribute to the general psychological literature, a rare and welcome event.

Issues in Task Performance in High-Turnover Workforces

Employee turnover has been studied for years for a variety of reasons, including the cost of turnover and the disruption to the process of work (Hillmer, Hillmer, & Roberts, 2004). For these reasons, most organizations are inclined to retain employees over a period of several months to years; however some industries, especially those that cater to entry-level employees, operate with workforces that have significantly shorter typical tenures, measured more accurately in weeks. Much available research focusing on job performance and production is restricted to performance over long periods of time and may not apply to organizations that have shorter typical employee tenures (Dalal, Bhave, & Fiset, 2014). Organizations that must survive despite a constantly-shifting workforce are vulnerable to a number of obstacles that make performance prediction more difficult and potentially misleading.

APA Style

First, research focusing on within-person productivity suggests that effective prediction over an extended period of time proves inaccurate due to numerous factors, such as self-efficacy, affect, self-control, shifts in trained abilities and measurement errors (Dalal, Bhave, & Fiset, 2014). But in organizations with high turnover, it is often difficult to look at performance over time at a person level. Thus, there is a lack of understanding about how performance evolves in a high-turnover organization with repetitive, structured tasks. Similarly, though historically researchers have assumed that performance data in organizations is represented by normal distributions, recent trends have shown that employee performance is best approximated by a Pareto distribution (O'Boyle & Aguinis, 2012). In a Pareto distribution the majority of work is completed by a small group of employees (the typically espoused ratio is 80/20). An objective determination of individual contributions to productivity is frequently challenging due to the non-normality of this distribution, but any examination of job performance in high-turnover environments must accept the reality of these data and attempt to understand them anyway.

Laboratory-based human performance literature largely focuses on basic tasks with very limited variations. Applied research into skilled performance is not as common, but laboratory simulations attempt to emulate these environments regarding skills such as driving and air traffic control (Ericsson & Williams, 2007). It is rare to have the opportunity to study real data on entry level non-manufacturing positions, so this study will provide insight into an area not heavily explored. This thesis describes such an opportunity. In cooperation with a host organization and in the context of a consultation event, we will examine how psychological theory interfaces with a rich and real-world database that contains information on actual employee performance on a complex task.

The Current Client: Overview and Description

Much of the information that follows in this section was derived from two primary sources: archival documents supplied by the organization and personal communication with the primary client. Any additional resources will be cited as they emerge.

The primary business of the host organization (hereafter referenced as HostCo) is public opinion research, predominantly concerning political issues and largely conducted over the telephone. HostCo utilizes computer telephony interface (CATI) software to conduct live telephone interviews throughout the United States. HostCo was established in the late 1990's and is currently one of the most prominent companies within its industry. HostCo is also a distinguished member of the American Association for Public Opinion Research (AAPOR), a professional organization that provides guidelines on best practices, ethics and codes of conduct for the industry. As a commercial enterprise, HostCo is not subject to external oversight or regulation and can (and often must) make operational allowances based on individual client preferences.

The nature of HostCo's business is seasonal, with workloads increasing substantially (by a factor of two or more) in even-numbered years, because political elections are the primary target issue for HostCo. Thus, the highest call volumes and staff demands occur between June and November of even-numbered years. Due to this cyclical nature, reductions in workforce (typically through voluntary turnover) occur in early November every other year. HostCo is not unique in this way; high turnover is characteristic of all members of this industry. HostCo counteracts these turnover effects by engaging in active hiring throughout the year. To illustrate the volatility of the workforce, over 50% of employees in the interviewer position terminate before thirty calendar days after hire (HostCo, personal

communication). Therefore, it was important to decide quickly in the consultation process how to extract the appropriate data based on these realities. In agreement with HostCo, the database included in this review begins on January 1, 2013 and ends on October 31, 2014. When appropriate in this introduction and beyond, calendar years will be referenced as YR13 (January 1, 2013 to December 31, 2013) and YR14 (January 1, 2014 to October 31, 2014).

HostCo's Physical Locations: Global Characteristics

HostCo operated a maximum of three call center locations within the United States during the timeframe targeted in this report. During YR13, two centers were in operation: Locations S1 and S2, both geographically located in the southern United States. During YR14, a third call center, Location NW, was in operation for a period of time and was located geographically in the Pacific Northwest. Corporate offices are located in the southern United States and are in a separate physical location from any of the call centers. Despite this separation, archival documents indicate that call centers have little discretion concerning day-to-day business operations. Company policies are centralized and rigid, and the culture surrounding them leaves little room for movement. These policies include all aspects of operation, including hiring, training, and performance expectations and evaluations, despite the fact that the areas surrounding the centers vary considerably in general economic conditions, job outlook and demographics. This centralized control structure is further supported by the dialing software, which links the geographical locations tightly into a sort of "virtual conglomerate", with ultimate oversight of the processes tasked to a project director located physically at Location S1.

Since this is an investigation of employee performance, it is useful to provide some context regarding the global characteristics of the environment from which these employees

were drawn. Both of the southern locations (S1 and S2) were in operation for the entire time period outlined previously and as such will comprise a considerable portion of the data. The two sites are in close physical proximity (separated by about 40 miles) and are both located in the same state. Despite this fact, the labor pools surrounding each site are quite different. Location S1 is in a small town outside of a large metropolitan area. The primary labor pool to which it has access is a mid-sized regional state university, which means it must usually settle for part-time workers. Location S2 is adjacent to a suburb of that same metropolitan area, and while it does employ students from surrounding universities, it does not rely as much on that demographic. For instance, archival records indicated that a meaningful portion of S2's employees were otherwise employed, using HostCo as a source of secondary income. Also, given its proximity to urban areas, Location S2 tends to draw more affluent and professional applicants. This is one reason why the host organization has requested that analyses be done in such a way that any site differences in performance might be isolated and explained.

In contrast, Location NW has experienced much less stability. It was in operation between July and November of 2012, and then reopened in February of 2014. Interestingly, many employees that worked there in 2012 returned in 2014, including both call center managers. It was expected that a center located in an earlier time zone (Mountain versus Central) would help to alleviate some turnover, since one of the reasons many former employees listed for leaving was the late hours (until 12am CST; HostCo, personal communication). However, this effect does not seem to have occurred, and turnover was similar to the two southern locations in 2014. The final capacity of Location NW was about 85 interview stations. Thus Location NW is not weighted in the overall values to the same

level as the other two locations in YR14. Location NW is located in a mid-sized town that also hosts a small private university. Median income and other economic indices for the region are lower than those for the two southern locations as well, so wages tend to be lower. In addition, unemployment in the area is quite high, which means that HostCo can attract some employees who are likely overqualified for the job. The university also provides a labor pool for the site, similar to Location S1; however, students comprise a smaller percentage of employees overall.

HostCo's Physical Locations: Employees and Management

The number of interviewers at each call center was relatively stable from the beginning of YR13 to the middle of YR14. HostCo endeavored to maintain at least 70 interviewing stations at each site, but in the middle of YR14, S1 and S2 doubled capacity, maintaining approximately 140 interviewing stations from that point forward. According to archival records, between 90 and 120 employees are needed to staff 70 workstations based on HostCo's definition of "staffed" (station is "on" 6 days per week, each employee works at least the minimum number of days per week). This is largely due to the overwhelming number of part-time employees versus those employed full-time. The variation occurs because Location S1's definition of "minimum days" is one fewer per week than the other locations. This volatility in staffing creates a unique analytical challenge, since the standard practice of making the individual employee the primary unit of analysis is not likely to bear much fruit.

Management across the locations is consistent at a policy level, but an investigation of the individuals who filled those positions revealed that some intriguing interdependencies existed among them. For instance, the most senior manager across the sites is employed at

Location S1; this person possesses two years more tenure than the manager at Location S2, and four years more tenure than the two managers at Location NW. This manager also was responsible for training the managers located at Locations S2 and NW, although due to operational limits the training opportunities and methodologies were not standardized (HostCo, personal communication). While management styles (Blake, 1977; Blake & Mouton, 1980; Peris-Ortiz, Willoughby & Rueda-Armengot, 2012; Yukl, 2012) might vary among these individuals, this "inbreeding" of training should have had the effect of synchronizing the managers' perspectives on the business and its vision. Another point of similarity between the managers is that they appear to have been new to the polling industry before appointment. This may be indicative of a belief in the organization that management should be able to take an "outsider's view" of the business in order to minimize policy stagnation, but it is unclear from the archival data available.

However, one of these managers in Location NW was terminated in the summer of 2014 and replaced by another former supervisor from 2012. The 2012 center manager did not return in any capacity in 2014, thus that experience was lost.

The Role of Interviewer: Comprehensive Overview

HostCo has requested that the interviewer position should be the focus of all analysis. Interviewers are assessed on objective and subjective performance metrics as they are engaged in public opinion polling via live telephone interviewing. In the paragraphs that follow, the nature of the interviewer job will be outlined, the characteristics of the interviewer workforce will be explored, and the important performance indices that HostCo uses will be reviewed.

Hiring and Training Practices

HostCo does not utilize strict screening requirements for new employees in response to the turnover it experiences and the cyclical nature of the business model. At a minimum, however, applicants must demonstrate acceptable reading skills on a standardized test and demonstrate that they can be available to work enough hours to meet the minimum work requirements inherent in the job. Once hired, interviewers are required to complete a 6-hour training session. The curriculum primarily emphasizes the protocols that govern interviews based on organizational expectations. For instance, the training includes an exercise using a simulated project which allows each interviewer to conduct approximately 60-90 minutes of "live interviews" using randomly-generated numbers. A second point of emphasis involves what are called "rebuttal techniques", which are lessons on how to both initially persuade respondents to participate and to complete the study in a manner acceptable to the organization and its clients. Trainers are not required to complete any formal education on training methods or techniques, but each trainer does receive cursory information on the role from more experienced trainers before assuming a training group. Each training group varies in size, but HostCo attempts to keep the ratio of trainees to trainers at no more than 16 to 1. Each training program lasts for three days (one day of training and two days of close supervision on live projects).

Interviewer Tools: CATI Software

In order to provide a detailed overview of the job tasks required of interviewers, it is important to provide foundational knowledge about the technology that interviewers use to conduct interviews. The software is part of a class of programs called Computer-Assisted Telephony Interfaces (CATIs), which perform the menial tasks necessary for interviewing

(i.e., dialing, recordkeeping, data collection, etc.). The software manages the process of dialing numbers according to rules included with each project profile. Telephone numbers connected to hard-wired telephones (hereafter called "landlines") are pre-screened by the dialer software so that the human interviewer never receives a busy signal or a "no-answer." Telephone numbers connected to cellular services, however (about 30% of all interviews across projects), cannot be screened in this matter by law, so it is possible that interviewers will have calls routed to them that are answered by voice mail. As a result, projects that include cellular numbers will likely require more effort and time on the part of the interviewer to achieve the expected number of completed interviews. Cellular projects also come with other challenges, most importantly that respondents tend to be less responsive to the initial request for the interview (though HostCo reports that this is not as much of a problem today as it was in previous years). Dialing most often occurs on weekdays between 1700 and 2100 (local time), although quite often dialing will persist until 2200 (local time) despite steps that HostCo takes to curtail this behavior. Archival data suggests that the quality of data and production after 2100 drops sharply, making this hour of activity rather cost-inefficient.

Descriptions of Interviewer Tasks

Once the CATI software is able to locate a potential respondent, the human interviewer has a set of tasks that he must attempt to complete. First, respondents are greeted and the interviewer briefly identifies the purpose of the call. All of these interactions are scripted by HostCo in cooperation with the client who is purchasing the data. The CATI software acts as a teleprompter, displaying the script to the interviewer during the process. It also acts as a survey device; when the interviewer must record a response to a question, she

can do this by clicking the appropriate software option. Interviewers rarely will have access to a survey script prior to execution because of the rapid nature of HostCo's business and because many interview projects have multiple forms and branching logic. Also, multiple projects are run daily and simultaneously within the organization, requiring interviewers to shift at times from project to project. This means that a given interviewer will probably work on about 2-3 different projects per day typically.

While the above description of an interviewer's work may sound relatively simple, further investigation into the job tasks that interviewers complete revealed additional layers of complexity as well as the need for interviewers to possess "soft skills" and traits such as emotional intelligence and agreeableness (Consiglio, Alessandri, Borgogni, & Piccolo, 2013; Srisuthisa-Ard, 2014; Ybarra, Kross, & Sanchez-Burks, 2012). First, interviewers must be able to rapidly and effectively produce behavioral responses to a variety of unusual communications from respondents. For example, a respondent may directly or by implication suggest to the interviewer that it is acceptable to shorten the interview process artificially (i.e., "go ahead and mark all Republicans for every question - I don't need to hear them"). Interviewers must be capable of gently but firmly redirecting the respondent and explaining that organizational policy will not allow this. Second, the interviewer must be ready to handle skepticism and suspicion from respondents about the true motives behind the call. Due to an oversaturation in society of telephone marketing since the 1980's and the increasing frequency of automated interactive calling, some respondents will connect unanticipated calls requesting information as scams, intended to steal identities or perpetrate other kinds of fraud.

Third, interviewers must be ready to correct and re-educate respondents on misconceptions concerning the nature of telephone surveys. For instance, many respondents are aware of the national "do not call" (DNC) list and will cite it as a reason why the interviewer is trespassing on their rights. However, this list only applies to callers actively soliciting sales for products. HostCo does not adhere to the DNC list restrictions for operational purposes. Numbers are frequently provided to HostCo by its clients, and wholesale exclusions of DNC list respondents would reduce the available population in biased ways. However, HostCo does maintain an internal DNC list for respondents who request to be placed on it. Fourth, the identities of end users (HostCo's clients) are not disclosed to respondents, which can present additional resistance for interviewers to overcome. A few states, however, will not allow this. Finally, respondents frequently complain about calls beyond 2100 as well as frequent redials, according to HostCo, which places an additional burden on the interviewer attempting to complete a survey event, and respondents may give bizarre answers or spew out foul language. All told, there is much more to the role of interviewer than reading questions and marking answers.

Another important task that interviewers must practice is what could be called "reflection." If a response is provided that is not entirely clear or does not conform exactly to a possible response option, the interviewer must reflect that ambiguity back to the respondent in an attempt to obtain clarity. For example, consider a question that culminates with choices that include "strongly likely" and "somewhat likely." The respondent, however, simply says, "likely." Interviewers must recognize that an incomplete answer was given and redirect the respondent by repeating the answer choices and asking for clarification. If this occurs frequently during an interview, the respondent may begin to feel irritated at the constant

corrections, which forces the interviewer to manage this effect while still focusing on the integrity of the survey.

Another important aspect of the interviewer's job is the variation in survey subject matter. While the majority of clients that HostCo serves are political in nature, a good number of projects may target a variety of other topic areas. It is unrealistic to expect a given interviewer to be an expert on all of these various domains, which can create a difficult situation at times. For example, an interviewer may be asked by a respondent to explain certain aspects of a question, provide a personal opinion, or confirm that the nature of the question is confusing. Unfortunately, the interviewer cannot provide these services, but must be loyal to the script and redirect the respondents to answer to the best of their ability. This can also jeopardize the interview, since respondents may interpret the refusal to comply with their request as rude and uncooperative and decide to disengage from the process.

Another challenging aspect of the interview process is that the interviewer must be aware of what they do not know. A key example of this is interview length. The typical range of lengths of a given interview script is between 3 minutes and 20 minutes. Branching logic (included in most projects) and variations related to respondent behaviors are two common sources of variance that are largely unpredictable. Average lengths are negatively correlated with project volume as well. This is largely due to the seasonal nature of the industry; shorter surveys become the norm rather than the exception during the election season. Of course, many respondents will ask how long the interview will take as an implied negotiation for their participation. Since interviewers cannot provide a definitive length, they must develop a skilled response that appeases the respondents but does not violate the integrity of the process.

In light of all of these obstacles and challenges, it is remarkable that surveys get completed at all. In fact, archival data from HostCo shows that respondent participation rates (defined as the ratio between initiated and completed surveys) range from approximately 3 to 10%. Projects that utilize cellular phone lists often come in at the lower end of this range.

Macro-Organizational Policies Affecting Interviewers

As in any organization, there are forces that constrain the activities of employees that are defined at a more global level. Examples that industrial-organizational psychologists can readily identify include team-level dynamics (e.g., Bartelt & Dennis, 2014; Jex & Britt, 2008; Schippers, 2014), organizational climate and culture (e.g., Jex & Britt, 2008; Murphy & Cleveland, 1995; Verbeke, Volgering & Hessels, 1998; Winiecki, 2004), and extra-organizational factors such as regulatory bodies (e.g., Dean & Rainnie, 2009; Walsh & Deery, 2006). During negotiations with HostCo for this project, it became apparent that there were important factors surrounding the interviewers that provided context for a more complete understanding of the job. The following section briefly outlines some of the more salient ones.

First, interviewers do not have any advance knowledge concerning the nature of the project(s) they must engage in a given workday. This reality serves to increase the perception among interviewers that the control of work is not theirs, but is instead centralized at higher levels. Project assignments are provided to interviewers by floor supervisors; they receive the assignments from call center managers who are in turn guided by a single project director (this person was the Location S1 manager as far as this thesis is concerned).

Second, there is considerable variability in work hours. Work shifts average 6.5 hours according to company data, but can be as long as 9 hours on many occasions. Interviewers

are expected to remain at their stations waiting for incoming calls throughout the work shift except for designated break periods. They are permitted to sit or stand, but they may not engage in reading material (oddly, crossword puzzles are allowed) or use their personal phones when "on task." Beverages are permitted at workstations provided they are in spill-proof containers, but food is not allowed on the floor. Often production levels and project demands on a given day do not warrant the need to retain all employees throughout the entire work shift. Management reduces the employee count based primarily on merit; however specific skill factors may apply as well, such as fluency in Spanish. Employees are not permitted to leave early on their own accord without management consent. The combination of these rules further enhances the perception of administrative control of work among the employees.

The factors outlined above have an impact on the analyses requested by HostCo. For instance, the organization maintains an operating policy about projects that determines how long each one should be "live" in a given day. Usually, a minimum time of three hours per shift is the goal, but it is common for this time to be extended, sometimes by as much as a factor of two. It is also common for a project to be "live" on multiple occasions within a given work day and for interviewers to be assigned to a project multiple times. This asynchrony between project and employee lifespans within each day creates the need for interviewers to flow fluidly in and out of project modes, creating a challenging work environment where each interviewer must be ready to shift into a different content domain and ask questions in such a way to create the impression of expertise and professionalism. Also, a given interviewer may not complete an entire survey with one respondent; this creates an interesting data problem in that interviewers cannot be mapped in a unitary fashion

to each data point. In short, discussions with HostCo and initial evaluation of archival data made it clear that defining a relatively standard performance metric, much less analyzing it inferentially, would be difficult.

Defining Performance at HostCo

Key Objective Performance Measures

According to HostCo, interview performance is tracked in real time with customized software that interfaces with the various CATI applications. While there are many metrics available in the database provided to us by HostCo, a measure called "completes per hour" (CPH) is clearly an index that the organization sees as paramount. In the method section of this paper, the mathematical realities of the CPH measure will be explained in detail, but essentially it is intended to be a ratio between the number of completed surveys credited to an interviewer and the amount of time needed to obtain those completed surveys, similar to an interviewer's "batting average". For example, consider an interviewer who is credited with 3 completed surveys over the course of 4 work hours on that project. The CPH value is easily calculable (3 surveys / 4 hours = 0.75 CPH). However, upon closer examination of this index within the context of the work processes outlined above, it became clear that the data were not quite so simple.

One methodological issue concerns the way in which an interviewer is "credited" with a completed survey. It is common for multiple interviewers to complete a single interview if the respondent is engaged for multiple sessions over time. HostCo reports that the interviewer who *finished* the interview is credited for it, regardless of the amount of actual time this person spent on the survey. On the surface, especially to organizational outsiders, this appears to be a very unfair system. However, given the way in which business

must be transacted given the organizational realities of the work, HostCo argues that there is no better way to index interviewer performance. Nevertheless, this further complicates the definition of performance, which is the primary focus of this consulting project.

A second issue has to do with CPH goals that employees receive and how they are perceived. Employees are instructed during their initial training that they will be expected to meet target CPH numbers for each assigned project. These values are designated daily on a project-by-project basis by executive staff using a complex algorithm of factors, the most significant of which are called "file type" (referring to landline vs. cellular records) and estimated survey length. These values, unfortunately, are estimates, and they frequently do not reflect the actual data once dialing is completed. Call centers will infrequently finish projects simultaneously and nightly production sheets are executed according to each center's various demands. Managers and supervisors can see these CPH statistics across the company in real time, but HostCo does not allow other employees to do so. Monthly evaluations are based on comprehensive company-level data which is updated as all locations submit final reports. By contrast, in order to avoid making employees in one center wait until all centers have finished work, CPH statistics are posted publicly only for that particular location. Summary sheets are posted daily for all projects that show CPH goal progress for all interviewers, which allow employees to see to what extent the goal estimates were "missed" by management. The goal literature suggests that goal-based motivation is strongest when the goal is a good match to the actor and to the environment (Baumeister & Bushman, 2013; Minjung & Fishbach, 2014); in the case of HostCo, this match appears to be almost impossible to achieve.

Nevertheless, policies indicate that the organization take CPH goals very seriously. Center managers and supervisors execute real-time evaluation of CPH data and will advise interviewers if their performance is poor. Usually this entails a "counseling session" of some kind in the form of a verbal interaction intended to make the interviewer aware of the deficit and motivated to alleviate it. These remedies can extend to the project workforce at large if its overall performance is below expectation. Counseling can be triggered for everyone assigned to a given project in such circumstances. Intriguingly, while the organization's administration seems to be clear about what levels of performance will trigger corrective action, it is not apparent that the interviewers are. The most likely hypothesis is that they are left to figure it out on their own, which creates the perception of an additional lack of control among them and could encourage a focus on numbers over quality. The organization makes reasonable efforts to limit this lack of knowledge, but operational constraints frequently do not allow for the delivery of a clear message.

Formal disciplinary action, triggered by unacceptable CPH data for a given interviewer, is not initiated until that employee has completed 30 calendar days of employment, all things being equal (this is different than actual work days). This amounts to a kind of probationary period. Variation in this policy is possible according to HostCo; hire dates, the severity of the poor performance, and/or the "business" of the season can reduce or extend the 30-day benchmark. All formal performance reviews are typically conducted during the first week of each calendar month, but since hiring and turnover is frequent and very fluid, this policy cannot always be adhered to. For instance, an interviewer hired on the 15th of a given month may not receive a review for 45 days or more. If the organization determines that the interviewer is not capable of the work due to persistent performance

failures, s/he may be terminated at the conclusion of a formal review by policy. HostCo reported that this is a relatively infrequent occurrence (HostCo, personal communication).

Finally, it was striking that no indication existed in the archival documents of any reinforcement-based approaches to employee motivation (i.e., tangible rewards, bonuses, etc.). All interactions with interviewers were based in either looming punishments or illusory social comparisons (i.e., "the other sites are making you look bad", etc.). This is unfortunate given the vast literature on reinforcement theory and its greater utility in real-world motivational situations (Mitchell, 2007; Pate, 1978; Vlaev & Dolan, 2015).

Key Subjective Performance Measures

HostCo has also spent time and energy in the development of a performance rating system that is independent of the objective measures, such as CPH. In its current form, it is a relatively recent addition to the organization's activities and has added a degree of quality control to the evaluations of interviewers. In this section, the elements of this system will be described, from their development to the processes that currently exist for their implementation.

Developmental Background. The organization originally established formal written rubrics for six subjective dimensions of performance in 2010. Precedent for the dimensions that were eventually identified already existed in HostCo's operational policies, but there was a distinct lack of formalization in both definition and measurement that limited usefulness. Call center managers, trainers, incumbent quality control specialists (QCSs), and executive staff were all involved in the discussions as part of a task force that led to the initial development of the measures. All members of the task force had served in the capacity of

telephone interviewer for this organization, even though some had not actively been conducting interviews for several years.

The entire development process lasted for several months, directed by an internal consultant with experience in developing performance evaluation instruments. First, a rudimentary job description was written for the interviewer position using archived training materials and subject matter expert (SME) input. The project leader reviewed dozens of random recordings to develop the initial rubric instruments in addition to a review of previous training materials and trainer input. Next, other SMEs provided feedback over multiple sessions, leading to the final evaluation structure. Frame of reference training was designed and established by the project leader and subsequently given to all QCSs and call center managers. In 2012, the organization decided to conduct a revision of the process. This was triggered by several factors, including changes in the composition of surveys, feedback from the first two years of use, and the growing influence of mobile phones. The original task force leader was also in charge of the revision. The process was similar to that from 2010, although shorter. Upon its completion, new frame of reference (FOR) training was once again provided to all QCSs and call center managers. The performance appraisal literature strongly supports the use of frame of reference (FOR) training to limit variance in rating levels (Murphy & Cleveland, 1995).

Beginning in late 2012, however, changes in operational realities within HostCo prevented the organization from conducting FOR training for new QCS employees. Only six of the 22 QCSs that will be included in these analyses received this FOR training. The remaining reviewers were hired after the last training occurred and were subsequently trained informally by peers, some of whom did not receive the original FOR training themselves. As

far as can be determined, no specific directions about how to train QCS employees were documented beyond 2012. This is disappointing, since we know that managers may be prone to using heuristics and cognitive modeling that incorrectly attributes universally good or bad perceptions across multiple metrics against individual employees based on a single metric (e.g., Nathan & Lord, 1983). Further, ratings may fail to capture true scores because critical events are unclear or not sufficiently defined so that raters can detect them (Matthews, Davies, Westerman & Stammers, 2000).

The Six Dimensions and Associated Ratings. Reviews are conducted anonymously by QCS employees by listening to live interviews. Limitations inherent in the CATI software only permit QCSs to conduct reviews of interviewers from their home location. The six dimensions of evaluation are loosely divided into two conceptual categories: data validity and organizational image. Within the data validity category are three finite dimensions: **verbatim** (*was the script followed exactly?*), **assuming** (*did the interviewer avoid making assumptions about the respondents' answers?*) and **leading** (*did the interviewer avoid suggesting answers to the respondents?*). Within the organizational image category are the final three dimensions: **pronunciation** (*did the interviewer correctly pronounce all words?*), **pace** (*did the interviewer work at a brisk but comfortable speed?*), and **professionalism** (*did the interviewer portray the organization favorably at all times?*).

The QCSs are responsible for scoring each dimension for every interview that they review. The organization uses a 5-point scale that is treated as interval-level but is often discussed and evaluated ordinally. A score of "4" indicates that the interviewer met minimum expectations. Any score less than that is unacceptable and has its own label attached to it (3 = minor issues detected; 2 = major issues detected; 1 = unacceptable performance). Any ratings

of "1" will usually result in the deletion of that interview from its associated project. A score of "5" indicates superlative performance, but HostCo indicated that these seem to be unusual due to a high standard (HostCo, personal communication). The QCS reviewers are provided with memory supports as they are able to see the descriptions of the dimensions as they listen to interviews; in fact, policy dictates that they must review the dimension descriptions before every review session, regardless of rater experience. This is potentially important because the performance dimensions can only be somewhat specific given the wide variety of possible behaviors that QCS reviewers might encounter. By refreshing the organization's definitions of the dimensions in the mind of the QCS, HostCo is attempting to minimize "definitional drift" which would jeopardize the integrity of the evaluations.

One way to understand these dimensions is to highlight ways in which interviewers may fail to meet the standards set forth in them. Violations classified as "assuming" are almost always indicative of an interviewer who inappropriately reacts to a respondent's failure to provide a clear answer from among the answer choices. Violations classified as "verbatim" are usually the result of attempts by the interviewer to artificially shorten the survey length, either due to protests from respondents or pressure to achieve a particular CPH goal. Violations classified as "leading" are usually less obvious and are often in response to the behavior of respondents that trigger the leading questions or statements. Note that these descriptions are of *typical* violations; the actual set of violations within each category is expansive and very difficult to define completely.

This ambiguity is even more applicable to the three "organizational image" dimensions. For instance, pace is almost entirely left to the judgment of the QCS. There are no established time ranges or pacing guidelines. Professionalism also requires a considerable

amount of subjective inference, since it is often unclear if what the QCS perceived as "unprofessional" was perceived by the respondent in the same manner. Also, a given interview may suffer from technical issues or a respondent may behave bizarrely. Neither of these events are the fault of the interviewer but they could contribute to perceptions of unprofessionalism in retrospect. Finally, pronunciation ratings also present an interesting challenge because sometimes words have more than one correct form or pronunciation. Also, a poorly-pronounced word could lead to a change in question meaning (e.g., "county' vs. "country"), which would make this error a "verbatim" violation rather than pronunciation. A true pronunciation error occurs if the interviewer said "axessed" instead of "accessed". The meaning of the word is not affected, but the professional image of the company may be tarnished.

Critical Confounds in Rating the Dimensions. HostCo revealed that it was very interested in an analysis of these ratings because of several possible ambiguities and confounding realities that could affect the nature of the data. Some of these are outlined here. First, the organization is aware that certain behaviors or verbiages could be theoretically assigned to multiple dimensions, leaving the QCS with a decision to make about where to assign them. For instance, it is possible for an interviewer to say something to a respondent that could be seen as both "assuming" and "leading", or even as a "verbatim" violation. Second, the workforce across all three locations is quite diverse in ethnicity and education. Some pronunciation issues may be tied to these ethnic and lingual differences, especially regarding names of people and places which are common in political surveys. Some names present particular challenges to employees who lack exposure to non-English phonemes, particularly names with High Germanic, French or Native American origins. Finally, each

rating is linked to an organizational action (discussed in the next section), which could encourage QCS reviewers to use scores that may not necessarily reflect the data that was reviewed but a desire to avoid organizational actions that would result from certain scores.

Organizational Actions Connected to QCS Ratings. HostCo indicated that the way in which interviews are selected for evaluation is random, but it seemed to be better characterized as "pseudo-random." Employees who are identified as "underperforming" based on objective measures (e.g., CPH) are more likely to be selected for review, since development is one core goal of the subjective evaluation process. However, higher producers are also selected more often because the organization maintains a per-project review quota and because interviewers who complete more surveys are likely to have more available interviews to review. By company policy, interviewers are notified when one of their interviews have been reviewed by a QCS, regardless of outcome. Feedback is also provided to interviewers on a regular basis at the discretion of each call center manager.

Superlative performance is always rewarded with verbal praise (and little else), but scores of "3" or below will always result in administrative action that can impact both the interviewer and the QCS. Thus, there appears to be a salient social and motivational cost for assigning anything other than ratings of "4", since that is the expected level of performance. The company operates under a developmental paradigm where managers are expected to assist employees in meeting performance and quality standards. Formal disciplinary action is only triggered after several attempts at remediation have been made. Basic counseling and coaching is the responsibility of floor supervisors who have some discretion within operating policy as to how to address the developmental needs of the employee. Call center managers may intervene in response to more severe or chronic failures. Regardless of how the

interventions take place, it was clear based on policy that low ratings on the subjective measures frequently cost the organization time, money and resources that had to be borrowed from the primary task of survey completion.

Research Questions

The two research questions that follow were developed with the dual purpose of providing insight to HostCo's business operations as well as to link this study to existing theory. These were selected from a list of questions directly requested by the client organization and based on an initial overview of the organizational realities of the situation. Effort has been made to ground the nature of the questions within relevant organizational theory.

The first research question is focused on the nature of interviewer performance over the first 30 workdays of activity. The host organization is very interested in how quickly employees "get" their new jobs and how they improve on their objective performance measures. We know that skilled performance often follows a power law, beginning at relatively low levels of accuracy and then improving monotonically to an asymptote (e.g., Neves & Anderson, 1981). It is expected that the same pattern will emerge here. However, HostCo is also interested in knowing at what point the interviewers' performance starts to reach that asymptotic level, and if that point varies depending on the location analyzed. Therefore, the first research question includes the following components:

> **RQ1: (a) Do employee production levels remain relatively stable after an initial period of acclimation? (b) How many work shifts does it take a new employee to stabilize? (c) Is this point estimate similar across employees? (d) Do location differences exist?**

The second research question will analyze in depth the data gathered from the subjective performance evaluation system. HostCo is interested primarily in the behavior of each QCS rater across the interviews they rated, and whether there is evidence of errors in rating that can be uncovered. It was also mentioned that there may be differences between locations in how leniently or severely each dimension is being evaluated. The literature provides some reason to hypothesize that location differences may exist given the tendency of groups of employees to become more homogeneous in their values and perceptions over time (Baumeister & Bushman, 2013; Murphy & Cleveland, 1995). If we consider the behavior of rating interviews as skilled behavior, then much of the logic applied to RQ1 is germane to this question as well, so ratings should become less variable as raters become more experienced, probably coalescing around the "4" score based on the social and organizational costs discussed previously. However, it is unclear whether the dimensions are clearly defined enough for this to occur, especially with respect to the three dimensions in the "organizational image" category. Given the frequency of reviews, QCS data will be evaluated by aggregating across multi-day intervals.

RQ2: (a) Do Quality Control specialists have similar quality ratings? (b) Are location differences present for the quality measures? (c) Are there trends in Quality Measure exceptions by workday intervals? (d) Do the six measures trend in the same ways between workdays?

Methods

Characteristics of the Data

Written permission has been provided by HostCo to use their data for these purposes. Datasets were taken from January 1, 2013 to December 31, 2013 (YR13) and from January 1, 2014 to October 31, 2014 (YR14) for RQ1. For RQ2 a single dataset comprised of quality ratings from this date range were extracted. Subsets were drawn as appropriate to each individual question and are defined in those sections. The master dataset contained over 200,000 production records, 45,000 quality records across more than 3,000 interviewers, three physically discrete call centers (S1,S2, and NW), over forty quality-control specialists (QCS), and four call-center managers.

Important Dependent Variables for RQ1

Analysis was planned for performance and quality metrics over interviewer tenure time. Tenure was defined as the number of completed shifts. In order to more appropriately capture actual skill progression and address confounding variables, a method was utilized to assign tenure in ordinal days rather than calendar days. Further details about these methods are located later in this document. The primary objective production measure was completed surveys per hour (CPH).

Completed surveys per hour (CPH) is a mathematical combination of total hours on a given project (H_p), time spent on breaks (T_b), and the total number of completed surveys on a given project *per day* (C_{pd}). The formula is constructed as follows:

$$CPH = H_p - (T_b / C_{pd})$$

Break time represents official breaks only, and the company only considers a totally executed survey as complete. Surveys are commonly conducted in multiple sessions and it is not

atypical for different interviewers to conduct each part, so the person who finishes the final portion of the survey receives full credit for the completion. While this reality does eliminate some analytic approaches, these situations are so common that it is not feasible to remove them from the data completely.

The number of completed surveys are extracted directly from data collection software (Command Center) and inserted into the production operations software (E-Manager) as each project is completed on a daily basis. Shift supervisors enter the hours worked on each project on a per interviewer basis which is compared against the timekeeping software (Timeforce). Break time is excluded from all entries. The software then calculates values of CPH for each employee.

Important Dependent Variables for RQ2

In addition to production metrics, the organization is heavily focused on answer quality. As noted in the introduction, six quality metrics (*leading, assuming, verbatim, professionalism, pace,* and *pronunciation*) are measured by quality control specialists (QCS) who quasi-randomly sample recorded calls. The first three measures (leading, assuming and verbatim) focus on avoiding inappropriate behavior that may bias survey data. Collectively they are *validity measures* because a violation of the first three types puts the validity of the survey at risk. Severe violations warrant deletion of the survey data. Some actions result in violations of more than one measure or frequently cascade into multiple violations.

Assuming errors occur when an interviewer selects a response that the respondent did not clearly indicate. Leading errors emphasize issues of bias with language and paralanguage, especially emphasis. It can also pertain to editing the script in a manner that creates bias. Verbatim errors occur when a deviation is made from the written script. It can include

omitting questions, omission/replacement/expansion of parts of questions, prompts and response choices or incorrect ordering of response choices as they are presented on the screen.

The second three measures (professionalism, pace, and pronunciation) target the company's image, legitimacy and professional conduct that can be inferred through interviewer actions. Errors of these types usually do not warrant data deletion. Professionalism is a catch-all category for use with anything that doesn't clearly fall into another area. It also addresses excessive use of transitions or the inclusion of inappropriate transitions that do not bias outcomes. Pace refers to reading the survey at the appropriate speed. This is an adaptive standard and the correct pace varies from respondent to respondent. It is common for an interviewer to have to vary pace throughout the survey based on respondent demands. Pronunciation errors occur when interviewers mispronounce words in a manner that does not change the meaning.

Raters utilize a 5-point scale to rate behaviors on recorded interviews, as described in the introduction. Each rating is an aggregate impression across the entire interview and should not be interpreted additively. For example, one interviewer may receive a score of "3" on a dimension due to a small number of important errors, whereas another interviewer might receive the same score for more frequent but less impactful violations.

General Procedures

Each research question, due to the complexity of the underlying data, required its own specific data screening and preparation strategy. Therefore, instead of describing these activities in this section (which is the usual practice) they will be outlined as each research question is presented in the next section.

Results

Research Question 1

The data utilized for these analyses were quite complex, and readers unfamiliar with the structure of the data (i.e., anyone not employed by the source company) may have difficulty following some of the strategies employed if the section were written in the traditional style of psychological research. Additionally, each research question upon which this investigation rests required a different cross-section of the larger dataset as well as its own unique data screening process. Therefore, this section is strictly organized into sub-sections by each research question, and each sub-section will include data screening procedures, pertinent sample characteristics, important descriptive data, and any inferential analyses that were utilized.

RQ1 - Sample Characteristics and Data Screening

These questions required data on the number of completed surveys per hour (CPH) for interviewers over the first thirty working days for two operating years, 2014 (YR14) and 2013 (YR13). Employees only generated production records on "live" projects. For the majority of employees the first production records occurred on the second actual working day since the first day is spent on a training project. This analysis classified the second working day as the first working day, unless the employee worked on "live" projects on his or her training day, which rarely occurred. Three physical business locations (Sites S1, S2, and NW) were included in the 2014 data, while the 2013 data included only two of these three (Sites S1and S2).

The concept of the "thirty workdays" was created for this set of analyses and may be difficult to grasp; thus, a detailed description follows. Interviewers and projects and calendar

dates made up each workday. Each interviewer record was assigned a workday value based on the number of days that interviewer had worked previously on projects. Therefore, "workdays" corresponded to many calendar days rather than single dates. For example, employees at Site S1 reached Workday 30, on average, after 61 calendar days; at Site S2, after 50 calendar days; and at Site NW, after 56 calendar days. Employees were continuously hired throughout the year; thus, each given workday was made up of large sets of individual calendar days (see Table 1).

For YR14, unique days comprising a workday ranged from 120 to 197 (M = 172.5, SD = 19.45). For YR13, the range was from 92 to 164 unique days (M = 131.23. SD = 22.47). The number of project worked per interviewer is also important to gain a basic understanding of these data displayed in Table 2. The data show a consistent pattern if interviewers being involved in multiple projects simultaneously and the number of project stayed relatively consistent across the 30 workdays studied in both years.

Since each workday was made up of a wide range of dates throughout the review period, a large number of different projects were also reviewed for quality on each workday. While projects (telephone surveys) typically run for 2 to 3 days, it was necessary to consider each project day as a separate project. Therefore, a 3-day project was represented as three separate projects. This allowed for the data to be consistent with the core data structure of the "workday" concept. After these adjustments, a YR14 project/workday range of 511 to 1036 was observed (M = 876.42, SD = 139.25), whereas a Y13 project/workday range of 223 to 469 (M = 373.32, SD = 72.61) was observed. Finally, it is worth noting that the number of

Table 1

Composition of Workday by Project/Interviewer/Dates

Workday	*Unique Projects*		*Unique Interviewers*		*Unique Dates*	
	YR 14	*YR 13*	*YR 14*	*YR 13*	*YR 14*	*YR 13*
D1	511	252	1084	616	120	102
D2	592	319	971	536	140	115
D3	914	399	900	495	174	140
D4	951	442	837	468	175	158
D5	1036	419	791	446	195	153
D6	1012	455	743	418	197	161
D7	997	450	706	398	195	164
D8	1026	453	673	377	193	154
D9	1032	432	645	362	197	148
D10	1014	469	610	342	194	160
D11	962	467	590	327	186	152
D12	968	411	572	312	196	150
D13	953	438	556	297	193	147
D14	967	401	539	283	187	147
D15	973	410	523	268	187	151
D16	946	390	507	260	187	151
D17	899	376	486	248	175	143
D18	909	361	472	231	169	131
D19	875	354	460	217	172	122
D20	861	335	450	207	170	123
D21	831	336	430	197	165	119
D22	815	322	408	189	164	117
D23	813	331	394	182	162	112
D24	771	314	375	171	160	114
D25	762	304	362	166	158	103
D26	729	281	344	160	155	108
D27	728	283	328	152	157	101
D28	693	249	318	145	156	101
D29	673	245	304	138	149	98
D30	668	223	291	132	147	92
Group M	876.43	373.32	574.07	302.50	172.50	131.23
Group SD	139.25	72.61	203.70	128.35	19.45	22.47

Table 2

Number of Projects Completed Per Workday Per Interviewer

Workday	***YR 14***		***YR 13***	
	M	*SD*	*M*	*SD*
D1	2.40	1.21	1.82	0.94
D2	2.48	1.17	1.96	1.01
D3	2.84	1.30	2.02	1.05
D4	2.79	1.20	2.06	1.03
D5	2.76	1.22	1.95	0.98
D6	2.65	1.23	2.02	1.02
D7	2.67	1.21	2.03	1.01
D8	2.81	1.25	2.06	1.03
D9	2.78	1.24	2.10	1.05
D10	2.74	1.22	2.13	1.08
D11	2.83	1.26	2.17	1.10
D12	2.78	1.28	2.05	1.07
D13	2.76	1.28	2.24	1.15
D14	2.93	1.24	2.16	1.14
D15	2.88	1.25	2.22	1.13
D16	2.80	1.30	2.03	1.08
D17	2.90	1.31	2.12	1.05
D18	2.91	1.35	2.19	1.13
D19	3.01	1.37	2.21	1.05
D20	2.86	1.27	2.21	1.05
D21	2.86	1.35	2.34	1.15
D22	2.91	1.34	2.21	1.10
D23	2.98	1.27	2.37	1.20
D24	2.94	1.34	2.47	1.27
D25	2.83	1.28	2.52	1.28
D26	2.87	1.31	2.28	1.20
D27	2.90	1.36	2.39	1.11
D28	2.87	1.27	2.21	1.17
D29	2.98	1.32	2.20	1.18
D30	3.00	1.41	2.18	1.14
Group M	2.81	1.28	2.16	1.09
Group SD	0.14	0.06	0.15	0.08

unique interviewers evaluated per workday declined rapidly as shown in Table 1 (Unique Interviewers). Therefore, analysis power diminished across workdays, a common occurrence in field data due to participant mortality, in this case largely due to all forms of employee turnover.

Several important data screening approaches were necessary. First, in early November of 2014, the company reduced staff considerably; therefore, only the YR14 data up to October 30, 2014 was used to reduce the influence of this event. There was no need to edit the data in this way for YR13. Second, the analysis was restricted to include employees hired on or after the starting date for each database and those hired up to 30 days prior to the ending date for the database. After these adjustments, the data for YR13 was comprised of 616 interviewers, while the data for YR14 included 1,084 interviewers. This is due to the seasonal nature of the business; even-numbered years tend to be higher in call volume. Third, one regular aspect of these data, regardless of year, is a significant turnover ratio. In YR14, 793 employees left the organization before or on Workday 30 (73.2%). In YR13, this number was 484 employees (78.6%). This reality made it unreasonable to examine interviewer behavior beyond the 30-day milestone.

Fourth, datasets include information from calls made on landlines as well as cellular lines. One project may include numbers of both types. When these data were collected, it was possible that an interviewer might "cross-login" so that she appeared to be working on both types of lines simultaneously. This resulted in some impossible records being created (i.e., 3 completed surveys, each of about 10 minutes, finished in less than 15 minutes). A search for these "impossible" results allowed for the removal of 237 records. Fifth, it was possible that an interviewer might have completed the final few questions of a survey that another

interviewer left unfinished, artificially inflating the first interviewer's performance statistics. Also, supervisors frequently completed the final portions of surveys at the end of a shift. To account for these practices, 537 records were tagged and removed from the database.

RQ1 - Data Analysis

In order to examine how employee performance changed during the 30-day work period, an unbiased dependent variable was constructed from the raw data on completed surveys per hour (CPH): the **weighted length modification factor (WLMF).** The rationale for and the calculation of this variable is described below.

First, CPH values are directly affected by survey length per project, which can range from 3 to 20 minutes. To remove bias due to this length factor, z-scores (Z_p) were calculated on a per project basis and then averaged for each interviewer for each workday:

$$CPH_{mod} = (Z_{p1} + Z_{p2} + Z_{p3} + Z_{px}) / \text{Projects Worked}$$

Any calculated value that fell beyond a specified range ($-3.5 < CPH_{mod} < 3.5$) was changed to reflect that limit (i.e., a value of "4" was changed to "3.5"). This affected less than 1% of overall records in either year.

Next, a weighted length modification factor (WLMF) was calculated by taking the z-score for each project for a given interviewer per workday (Z_p) and adjusting it by the number of hours (H_n) worked on the project:

$$WLMF = ((Z_p * H_1) + (Z_p * H_2) \ldots + (Z_p * H_n)) / \sum (H_1..H_n)$$

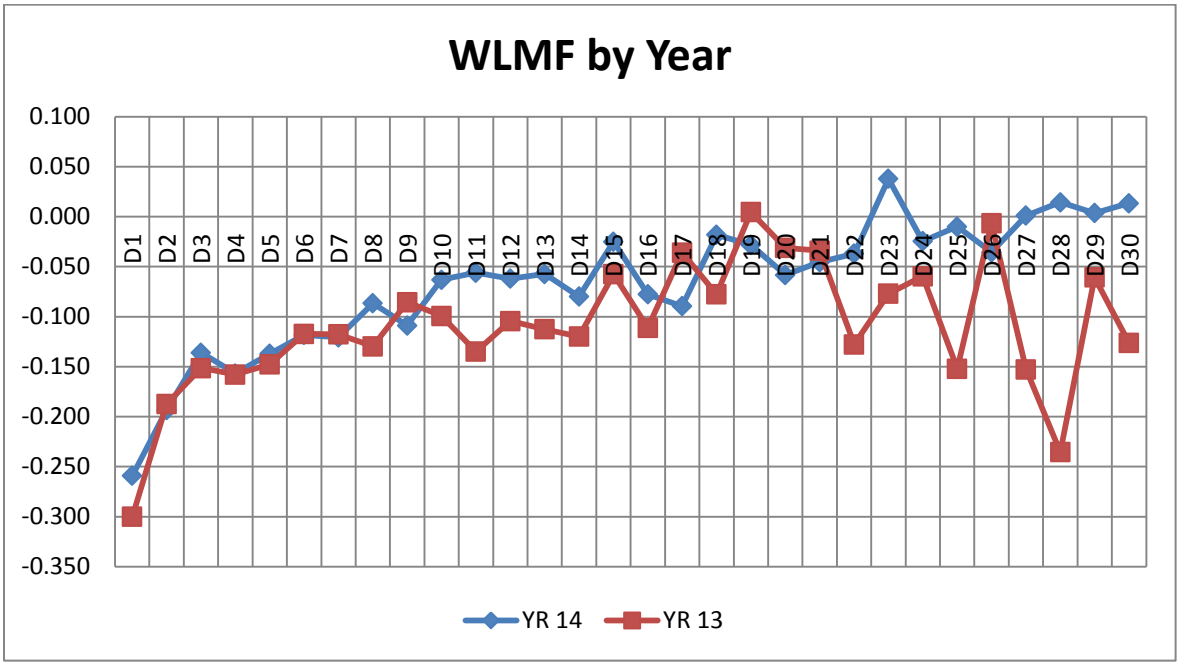
In the formula, each term in the dividend is a separate project. Within interviewers, LMF will be equal to WLMF if only one project was worked for a given workday or if no length variability occurred across worked projects. Since some days were comprised of only a single

Table 3

Pairwise Means and Standard Deviations of LMF and WLMF for YR 14 & YR 13

	YR 14 CPH				***YR 13 CPH***			
Workday	*LMF*		*WLMF*		*LMF*		*WLMF*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
D1	-0.256	0.661	-0.259	0.639	-0.288	0.620	-0.300	0.591
D2	-0.202	0.625	-0.193	0.625	-0.214	0.622	-0.187	0.611
D3	-0.132	0.595	-0.136	0.601	-0.171	0.641	-0.151	0.644
D4	-0.156	0.592	-0.157	0.590	-0.176	0.646	-0.158	0.636
D5	-0.152	0.634	-0.137	0.627	-0.173	0.678	-0.148	0.678
D6	-0.124	0.635	-0.118	0.622	-0.123	0.708	-0.117	0.676
D7	-0.139	0.649	-0.121	0.641	-0.121	0.666	-0.118	0.648
D8	-0.093	0.610	-0.087	0.585	-0.129	0.647	-0.130	0.647
D9	-0.107	0.656	-0.109	0.634	-0.082	0.731	-0.085	0.718
D10	-0.065	0.664	-0.063	0.657	-0.107	0.686	-0.099	0.662
D11	-0.069	0.647	-0.056	0.624	-0.147	0.658	-0.135	0.663
D12	-0.066	0.657	-0.062	0.627	-0.133	0.662	-0.104	0.658
D13	-0.059	0.685	-0.057	0.676	-0.107	0.655	-0.113	0.644
D14	-0.087	0.629	-0.080	0.609	-0.149	0.679	-0.120	0.668
D15	-0.055	0.656	-0.025	0.656	-0.074	0.690	-0.058	0.695
D16	-0.077	0.664	-0.078	0.640	-0.107	0.682	-0.111	0.671
D17	-0.111	0.592	-0.089	0.575	-0.036	0.703	-0.036	0.710
D18	-0.007	0.605	-0.018	0.593	-0.106	0.739	-0.078	0.703
D19	-0.034	0.629	-0.028	0.625	-0.022	0.718	0.005	0.697
D20	-0.060	0.579	-0.058	0.585	-0.063	0.663	-0.031	0.627
D21	-0.058	0.664	-0.046	0.649	-0.046	0.723	-0.034	0.708
D22	-0.054	0.637	-0.037	0.629	-0.125	0.711	-0.128	0.693
D23	0.014	0.603	0.038	0.591	-0.061	0.708	-0.077	0.699
D24	-0.037	0.648	-0.025	0.642	-0.007	0.777	-0.060	0.709
D25	-0.007	0.664	-0.010	0.640	-0.156	0.637	-0.152	0.645
D26	-0.037	0.650	-0.036	0.637	-0.013	0.709	-0.007	0.690
D27	0.010	0.637	0.001	0.619	-0.171	0.633	-0.153	0.629
D28	-0.007	0.707	0.014	0.687	-0.238	0.663	-0.235	0.639
D29	-0.034	0.663	0.004	0.642	-0.073	0.674	-0.061	0.652
D30	0.032	0.637	0.013	0.636	-0.119	0.682	-0.126	0.649

Figure 1.
WLMF Trends for 2013 and 2014 Across the Thirty Workday Period



project/interviewer, extreme WLMF scores were still possible, so score adjustment based on range was again utilized (-3.5 < WLMF < 3.5). Table 3 shows descriptive statistics relevant to this unbiased measure, and Figure 1 displays the trend in WLMF graphically over the 30 workdays studied.

As seen in Figure 1, YR14 WLMF values increased until about the middle of the time interval, where they reached an asymptote. As shown in Table 4, an examination of the WLMF data for YR14 revealed a significant within-days main effect (F (1, 290) = 40.81, η^2=.12). For YR13, the increase in WLMF was less obvious and the asymptote seemed to occur earlier in the studied time interval. Additionally, YR13 data revealed more variation in WLMF values at the end of the time interval. However, an examination of the WLMF data for YR13 also revealed a significant within-days main effect (F (1, 131) = 4.13, η^2=.03), displayed in Table 5. Thus, data for both years show a significant increase in performance followed by an asymptote.

Table 4

YR14 Repeated Measures ANOVA For Work Day Intervals

	SS	*df*	*MS*	*F*	*p*	η^2
Overall	21.214	1	21.214	40.813	< .001	0.123
Error	150.737	290	0.52			
Day 1-6	3.251	1	3.251	10.027	0.002	0.013
Error	239.943	740	0.324			
Day 7-12	1.24	1	1.24	4.108	0.043	0.007
Error	172.273	571	0.302			
Day 13-18	0.168	1	0.168	0.589	n.s.	n.s.
Error	134.231	471	0.285			
Day 19-24	0.171	1	0.171	0.616	n.s.	n.s.
Error	103.493	374	0.277			
Day 25-30	0.141	1	0.141	0.508	n.s.	n.s.
Error	80.411	290	0.277			

Note: N values vary due to employee mortality.

Table 5

YR13 Repeated Measures ANOVA For Work Day Intervals

	SS	*df*	*MS*	*F*	*p*	η^2
Overall	2.113	1	2.113	4.129	0.044	0.031
Error	67.027	131	0.512			
Day 1-6	4.4415	1	4.415	12.691	< .001	0.03
Error	145.052	417	0.348			
Day 7-12	0.012	1	0.012	0.035	n.s.	n.s.
Error	104.263	311	0.335			
Day 13-18	0.061	1	0.061	0.174	n.s.	n.s.
Error	81.343	230	0.354			
Day 19-24	0.625	1	0.625	1.624	n.s.	n.s.
Error	64.4	170	0.56			
Day 25-30	0.022	1	0.022	0.076	n.s.	n.s.
Error	38.416	131	0.293			

Note: N values vary due to employee mortality.

With respect to Part (b) of RQ1, in order to determine the point at which the observed linear effect vanishes, the 30-day time interval was broken into five sequential intervals of six workdays each. Stabilization of performance would be indicated at the interval where the significant main effect was no longer observed. Tables 4 and 5 show the results of separate ANOVAs on each interval for YR13 and YR14 data. For YR14, the significant effect disappears at the third interval (Days 13-18), whereas for YR13, the effect disappears at the second interval (Days 7-12). Therefore, it is reasonable to conclude that, across the entire database represented in these analyses, the company can expect that performance will generally asymptote at or around Day 12.

Regarding Part (c) of RQ1, no significant differences were observed in WLMF values between call center locations across the entire time interval studied in either YR13 or YR14. However, when the sample was divided into the 6-day intervals as previously described,

significant interval level differences do exist in YR13 for the 1st and 2nd intervals ($F(1, 416) = 5.67$, $\eta^2 = 0.01$; $F(1, 310) = 6.94$, $\eta^2 = 0.02$).

Finally, with respect to Part (d) of RQ1, the goal was to compare changes over time in WLMF for the bottom, middle and top 25% of WLMF scores across locations. Mixed-model ANOVAs were conducted for YR13 and YR14 to examine similarities and differences between these three *ad hoc* segments of the sample. Thus the first group represented the lowest 25% (0-25%) of performers, the second group represented the middle 25% of performers (from 37.5% to 62.5%) and the third group represented the highest 25% of performers (75% to 100%). Due to the manner in which the dataset was built these groups do not include mutually-exclusive sets of interviewers as movement between groups across intervals is possible.

Results indicated that, as expected, means for these ad hoc groups were significantly different, which suggested that sufficient variability in WLMF scores existed in the sample. More importantly, results revealed that WLMF values did not change similarly across the time interval studied within these three groups. The significant main effect for WLMF over time documented earlier in this section appears to be driven primarily by those scoring in the top 25% of WLMF values. There was no main effect of time on WLMF for the bottom 25% of performance records.

Table 6

YR14 Performance Group Repeated ANOVAs

	SS	*df*	*MS*	*F*	*p*	η^2
G1 (Bottom 25%)	1.384	1	1.384	2.881	n.s	n.s
Error	34.578	72	0.48			
G2 (Middle 25%)	6.306	1	6.306	12.566	< .001	0.149
Error	36.129	72	0.502			
G3 (Highest 25%)	15.187	1	15.187	31.761	< .001	0.309
Error	33.95	71	0.478			

Table 7

YR13 Performance Group Repeated ANOVAs

	SS	*df*	*MS*	*F*	*p*	η^2
G1 (Bottom 25%)	0.123	1	0.123	0.232	n.s	n.s
Error	16.909	32	0.528			
G2 (Middle 25%)	0.342	1	0.342	0.895	n.s	n.s
Error	12.626	33	0.383			
G3 (Highest 25%)	3.304	1	3.304	5.344	0.027	0.143
Error	19.788	32	0.618			

Research Question 2

RQ2 - Data Screening Considerations

The second research question centered on a system of subjective performance ratings that the company calls "Quality Control". Employees designated as Quality Control Specialists (QCS) provide ratings of recordings generated by quasi-randomly selected interviewers, a practice that is becoming standard in many call-based industries. To qualify to be a QCS, the employee must have served as an interviewer for no less than 30 days; most QCSs meet this criterion quite easily. Approximately 10% of all daily interviews are selected for review.

However, as explained previously in this paper, the process is more constrained and complex than it might appear on the surface. First, it is very unlikely that a specific interviewer will receive feedback on his performance on consecutive days, due to manpower limitations it is not possible to review every survey. Additionally, the selection process is "quasi-random" because interviewers who receive poorer ratings are more likely to be selected for review later. This leads to two important points: first, there are ultimately fewer data points from which to draw, and second, it is impossible to use the individual interviewer as the unit of analysis. Thus, YR13 and YR14 data were combined for the current research question, and the unit of analysis was shifted to the raters (QCSs), which is more consistent with the language used in the research questions.

The master quality dataset for the review period of January 1, 2013 to October 31, 2014 contained 41,737 unique records for 1,929 unique interviewers across 525 unique calendar dates. Forty-nine unique QCSs contributed ratings to the database; however, 22 of them were removed from the database before analysis. Two of those removed were

managers, due to concerns about variation in the frames of reference that may have been used to create the ratings. The rest were removed because of a brief tenure (< 25 workdays) and/or a limited number of reviews (< 200). While these cutoffs were admittedly arbitrary, the goal was to reduce random variability in the ratings by emphasizing raters with considerable experience and practice at the task. As a result of this screening, the final database contained 40,184 records (a reduction of 3.7%).

Quality Rating Scale Metrics

Ratings are assigned independently for the six ratings for a given interview. For all six rating dimensions, the company defines "meeting expectations" as a score of 4 (on a 5-point scale). A score of 5, while possible, is rare because of the remarkably difficult standard associated with it, and as a result of unintended consequences of organizational policy. Any score on any dimension less than 4 is considered below standard and considered an "exception". There are three possible ratings in this broad class of exceptions. First, a rating of "3" indicates an infrequent and minor departure from expectations that does not significantly impact data validity or the organization's reputation. Second, a rating of "2" indicates either an infrequent but significant departure or frequent but insignificant departures from expectation that *could* affect data validity but usually does not require the deletion of the survey. Both "2" and "3" ratings trigger interventions by a floor supervisor immediately following completion of the offending interview. Finally, a rating of "1" indicates serious deficiencies that require immediate and direct action by the call center manager and deletion of the survey data. Table 8 shows the frequency of each rating within dimension.

Table 8

Ratings Frequency by Quality Measure

	Unacceptable	*Needs Improvement*	*Minor Issues*	*Meets Expectations*	*Superior*
Validity					
Verbatim	2,371	1,290	3,555	34,502	19
Assuming	1,230	4,142	643	35,707	15
Leading	407	1,813	1,737	37,774	6
Company Image					
Pace	25	36	193	41,464	19
Pronunciation	97	312	2,116	39,200	12
Professionalism	246	609	929	39,895	58

Notes: N = 41,737

RQ2 - Examining Ratings between Locations

The second part of this research question was focused on how ratings differed depending on the physical location of the QCS. Examination of the database within locations revealed that Location NW contained a disproportionate number of records that were generated by inexperienced raters. Thus, it was necessary to use the screened database as described above for this part of the question as well.

One-way analysis of variance was used to analyze differences in each rating dimension between locations. Significant main effects were observed for all analyses ($p < .05$), so post hoc contrasts were subsequently conducted. Bolded values in Table 9 indicate paired means that were different (Tukey's LSD, $p < .05$). For all six dependent measures there was as significant main effect of location (for Assuming, $\eta^2=.026$; for Leading, $\eta^2=.035$; for Verbatim, $\eta^2=.022$; for Pace, $\eta^2=.003$; for Pronunciation, $\eta^2=.003$; for Professionalism, $\eta^2=.004$).

Table 9

Quality Metrics by Locations

	Location S1		*Location S2*		*Location NW*		η^2
	M	*SD*	*M*	*SD*	*M*	*SD*	
Verbatim	**3.54**	0.96	**3.74**	0.69	**3.88**	0.46	.022
Assuming	**3.54**	0.92	**3.76**	0.69	**3.91**	0.41	.026
Leading	**3.75**	0.66	**3.88**	0.46	**3.96**	0.26	.035
Pace	3.99	0.11	4.00	0.08	**3.98**	0.20	.003
Pronunciation	**3.95**	0.27	3.91	0.34	3.92	0.34	.003
Professionalism	3.94	0.36	3.95	0.31	**3.87**	0.47	.004

Note: Bold values indicate significant differences between mean pairs, (Tukey LSD, $p < .05$).

RQ2 - Examining Individual QCS Differences in Ratings

The third part of this research question was focused on individual raters (QCSs) across the six rated dimensions. Again, the screened database as described above was used for this analysis, and no further screening was necessary. Tables 10, 11 and 12 show descriptive statistics for each dimension's ratings broken down by location. While there seems to be considerable consistency among raters, it was noted that QCS #104 and QCS #109 were much more critical than any other QCS at any other location. It is likely that these two raters may be responsible for much of the rating variation in these dimensions.

Table 10

Location S1 QCS Mean Ratings By Measure

QCS		*VB*	*ASM*	*LD*	*PC*	*PRN*	*PFL*	*N*
100	*M*	3.65	3.57	3.66	3.99	3.89	3.78	443
	SD	0.82	0.85	0.78	0.16	0.36	0.66	
101	*M*	3.61	3.69	3.75	3.99	3.94	3.94	3460
	SD	0.9	0.78	0.66	0.1	0.27	0.34	
102	*M*	3.82	3.88	3.98	4.00	3.96	3.97	824
	SD	0.57	0.46	0.17	0.1	0.21	0.21	
103	*M*	3.72	3.62	3.81	3.99	3.93	3.92	639
	SD	0.72	0.84	0.54	0.13	0.28	0.42	
104	*M*	3.06	3.26	3.63	3.99	3.97	3.92	1851
	SD	1.25	1.09	0.77	0.14	0.19	0.43	
105	*M*	3.46	3.5	3.65	3.99	3.91	3.96	940
	SD	1.00	0.95	0.81	0.15	0.34	0.35	
106	*M*	3.46	3.2	3.78	4.00	3.98	3.96	1099
	SD	0.99	1.09	0.58	0.07	0.18	0.33	
107	*M*	3.54	3.52	3.71	4.00	3.96	3.95	3080
	SD	0.92	0.91	0.66	0.04	0.22	0.29	
108	*M*	3.79	3.64	3.95	4.00	3.99	3.97	2907
	SD	0.74	0.86	0.33	0.10	0.15	0.32	
109	*M*	2.96	3.01	3.12	3.99	3.66	3.89	552
	SD	1.24	1.15	1.08	0.10	0.65	0.49	

Note: VB = Verbatim; ASM =Assuming; LD = Leading; PC = Pace; PRN = Pronunciation; PFL = Professionalism

Table 11

Location S2 QCS Mean Ratings By Measure

QCS		*VB*	*ASM*	*LD*	*PC*	*PRN*	*PFL*	*N*
200	*M*	3.73	3.69	3.92	4.00	3.92	3.93	6042
	SD	0.62	0.76	0.30	0.07	0.29	0.29	
201	*M*	3.85	3.91	3.94	4.00	3.89	3.94	3977
	SD	0.53	0.43	0.36	0.11	0.36	0.34	
202	*M*	3.68	3.74	3.94	4.00	3.85	3.93	1,111
	SD	0.88	0.75	0.34	0.10	0.51	0.41	
203	*M*	3.82	3.72	3.89	4.00	3.99	3.99	1,554
	SD	0.63	0.73	0.45	0.00	0.13	0.13	
204	*M*	3.75	3.78	3.84	3.99	3.96	3.94	1,710
	SD	0.68	0.69	0.54	0.10	0.21	0.38	
205	*M*	3.82	3.70	3.85	3.99	3.94	3.94	978
	SD	0.60	0.78	0.55	0.11	0.33	0.37	
206	*M*	3.43	3.42	3.77	4.00	3.85	3.91	514
	SD	0.87	0.94	0.51	0.06	0.40	0.40	
207	*M*	3.51	3.87	3.78	3.99	3.62	3.95	498
	SD	0.97	0.54	0.63	0.08	0.67	0.30	
208	*M*	3.55	3.76	3.66	4.00	3.93	3.98	2,032
	SD	0.92	0.68	0.75	0.06	0.30	0.21	

Note: VB = Verbatim; ASM =Assuming; LD = Leading; PC = Pace; PRN = Pronunciation; PFL = Professionalism

Table 12

Location NW Quality Measures by QCS

QCS		*VB*	*ASM*	*LD*	*PC*	*PRN*	*PFL*	*N*
300	*M*	3.89	3.97	4.00	3.99	3.99	3.88	283
	SD	0.41	0.19	0.00	0.08	0.10	0.35	
301	*M*	3.93	3.93	3.93	3.97	3.95	3.90	569
	SD	0.33	0.38	0.34	0.24	0.29	0.41	
302	*M*	3.90	3.90	3.96	3.99	3.98	3.96	341
	SD	0.46	0.45	0.23	0.18	0.16	0.29	
303	*M*	3.95	3.97	3.98	3.99	3.97	3.96	1,914
	SD	0.33	0.30	0.22	0.12	0.24	0.29	
304	*M*	3.75	3.76	3.94	4.00	3.87	3.65	433
	SD	0.56	0.65	0.31	0.17	0.39	0.81	
305	*M*	3.95	3.93	3.99	3.99	3.89	3.92	1,149
	SD	0.30	0.34	0.11	0.13	0.39	0.36	
306	*M*	3.91	3.96	3.98	3.99	3.94	3.75	382
	SD	0.38	0.27	0.14	0.07	0.33	0.63	
307	*M*	3.72	3.91	3.93	3.91	3.80	3.76	902
	SD	0.64	0.38	0.30	0.36	0.51	0.59	

Note: VB = Verbatim; ASM =Assuming; LD = Leading; PC = Pace; PRN = Pronunciation; PFL = Professionalism

RQ2 - Examining Dimension Ratings Over Time

It was important to examine rating data longitudinally as well, focusing on how the number of observed exceptions changes over the first 30 workdays. However, this analysis required that the database be rebuilt and re-screened. The first step in data preparation was to re-insert excluded raters so that all 49 QCSs were considered. Secondly, since only records within the 30-workday time frame were targeted, the database was reduced to 20,971 records (1,816 unique interviewers, 516 unique calendar dates). If an interviewer received more than one review on a given workday, those ratings were averaged for each of the six quality dimensions.

The unit of analysis for this study remained the interviewer since calculations were derived with interviewer daily averages rather than raw data per record. Review by interval was utilized rather than a daily review for several reasons. First, since the reviews represent a small percentage (about 10%) of the overall surveys completed by the organization, aggregation across intervals was made more sense. This approach provided a more robust measure of performance variance. A daily review would not have been practical since interviewers are evaluated on quality too sporadically. Second, proprietary and location-based policy differences in observation protocols for the first two workdays would result in artificial bias for those days, which can be reduced by reviewing the first six workdays as a group.

Tables 13 and 14 show that exceptions generally decrease across time within all rating dimensions while showing signs of stabilizing near the end of the 30-workday period. Exceptions also appeared to be more common for the three data validity dimensions (*verbatim, assuming, leading*) than for the three dimensions associated with organizational reputation (*pace, pronunciation, professionalism*).

Table 13

Exception Analysis by Interval by Measure

		Day 1-6	Day 7-12	Day 13-18	Day 19-24	Day 25-30	Overall
	Total Ratings	5667	3795	3027	2396	1866	16751
Verbatim	Meets	3879	2856	2362	1932	1494	12523
		68.4%	75.3%	78.0%	80.6%	80.1%	74.8%
	Exception	1788	939	665	464	372	4228
		31.60%	24.70%	22.00%	19.40%	19.90%	25.20%
Assuming	Meets	4216	2980	2422	2010	1563	13191
		74.4%	78.5%	80.0%	83.9%	83.8%	78.7%
	Exception	1451	815	605	386	303	3560
		25.60%	21.50%	20.00%	16.10%	16.20%	21.3%
Leading	Meets	4731	3302	2672	2170	1664	14539
		83.5%	87.0%	88.3%	90.6%	89.2%	86.8%
	Exception	936	493	355	226	202	2212
		16.50%	13.0%	11.70%	9.40%	10.80%	13.20%
Pace	Meets	5587	3761	3003	2378	1858	16587
		98.6%	99.1%	99.2%	99.2%	99.6%	99.0%
	Exception	80	34	24	18	8	164
		1.4%	.90%	.80%	.80%	.40%	1.0%
Pronunciation	Meets	4924	3403	2761	2224	1751	15063
		86.9%	89.7%	91.2%	92.8%	93.8%	89.9%
	Exception	743	392	266	172	115	1688
		13.1%	10.3%	8.8%	7.2%	6.2%	10.1%
Professionalism	Meets	5244	3515	2832	2271	1755	15617
		92.5%	92.6%	93.6%	94.8%	94.1%	93.2%
	Exception	423	280	195	125	111	1134
		7.5%	7.4%	6.4%	5.2%	5.9%	6.8%

Note: Meets = Ratings of 4 or greater. Exception = Ratings of less than 4.

Table 14

Exception Percent Change Between Intervals

		Int 1 to 2	*Int 2 to 3*	*Int 3 to 4*	*Int 4 to 5*
Validity					
	Verbatim	-21.84%	-10.93%	-11.82%	2.51%
	Assuming	-16.02%	-6.98%	-19.50%	0.62%
	Leading	-21.21%	-10.00%	-19.66%	12.96%
Company Image					
	Pace	-35.71%	-11.11%	0.00%	-100.00%
	Pronunciation	-21.37%	-14.56%	-18.18%	-16.13%
	Professionalism	-1.33%	-13.51%	-18.75%	11.86%

Chi-square analysis revealed that the differences in ratings across the five time intervals were significant for assuming (χ^2 (1,4) = 133.04, $p < .05$); leading (χ^2 (1,4) = 199.11, $p < .05$); verbatim leading (χ^2 (1,4) = 208.92, $p < .05$), pace (χ^2 (1,4) = 19.41, $p < .05$), pronunciation (χ^2 (1,4) = 117.15, $p < .05$), and professionalism (χ^2 (1,4) = 18.218, $p < .05$). The magnitude of change for the three image measures may appear dramatic, but in terms or raw values are much lower than those for the validity measures since the exception percentage values are significantly higher for validity than image.

RQ2 - Global Changes in Ratings Over Time

The final part of this research question was focused on the changes in ratings over a period of time (the 30-workday period). Due to the nature of the question, the unit of analysis needed to be individual records rather than individual raters. Table 15 shows means and variation for each dimension within the five 6-day intervals used in previous analyses and Figure 2 displays these means graphically. Sample size decreases across intervals due to the availability of interviewers to review as well as the tendency to place greater emphasis on less experienced employees. Generally, the so-called "validity" dimensions (verbatim, assuming and leading)

Table 15

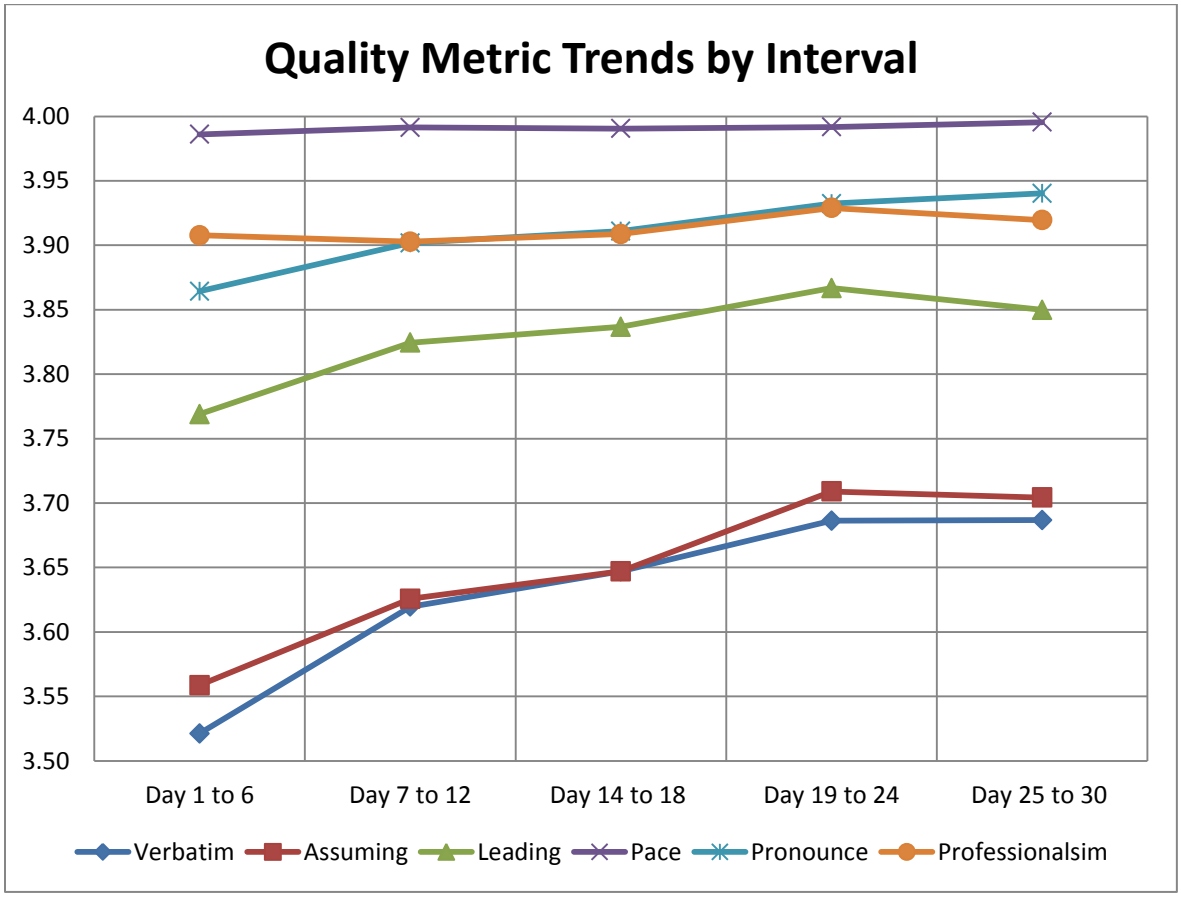
Organizational Trends for Quality Metrics by Intervals

Interval		*VB*	*ASM*	*LD*	*PC*	*PRN*	*PFL*	*N*
Day 1 to 6	*M*	3.52	3.56	3.77	3.99	3.86	3.91	
	SD	0.92	0.9	0.64	0.15	0.43	0.41	7323
Day 7 to 12	*M*	3.62	3.63	3.82	3.99	3.90	3.90	
	SD	0.84	0.84	0.56	0.12	0.35	0.43	4764
Day 13 to 18	*M*	3.65	3.65	3.84	3.99	3.91	3.91	
	SD	0.82	0.82	0.53	0.14	0.34	0.42	3717
Day 19 to 24	*M*	3.69	3.71	3.87	3.99	3.93	3.93	
	SD	0.79	0.77	0.49	0.13	0.29	0.38	2942
Day 25 to 30	*M*	3.69	3.70	3.85	4.00	3.94	3.92	
	SD	0.78	0.77	0.51	0.09	0.27	0.39	2225

Note: VB = Verbatim; ASM =Assuming; LD = Leading; PC = Pace; PRN = Pronunciation; PFL = Professionalism

Figure 2.

Interval-Based Trends in Subjective Performance Ratings Across A 30-Workday Period



increased over time, whereas the “company image” dimensions were rated consistently across the time frame.

Discussion

The study reported here was challenging because there were two sets of goals operating simultaneously. First, the preceding study was commissioned by HostCo in hopes of gaining a deeper understanding of the performance dynamics at work within its employees. But second, the study was also constructed as an opportunity to apply known psychological theory to a real-world data set and examine its validity in that light. As a result, this section must not only relate the analyses to extant psychological theory, but must also serve as a summary to the host organization. The section will be organized by research question and, where possible, elaborations on theoretical implications will be separated from the consultative summaries associated with the project.

Employee Performance Over Time

The data suggest that, despite this lack of task standardization and the uncertainty surrounding who will ultimately answer the phone, employees in this company generally show the expected trend based in psychological theory. Over the months studied, we see a clear power law function that is typical of skill development in complex tasks (e.g., Neves & Anderson, 1981). It is important to understand that the power function is not dependent on the specific metric used. As explained earlier, the nature of these data demanded that a very unique and aggregated performance index was utilized (CPH). Nevertheless, average performance across the 30-workday period was as expected – slow but steady improvements leading ultimately to a "performance plateau." This conclusion was statistically supported by breaking the time series into five 6-day intervals and examining the slope of the regression line within each interval to find the point at which the slope in non-significant. Interestingly, however, when examining the trends after splitting the sample into three "performance tiers," it was observed that the poorest

performers showed no significant improvement in performance over the workday period, whereas the best performers did. This is most likely due to an ability ceiling effect. The motivation literature is built on the premise that performance is a combination of motivation, knowledge, and competence (Matthews, Davis, Westerman, Stammers, 2000; Minbaeva, 2013). If one of these components is missing, then performance tends to degrade quickly. It is possible that the poorest performers were just unable to improve because they lacked one or more of these performance components.

However, it should be noted that for one of the time series (YR13), there was considerably more variation at the end of the time period than would be expected based on theoretical predictions. Skilled performance theory (i.e., Debarnot, Sperduti, Rienzo, & Guillot, 2014; Ericsson & Charness, 1994; Ericsson, Krampe & Tesch-Römer, 1993; Krampe & Ericsson, 1996) suggests that the variation in performance as practice time accrues should decrease, leading to not only better mean performance but also smaller standard deviations around that mean performance. Standard deviations within days do not vary considerably, although they do decrease slightly over time for YR13, but they do not for YR14.

Attempts were made to seek explanations by contacting the company that generated the data (Hostco, personal communication). For example, an unusually steep downward spike occurred on workday 28 (YR13). In cooperation with the company, an examination of the underlying dates and projects making up this workday was conducted, but no obvious reason stood out as a potential confound. However, hypotheses were generated from these conversations. First, for internal reasons, employees in YR14 were not evaluated as frequently. The literature on performance feedback suggests that supervisors can over-evaluate employees, especially in situations that are not routine and static, which could lead to confusion rather than

clarification of duties and protocols (Kim & Hamner, 1976; Latham & Wexley, 1981). Also, the maximal/typical performance literature (Deadrick & Gardner, 2008; Klehe & Anderson, 2007; Sackett, 2007) suggests that performance can lag immediately following a performance review. It is possible that the variation in YR13 is a reflection of these effects. A second hypothesis was centered on manager experience. Recall that YR13 contains data for two call centers, whereas YR14 contains data from three centers. In YR14, the center managers for Location S1 trained the managers in locations S2 and NW, but the company reports that the training was not as long for the Location NW manager, largely because this person became a manager midway through the year. Thus, this manager possessed considerably less experience than those at the other locations, who each had at least two years of management experience. Unfortunately, it would be more likely based on this logic that YR14 data (in contrast with YR13) would be more variable, so this hypothesis is ultimately somewhat unsatisfying. Finally, it was noted that the number of interviewers contributing to the CPH performance measure in YR13 was significantly smaller than in YR14. This difference could contribute to the increased performance variation as a statistical artifact, which would mean that no further explanation could be uncovered. Nevertheless, additional studies in future years would help to clarify the result.

Another set of hypotheses regarding the variation in performance is based on the possibility that either external forces (external to the employee but within the company) created a perturbation in the system that disrupted performance, or there are other employee-level factors that could have been activated during that year but were either not measured or not analyzable in the data available. For instance, it is probable that other skills are important in predicting performance as well. An example might be the social-cognitive phenomenon known as *social influence* (e.g., Kelman, 1961). Interviewers must in the end convince a respondent to participate

through an entire survey. This challenge changes in difficulty based on the survey length, a factor captured by the way WMLF index was calculated in this study, but other variables could make this more or less difficult as well. Despite just the prevailing cultural attitude about telephone surveys, variables like survey topic, respondent age, time of day, and competing respondent interests (e.g. Monday Night Football, dinner) could make the incentives associated with participation less valuable. Also, the employee does not know how many other interviewers have dialed the number recently, which could introduce respondent fatigue. Employees must also seek cooperation from abusive and threatening respondents, while refraining from retaliation and navigating uncooperative respondents to acceptable answers. Emotional labor theory (Hülsheger & Schewe, 2011; Hülsheger, Alberts, Feinhold, & Lang, 2013; Schreurs, Guenter, Hülsheger, & van Emmerik, 2014) would predict that this probably contributes to considerable fatigue, which may in turn disturb performance and elevate variation around the mean. Interestingly, social influence success has also been linked to individual differences in personality (Grant, 2013), which is not measured by HostCo because of the labor market factors discussed in the introduction.

Another key issue that could have affected performance is the distinction between landlines and cellular lines. As the cellular telephone slowly gains favor in society over land-based lines, more and more interviewers will be dialing cellular numbers. This increases the possible environments surrounding the respondent at the time that the phone call is answered, most likely making it harder to persuade the respondent to stay on the line. Any further studies of data within this organization and/or industry should be constructed *a priori* to focus on the impact of the cellular telephone's popularity on interviewer performance.

Regarding the analyses of performance between locations, the lack of difference among the three sites is telling, albeit unexpected. As previously mentioned, at issue in these data is an apparent conflict between the regimented and standardized nature of the work and a set of important differences between the labor markets to which the locations had access and clear differences in management styles as reported by the host organization. On the surface, these appeared to be competing forces: standardization implying a lack of difference between sites, whereas leadership and labor market differences implying the opposite. The data clearly support the notion that the culture of standardization was stronger than demographic and management variations. The literature on organizational culture suggests that, in some cases, organizations can construct a set of cultural norms and expectations that are very salient and palpable, even though they may not be explicitly communicated to the employees (Hofstetter & Harpaz, 2015; Jex & Britt, 2008). Often, the rigidity of the culture will be implied through the use of extensive rules, prescriptions/proscriptions, and punishments, and the way in which the host organization described their basic work processes and structures certainly suggested that a strong culture was in place.

With respect to consultative recommendations, the organization in question should be pleased that evidence suggests its employees are becoming skilled at their jobs in a relatively short period of time. Because expert performance typically requires several years of practice (Ericsson, Krampe, & Tesch-Römer, 1993), employees could still improve slightly from the levels shown in these data *if* the organization can retain them. The telephone interview industry is notoriously one of high turnover, and so it is encouraging that asymptotic performance appears to set in within the first 30 days at work (given that the employee stays that long).

A Discussion of Subjective Performance Evaluation Data

Research Question 2 was focused on the way in which the organization's subjective rating system was functioning. Specifically, the organization was interested in finding any systemic variance that was unrelated to employee performance, and whether there was evidence that the system needed revision. Six dimensions of performance were rated: three dimensions (verbatim, assuming and leading) were considered to be reflective of how employees followed proper survey procedure and ensured data validity, whereas the other three dimensions (pronunciation, pace and professionalism) were more concerned with how the employee portrayed the organization through his/her behavior.

The data showed that ratings for all dimensions were tightly grouped around a score of "4" on the 5-point scale that was utilized. While scores of 3 or less were observed, primarily for the three validity-based rating dimensions, such ratings were relatively uncommon. Of note was the finding that these lower scores were almost entirely produced by a small subset of raters. The data also showed that individual raters were quite consistent over time, showing little variability within raters. It should be made clear at this point that inter-rater reliability was not at issue here given that only one rater was assigned to each behavioral sample.

When looking at trends for exceptions (values other than "4") across the time intervals they appear fairly similar to those observed in RQ1 for objective measures. In almost all cases a sharp decrease occurred between the first and second interval followed by less dramatic decreases. Exceptions on Verbatim, Assuming, Leading and Professionalism dimensions slightly but non-significantly increased within the final interval. This may suggest that skill acquisition is in progress and that it becomes stable around the fifth interval. The slight increase in the final interval may occur due to rise in comfort level which could reduce vigilance (Espedal, 2006).

Overcoming resistance is an essential element of the interviewer position. As time on task increases, interviewers able to endure respondent resistance will remain on the job and will be better able to appropriately and strategically address problematic behaviors. Thus, it is not surprising that average ratings are higher for the more concrete measures of Verbatim, Assuming and Leading since these dimensions are connected to direct modifications of respondent behaviors.

Consultative Recommendations Regarding HostCo's Business

Recommendation 1: Perform more frequent evaluations of employee performance during odd years to alleviate unnecessary volatility.

Unlike YR14, unexplained CPH volatility was observed over the final 7 workdays in YR13. Due to the practice of continuous hiring, this effect may not be evident on daily reports generated by the E-Manager software as those reports are based on calendar days rather than workdays as defined in these analyses. One key difference between the two years during this time frame could have been the average survey length, higher in YR13 than YR14 due to the seasonal nature of the business and sensitivity to election cycles. Thus, this effect may only occur in odd years. The variation also coincides roughly with the first performance review for new employees. It is possible that managers may be allocating so much attention to newer employees at this time that more veteran employees may "relax", becoming less vigilant metacognitively. Employees who already possess an acceptable review may lower their attention to the task because of this.

Recommendation 2: Consider the use of a realistic job preview as part of the initial hiring process.

Splitting the sample into "performance levels" showed that the lower 25% of performers did not improve in either year across the time period analyzed. In contrast, the top 25% of performers showed significant performance improvement in both years. HostCo should consider reasons why the poorer performers seem to stagnate at low levels. One factor could be that many potential employees are not keenly aware of the challenges and obstacles that the interviewer must overcome. Research shows that a realistic job preview (RJP: e.g., Phillips, 1998) can enhance perceptions of the job, improve performance and decrease voluntary and involuntary turnover. Right now, HostCo does not use any screening tools that provide an ecologically valid view of operations. For example, a video could be shown with an actor reading a series of sample survey screens and handling difficult respondents would be useful. Similar videos could be produced and used in training to further immerse the new employees into the job and inoculate them against the stressors that they are likely to face. This idea is further explored in Recommendation 3 below as it relates to quality control specialists (QCS).

Recommendation 3: Establish a web-based self-guided training program for Quality Control Specialists. This training should be required quarterly.

The variance seen between locations and across raters in RQ2 is most likely caused by insufficient mental models, as the original frame of reference (FOR) training in not emphasized at this time. It is unclear if rating variance is due to competency, vigilance, leniency or other rater-based factors as a result. The effectiveness of FOR training is strongly supported in the literature over a long period of time (Gorman & Rentsch, 2009; Melchers, Lienhardt, Von Aarburg, & Kleinmann, 2011). Several software tools exist to facilitate this kind of training (i.e.,

Lectora), which should make implementation relatively easy. If HostCo is intent on using the quality metrics to guide their evaluations of employees, then more care must be taken to ensure that the data generated by these activities is maximally valid.

Recommendation 4: Implement a system of positive reinforcements to support good employee behavior and de-emphasize the reliance on looming punishment as a behavioral modifier.

While a systematic notification system exists for quality measures, no such system exists for objective measures such as CPH. The psychological literature is remarkably clear regarding the substantial benefits of reinforcement for behavioral modification, but it is also quite clear that punishment is not as effective in real-world environments (e.g., Mitchell, 2007; Pate, 1978; Vlaev & Dolan, 2015). Punishment requires immediate and consistent application, as well as clear contingencies between the undesired behavior and the punishing stimulus, and these criteria are impossible to meet reliably within the organizational context. Therefore, it is recommended that a systematic system of positive reinforcement be implemented that is triggered when employees reach goals set on objective performance criteria. For example, employees that surpass a daily CPH goal could receive verbal praise, a "break ticket", or other rewards as incentives to reach the goal. Employees who are able to achieve these goals on a regular basis could be eligible for more substantial incentives, such as gift certificates or other tangible rewards. It is important that this reinforcement system be designed carefully to control the costs to the company, however. Finally, discontinue the use of posted statistics, as it is just as likely to create unhealthy competition among employees as it is to motivate them.

References

Bartelt, V.L., & Dennis, A.R. (2014). Nature and nurture: The impact of automaticity and structure of communication of virtual team behavior and performance. *MIS Quarterly, 38*, 521-A4.

Baumeister, R.F & Bushman, B.J. (2013). Social Psychology and Human Nature. Belmont, CA: Wadsworth.

Blake, R.R., & Mouton, J.S. (1980). The Board Grid: An interview with Blake and Mouton. *Directors & Boards, 5*, 19-27.

Blake, R.R. (1970). Fifth achievement. *The Journal of Applied Behavioral Science, 6*, 413-426.

Burns, J.S. (1996). Defining leadership: Can we see the forest for the trees? *Journal of Leadership Studies, 3*, 148-157.

Consiglio, C., Alessandri, G., Borgogni, L., & Piccolo, R.F. (2013). Framing work competencies through personality traits: The Big Five competencies grid. *European Journal Of Psychological Assessment, 29*, 162-170.

Dalal, R.S., Bhave, D.P., & Fiset, J. (2014). Within-person variability in job performance: A theoretical review and research agenda. *Journal of Management, 40*, 1396-1436.

Deadrick, D.L., & Gardner, D.G. (2008). Maximal and typical measures of job performance: An analysis of performance variability over time. *Human Resource Management Review, 18*, 133-145.

Dean, A.M., & Rainnie, A. (2009). Frontline employees' views on organizational factors that affect the delivery of service quality in call centers. *Journal of Services Marketing, 23*, 326-337.

Debarnot, U., Sperduti, M., Rienzo, F.D., & Guillot, A. (2014). Experts' bodies, experts' minds: How physical and mental training shape the brain. *Frontiers in Human Neuroscience, 8,* 280. doi:10.3389/fnhum.2014.00280.

Ericsson, K.A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49,* 725-747.

Ericsson, K.A., Krampe, R., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363-406.

Ericsson, K.A., & Williams, A.M. (2007). Capturing naturally occurring superior performance in the laboratory: Translational research on expert performance. *Journal of Experimental Psychology: Applied, 13,* 115-123.

Espedal, B. (2006). Do organizational routines change as experience changes? *Journal Of Applied Behavioral Science, 42,* 468-490.

Gorman, C.A., & Rentsch, J.R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology, 94, 1336-1344.*

Grant, A.M. (2013). Rethinking the extraverted sales ideal: The ambivert advantage. *Psychological Science, 24,* 1024-1030.

Hofstetter, H., & Harpaz, I. (2015). Declared versus actual organizational culture as indicated by an organization's performance appraisal. *The International Journal of Human Resource Management, 26,* 445-466.

Kelman, H.C. (1961). Processes of opinion change. *Public Opinion Quarterly, 25*(1), 57-78.

Hillmer, S., Hillmer, B., & McRoberts, G. (2004). The real costs of turnover: Lessons from a call center. *Human Resource Planning, 27,* 34-41.

Hülsheger, U.R., Alberts, H.J.E.M., Feinhold, A., & Lang, J.W.B. (2013). Benefits of mindfulness at work: The role of mindfulness in emotion regulation, emotional exhaustion, and job satisfaction. *Journal of Applied Psychology*, *98*, 310-325.

Hülsheger, U.R., & Schewe, A.F. (2011). On the costs and benefits of emotional labor: A meta-analysis of three decades of research. *Journal of Occupational Health Psychology*, *16*, 361-389.

Jex, S.M., & Britt, T.W. (2008) *Organizational Psychology: A scientist-practitioner approach*. Hoboken, NJ: John Wiley & Sons, Inc.

Kim, J.S., & Hamner, W.C. (1976) Effect of performance feedback and goal setting on productivity and satisfaction in an organizational setting. *Journal of Applied Psychology, 61*, 48-57.

Klehe, U., & Anderson, N. (2007). Working hard and working smart: Motivation and ability during typical and maximum performance. *Journal of Applied Psychology, 92*, 978-992.

Krampe, R.T., & Ericsson, K.A. (1996). Maintaining excellence: Deliberate practice and elite performance in young and older pianists. *Journal of Experimental Psychology: General*, *125*, 331-359.

Latham, G.P., & Wexley, K.N. (1981). *Increasing productivity through performance appraisal*. Reading, MA: Addison-Wesley.

Matthews, G., Davies, D.R., Westerman, S.J., & Stammers, R.B. (2000). *Human performance: Cognition, stress and individual differences*. Philadelphia, PA: Psychology Press.

Melchers, K.G., Lienhardt, N., Von Aarburg, M., & Kleinmann, M. (2011). Is more structure really better? A comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers' rating quality. *Personnel Psychology, 64*, 53-87.

Minbaeva, D. B. (2013). Strategic HRM in building micro-foundations of organizational knowledge-based performance. *Human Resource Management Review, 23*, 378-390.

Minjung, K., & Fishbach, A. (2014). The small-area hypothesis: Effects of progress monitoring on goal adherence. *Journal of Consumer Research*, 39 (3), S138-S154.

Mitchell, J.M. (2007). An analysis of reinforcement sensitivity theory and the five-factor model. *European Journal of Personality, 21*, 869-887.

Moradi, S., Nima, A.A., Ricciardi, M., Archer, T., Garcia, D., & Andersson Arntén, A. (2014). Exercise, character strengths, well-being, and learning climate in the prediction of performance over a 6-month period at a call center. *Frontiers in Psychology, 5*, 1-37.

Murphy, K.R., & Cleveland, J. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Nathan, B.R., & Lord, R.G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. *Journal of Applied Psychology, 68*, 102-114.

Neves, D.M. & Anderson, J.R. (1981). Knowledge compilation: Mechanisms for the automatization of cognitive skills. In J.R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 127-203). Hilldale, NJ: Lawrence Erlbaum.

O'Boyle Jr., E., & Aguinis, H. (2012). The best of the rest: Revisiting the norm of individual performance. *Personnel Psychology, 65*, 79-119.

Pate, L.E. (1978). Cognitive versus reinforcement views of intrinsic motivation. *Academy Of Management Review*, *3*, 505-514.

Peris-Ortiz, M., Willoughby, M., & Rueda-Armengot, C. (2012). Performance in franchising: the effects of different management styles. *Service Industries Journal, 32,* 2507-2525.

Phillips, J.M. (1998). Effects of realistic job previews on multiple organizational outcomes: A meta-analysis. *Academy of Management Journal, 41*, 673-690.

Rice, S., Geels, K., Hackett, H., Trafimow, D., McCarley, J.S., Schwark, J., & Hunt, G. (2012). The harder the task, the more inconsistent the performance: A PPT analysis on task difficulty. *Journal of General Psychology, 139,* 1-18.

Sackett, P.R. (2007). Revisiting the origins of the typical-maximum performance distinction. *Human Performance*, *20,* 179-185.

Schippers, M.C. (2014). Social loafing tendencies and team performance: The compensating effect of agreeableness and conscientiousness. *Academy Of Management Learning & Education, 13*, 62-81.

Schreurs, B., Guenter, H., Hülsheger, U., & van Emmerik, H. (2014). The role of punishment and reward sensitivity in the emotional labor process: A within-person perspective. *Journal Of Occupational Health Psychology, 19*, 108-121.

Sweney, A. B., Fiechtner, L.A., & Samores, R.J. (1975). An integrative factor analysis of leadership measures and theories. *Journal of Psychology, 90*, 75-85.

Srisuthisa-Ard, A. (2014). The impact of interaction with the public on work outcomes: Role of agreeableness and job complexity. *Academy Of Management Annual Meeting Proceedings*, 1540-1544.

Walsh, J., & Deery, S. (2006). Refashioning organizational boundaries: Outsourcing customer service work. *Journal Of Management Studies*, *43*, 557-582.

Winiecki, D.J. (2004). Shadowboxing with data: Production of the subject in contemporary call centre organisations. *New Technology, Work & Employment*, *19*, 78-95.

Verbeke, W., Volgering, M., & Hessels, M. (1998). Exploring the conceptual expansion within the field of organizational behavior: Organizational climate and organizational culture. *Journal of Management Studies*, *35*, 303–329.

Vlaev, I., & Dolan, P. (2015). Action change theory: A reinforcement learning perspective on behavior change. *Review Of General Psychology*, *19*(1), 69-95.

Ybarra, O., Kross, E., & Sanchez-Burks, J. (2014). The "Big Idea" that is yet to be: Toward a more motivated, contextual, and dynamic model of emotional intelligence. *Academy Of Management Perspectives*, *28*, 93-107.

Yukl, G. (2013). *Leadership in organizations* (8th Ed.). New York, NY: Pearson.

VITA

John R. Starne has worked in a variety of industries in dual roles serving as an internal consultant on human resource assessment and development issues as well as building and supporting the information management systems necessary to execute solutions. He has also served formally in the capacity of a human resources manager as well as a system administrator. Completing his bachelor's degree in psychology at Sam Houston State University in 2007 he returned to conduct master's level work in I/O Psychology at Angelo State University in 2013 where he conducted research associated with performance appraisal using eye tracking technology and frame of reference style treatments to understanding rating decision processing. Raised in Houston, Texas he is a first generation college student.