DO EXERCISES OR DIMENSIONS EXPLAIN THE VARIANCE IN OVERALL ASSESSMENT RATINGS IN A CALL CENTER BASED ASSESSMENT CENTER?

by

Brannan W. Mitchel-Slentz

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Sciences in Industrial/Organizational Psychology

Middle Tennessee State University

December 2015

Thesis Committee:

Michael Hein, Ph.D., Chair

Mark Frame, Ph.D., Committee Member

Judith Van Hein, Ph.D.

ABSTRACT

This study addressed the on-going question of whether trait-based dimensions or task-based dimensions best explain the variance in the Overall Assessment Ratings (OARs) in Assessment Centers (ACs) by using Confirmatory Factor Analysis to compare the fit of three theoretical models. Participants took part in several phone calls, in which they were interviewed and engaged in role-play scenarios to elicit behavioral data and were then rated on trait-based dimensions by a single assessor. The data collected was then used to test models in which trait-based dimensions, task-based dimensions, and one-general dimension were predicted to explain the variance in OARs. The trait-based and one-general dimension models had a poor fit, while the task-based dimension model resulted in an acceptable fit with the data. The results suggest a need to re-evaluate AC design methodology, with a focus on job relevant tasks rather than job relevant traits.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

INTRODUCTION

Assessment Centers (ACs) are tools used to observe and rate potential or current employees by presenting the participants (those being assessed) with multiple tasks, which are used to elicit behaviors indicative of specific traits. These traits are then rated by multiple trained observers (Kuncel & Sackett, 2014). Tasks used in ACs include: In-Baskets, Role Plays, Leaderless Group Discussions, and other exercises that represent work place behaviors. Traits measured in ACs are more varied but include: Leadership, Communication, Agreeableness, Innovation, and other traits determined by a job analysis. In all, Arthur, Day, McNelly, and Edens (2003) found 168 trait-based dimensions in a meta-analysis of ACs, while Rupp, Gibbons, Runnels, Anderson, and Thornton (2003) found 1,095 separate trait-based dimensions. Assessment Centers were first developed in the 1940s by British Civil service agencies, but underwent much more rigorous study in the 1950s by AT&T (Thornton & Rupp, 2006). It was during this time that many of the aforementioned basic components that define an AC were determined. Although, ACs have varied since their inception, in part because of changing complexities in work, changing workplace demographics, and new applications of the assessment center method, some basic assumptions have gone relatively unchanged. One of the underlying assumptions since the first organizational studies of Assessment Centers (Bray & Grant, 1966) has been that the traits have predictive validity in determining future job success for the participants; typically in terms of job performance or salary progression. Multiple studies, including meta-analyses, have shown strong

criterion-related validity for Assessment Centers in work place settings (Arthur, Day, McNelly, & Edens, 2003; Gaugler, Rosenthal, Thornton, & Bentson, 1987; Meriac, Hoffman, Woehr, & Fleisher, 2008).

Assessment Centers (ACs) are widely used for both administrative and developmental purposes (Chen, 2006; Thornton & Rupp, 2006) as they provide a mix of cognitive and non-cognitive tests by which a job candidate or incumbent can be rated. The function of an assessment center is to provide a comprehensive type of testing involving behavioral observations by multiple assessors across a variety of job related tasks. A traditional assessment center is designed to elicit behaviors that are indicative of specific desirable traits for job candidates. These traits are quantified by each assessor, then aggregated across multiple assessors and different simulations, either through facilitated discussion or with the aid of a regression formula, to create Overall Assessment Ratings (OARs). According to Thornton and Rupp (2006) the best ACs choose traits based on job analysis or competency modelling, and are tailored for different purposes, different positions, and the organizational culture affects the way these traits are defined. For a full review of what exercises and activities are generally considered to be a part of an AC, the reader is referred to the Guidelines and Ethical Considerations for Assessment Center Operations (International Taskforce on Assessment Center Guidelines, 2014).

Assessment Center Advantages

A large part of Assessment Centers' popularity comes from face validity. Participants regard Assessment Centers as fair because the tasks are designed to simulate

job-related situations (Thornton & Rupp, 2006). It follows that someone who performs well in the simulation would perform well at her or his job. Klimoski and Brickner (1987) also found that ACs generally have good content validity, as the tasks and skills that are typically used on the job are adequately reflected by the tasks required for the Assessment Center. In addition, a consistently high criterion-related validity has been found for OARs' ability to predict future job success (Gaugler, Rosenthal, Thornton, & Bentson, 1987; Meriac et al. 2008). Hunter and Hunter (1984) found a .43 correlation between OAR scores and job performance; and Jansen and Stoop (2001) found a corrected mean validity of .39 in predicting career advancement over a 7 year period. Finally, as a practical matter, Assessment Centers have very low adverse impact compared to some other selection tools, especially compared to only using cognitive ability tests (Petrides, Weinstein, Chou, Furnham, & Swami, 2010; Robertson, Iles, Gratton, & Sharpley, 1991; Robie, Osburn, Morris, Etchegaray, and Adams, 2000; Thornton & Rupp, 2006).

Assessment Center Disadvantages

ACs have some drawbacks, however, that have kept them from being the primary selection tool for most organizations. First, ACs are relatively expensive. The time and expertise required to develop an AC is significant, and they require training of the assessors which will require contracting or hiring psychologists or other testing administration specialists. Often the assessors will be high level employees, such as department managers who will need to be compensated for their time. Thus, Assessment Centers may not be the most efficient way to make decisions due to time and cost. While

the fact that ACs are both expensive and time consuming does make them less desirable, they have been shown to have a positive return on investment (Thornton & Rupp, 2006) both financially – between $2,500 and $21,000 per selected manager – and as a training tool for the raters. By participating in assessment center rater training, raters can learn effective behavioral observation skills, get an opportunity to interact with other managers, and can improve their ability to appraise performance. Macan, Mehner, Havill, Meriac, Roberts, and Heft (2011) found that managers trained as raters for assessment centers used more specific behavioral descriptions during performance appraisals than untrained managers. However, the potential of this positive return on investment depends on how well the AC is measuring what it is intended to measure.

Construct Validity

Construct Validity is the degree to which a test or instrument actually measures what it purports to measure (Cronbach & Meehl, 1955). In the case of ACs, the constructs are typically trait-based dimensions, which are operationally defined by theoretically indicative behaviors. According to Cronbach and Meehl it is important to develop a nomological network around a purported measure in order to establish the construct validity of that measure. A nomological network is developed by having researchers measure the relationships between different constructs that one would expect to be similar or dissimilar. If there is are two constructs that we would expect to be similar or the same, we would expect them to highly correlate – show convergent validity. If a construct should have an opposite or dissimilar meaning than the construct we are measuring, then we should expect it to not correlate or negatively correlate – show

discriminant validity. According to Cronbach and Meehl, no single study is sufficient to establish construct validity. Construct validity is established through a continuous process in which it is evaluated against new information, refined, and then reevaluated.

While ACs appear to be valid in that they are effective at predicting the success of job candidates (Arthur, Day, McNelly, & Edens, 2003; Jansen & Vinkenburg, 2006; Meriac, Hoffman, Woehr, & Fleisher, 2008), there is some debate as to how they work, and whether they are construct valid – the degree to which they are measuring what they purport to measure. The typical method of determining predictive validity is to look at the OARs and compare them with later markers for success: such as raises, promotions, and job reviews. However, the means by which the OARs are derived have come into question (Sackett & Dreher, 1982). The OARs are determined by aggregating trait-based dimension scores across different tasks. To show convergent validity, one should find that these trait-based dimensions correlate with the same trait-based dimensions across the different tasks at a higher rate than they correlate with different trait-based dimensions in the same task. In addition, to show discriminant validity, one should find that trait-based dimensions will have low correlations with unrelated trait-based dimensions. However, since the release of Sackett and Dreher's highly influential study in 1982, a large number of studies (Bowler & Woehr, 2006; Lance, Lambert, & Gewin, 2004; Petrides et al. 2010) have found the opposite to be the case. In fact, the trait-based dimensions that should have correlated with the same trait-based dimensions in other tasks were often more highly correlated with seemingly unrelated trait-based dimensions

in the same task. These results suggest a problem with the construct validity of the way ACs are traditionally developed.

Lance et al. (2004) noted that while trait-based dimension factors tended to fall apart, there was strong evidence that task-based dimensions were more useful distinct factors. Jackson, Stillman, and Atkins (2005) found similar evidence that task-based factors tended to explain a greater degree of the variance in OARs than trait-based factors. Petrides et al. (2010) attempted to use best practices to reduce potential exercise effects, but still found heterotrait-monomethod correlations (the correlations between different trait-based dimensions across a single task) to be significantly greater than heteromethod-monotrait correlations (the correlations between the same dimensions across different tasks), as well as a factor structure that showed tasks to be the drivers of variance in their Assessment Center. A growing number of studies support these findings, and suggest that the task-based dimensions are the unique constructs being measured rather than the trait-based dimensions around which the ACs are traditionally developed.

More recently Kuncel and Sackett (2014) posited that trait-based dimensions still have value, but that the focus needs to change from the intermediary post-exercise dimension ratings – ratings determined after each exercise – to Overall Dimension Ratings – ratings determined after the completion of all exercises. They found that, with an increase in the number of tasks across which the trait-based dimensions were aggregated, the degree to which task-based variance affected the OARs was seriously mitigated. The authors suggest that a general dimension factor, one which could not be attributed to specific measured trait-based dimensions, is the primary driver of variance

in Assessment Center ratings. This would minimize the effect of specific trait-based factors on the variance of the OARs, which could explain why tasks have consistently been found to explain more of the variance. A common method of testing the construct validity of ACs involves using a Multi-Trait Multi-Method matrix (Campbell, 1959) to compare correlations and establish convergent and discriminant validity (Jackson, Stillman, and Atkins, 2005; Petrides et al., 2010; Sackett & Dreher, 1982). However, this method has some issues in that the correlations being compared have varying degrees of error, making comparisons less reliable. Instead, in what some consider to be a superior method for testing the degree to which task-based dimensions and trait-based dimensions affect the variance of the OARs, many studies have begun using confirmatory factor analysis (CFA) to test different models for goodness of fit. Confirmatory Factor Analysis, developed by Jöreskog (1969), is a way of testing whether data fits with a hypothesized measurement model (Bowler & Woehr, 2006). Using CFA allows for proposing models that control for error variance.

The authors of several studies have proposed different factor models that might better explain how ACs work. Bowler and Woehr (2006) tested multiple models in a large meta-analysis and found that trait-based dimension and task-based dimension models both fit to some degree, but that task-based models explained more of the variance in AC scoring than trait-based models. A study by Meriac, Hoffman, Woehr, and Fleisher (2008) tested multiple models for best fit, including a model that proposed that trait-based dimensions had the greatest effect on variance in the OARs; one that proposed task-based dimensions had the greatest effect; one that suggested equivalent

effects; and one that aggregated the trait-based dimensions into a single factor and then combined it with discrete task-based dimensions. Other studies have tried to strengthen the trait-based effects by more tightly controlling the type of training that the raters receive (Lievens, 2002). They suggest that high levels of cognitive load might make it difficult for raters to focus on individual traits during each task, and instead cause them to conflate their ratings.

Assessment Center Development

The traditional process of developing an AC (Thornton & Rupp, 2006) first involves performing a job analysis or using competency modeling to determine job relevant trait-based dimensions. These dimensions are chosen very carefully and defined very carefully because something as imprecise as "leadership" could have different meaning in different job settings. The next part of the process is to choose tasks that elicit dimension relevant behaviors. Different tasks should be selected in order to reveal all facets of a given job-relevant trait-based dimension. These steps are logical if the trait-based dimensions are the true predictors of job success, and are followed by virtually every organization that uses ACs. However, with the growing evidence (Bowler & Woehr, 2006; Lance, Lambert, & Gewin, 2004; Petrides et al. 2010; Sackett & Dreher, 1982) that task-based dimensions, not trait-based dimensions, are the drivers of variance in future job performance, this suggests an alternative design would be superior. According to Thornton & Rupp (2006), one of the unique benefits of assessment centers are their use of simulations to gather data. The alternative design would center on the simulations in terms of relevance to the job. Instead of using a worker-oriented job

analysis or competency modelling to develop a list of job relevant traits, it would make sense to perform a task-oriented job analysis to sample the job for relevant tasks. The ability to perform the tasks most relevant to the job may be a more direct way of measuring job performance than creating a list of traits that are common among individuals capable of performing the tasks that comprise the job.

Despite the evidence that trait-based dimensions are less relevant to the design and interpretation of ACs than task-based dimensions are, why haven't there been broad changes to how ACs are designed and put into practice? Thornton and Rupp (2006) propose that the findings are not so clear as to require such changes. They suggest that assessor effects could be included in the task-related variance, that the different tasks do not actually reflect different methods, and that it is flawed to be comparing traits for each task – traits should be rated after all of the relevant behavioral data has been collected from the different tasks. Indeed, this would mean that within-task comparisons of different traits are not possible and therefore Multi-Trait Multi-Method Matrices are methodologically unsound for Assessment Center construct validation.

Regardless of whether the evidence that tasks are the true drivers of variance in performance is sufficient to change the design of ACs, they appear to have criterion, content, and face validity in their current design. Given the current state of the task-trait debate as pertains to the construct validity of assessment centers, it is important that more research on the internal structure of ACs – research that provide strong evidence one way or the other – be performed in order to create a consensus.

Our research used confirmatory factor analysis to test three competing models for best fit in evaluating the construct validity of Assessment Centers. The first model is the theoretical model on which Assessment Centers have been based and assumes that trait-based dimensions are the factors being measured. Two alternative models have been proposed as more accurate representations of what are being measured in Assessment Centers: a single-trait (one-dimension) model and a task-based model. The goal of this research is to determine whether the traditional trait-based model, a single-trait model, or a task-based model best represents the underlying factors in an Assessment Center. By comparing these three models using a large sample size, we hope to shed further light on the task-trait debate.

CHAPTER II

METHODS

Participants

Assessment Center data were collected from 756 participants applying for the position of Financial Advisor at a large financial services company. Participants performed multiple-tasks, including an interview and role plays, during phone calls that took place over several days. All identifying information for participants was purged from the data to protect anonymity and replaced with randomly generated ID numbers. Finally, any participants with missing data were removed from the data to produce a final sample of 703 participants. Data for participants used in the analysis were collected from January 2004 to January 2007.

Materials and Measures

The assessment center measured a set of company-specific competencies across a common set of tasks. These following 7 competencies were rated in the course of this AC: Demonstrates Personal Impact; Demonstrates Stress Tolerance/Adaptability; Drives Towards Success; Establishes Credibility and Builds Trust; Has High Personal Work Standards; Is Focused; and Operates with a Quality Service Mind Set. The 5 tasks used for the AC included: an Accountability Meeting and Debrief Interview; a Mini Simulation; a Phone Call with a New Prospect Role Play; and a Sales Presentation Role Play. However, only 3 competencies and 3 tasks were completed and rated for all 703 candidates, so were the only ones included in the analysis. The three competencies measured in the final analysis were: Demonstrates Personal Impact, Drives Towards

Success, and Operates with a Quality Service Mind Set. The three tasks included in the analysis were: Accountability Meeting and Debrief Interview, Phone Call with a New Prospect Role Play, and Sales Presentation Role Play. The raters scored each competency on a 3-point scale --Highly Effective, Effective, and Ineffective. The task-based dimensions used in the analysis, along with the corresponding used trait dimensions – the competencies, can be found in Appendix A.

Raters

Data for each participant were collected by an individual rater, who was located in a call center, during a series of phone interviews. The rater used a list of behavioral examples against which the applicant's behavior was compared for each "Success Factor", or competency, on which the applicant was to be rated. Based on the 3-point Likert scale for each competency – Highly Effective, Effective, or Ineffective -- the interviewer then determined an overall rating of the applicant's readiness for the position, rating her/him using a 3-point Likert scale as either "Not Ready at this Time", "Satisfactory", or "Strong."

Procedures

Confirmatory Factor Analysis. In order to test the model of best-fit for how assessment center ratings are scored, three existing factor models were selected for comparison to be analyzed using confirmatory factor analysis. The trait-based dimension model assumes that the items designed to measure specific trait-based dimensions load on the traits across each task, as shown below in Figure 1.

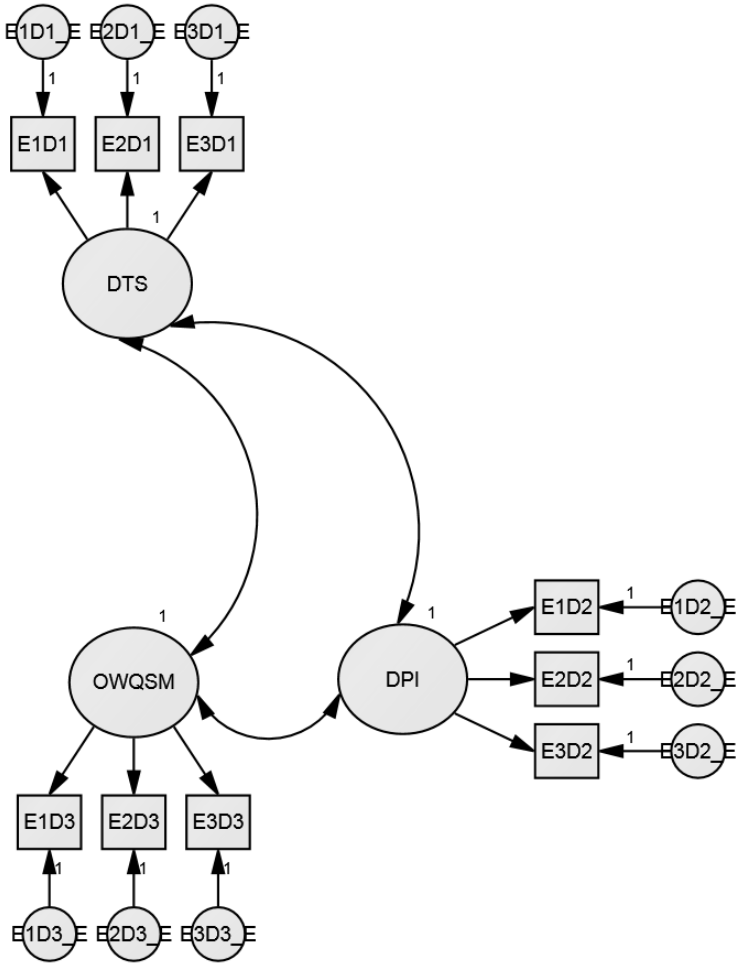

Figure 1. Trait-based Factor Model.

*DTS = Drives Towards Success; DPI = Demonstrates Personal Impact; OWQSM = Operates with a Quality Service Mindset

The single dimensional model assumes that each individual item loads directly on a single general dimension, as shown below in Figure 2.

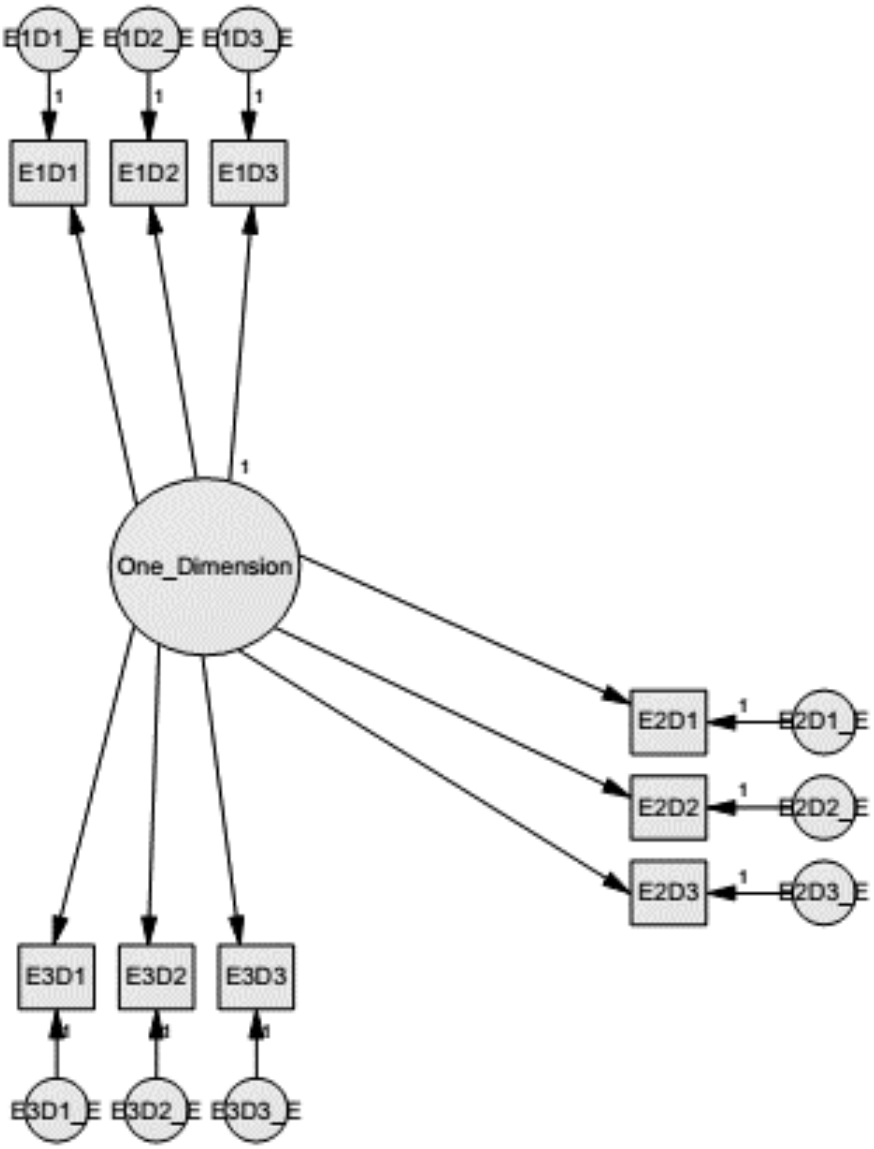

*Figure 2***.** One-dimension Factor Model

The task-based model suggests that each individual item loads on the task in which it was measured, rather than the traits across each task, as shown below in Figure 3.

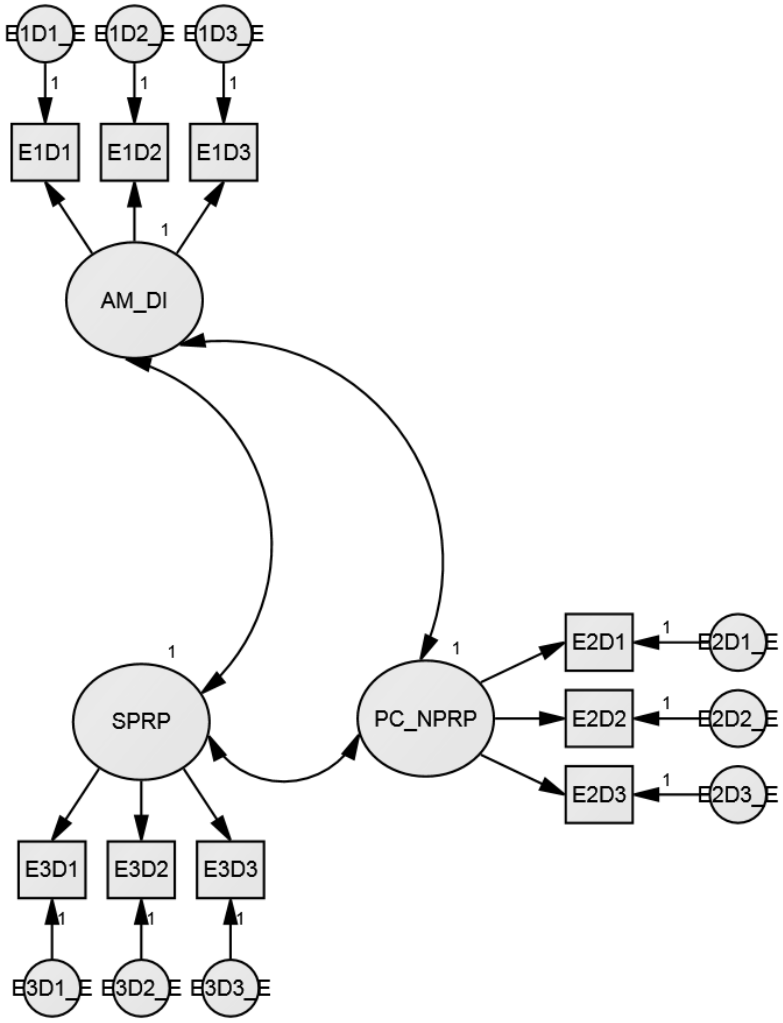

Figure 3. Task-based Factor Model

Note. AM_DI =Accountability Meeting and Debrief Interview; PC_NPRP = Phone Call with a New Prospect Role Play; SPRP = Sales Presentation Role Play

A confirmatory factor analysis was performed using the IBM Statistical Package AMOS. The following measures of fitness were tested for each model: Chi-square Goodness of Fit, RMSEA, AIC and CFI. The Chi-square was selected for parsimonious fit, the RMSEA to account for sample size, and the CFI for incremental fit over the null hypothesis. Akeike information criterion was measured, not as an absolute measure of fit, but as an additional measure to compare estimates of information loss from each candidate model.

CHAPTER III

RESULTS

When analyzing the data using the AMOS software, we fixed the variance of the factors to 1.0. Descriptive statistics are displayed below on Table 1.

Table 1.

Means and Standard Deviations for Trait-based Dimensions

Measure	*n*	*M*	*SD*
Accountability Meeting and Debrief Interview			
Drives Towards Success	703	1.87	.706
Demonstrates Personal Impact	703	1.82	.688
Operates with a Quality Service Mindset	703	2.08	.656
Phone Call with a New Prospect Role-play			
Drives Towards Success	703	1.92	.749
Demonstrates Personal Impact	703	1.91	.732
Operates with a Quality Service Mindset	703	2.11	.705
Sales Presentation Role Play			
Drives Towards Success	703	2.19	.744
Demonstrates Personal Impact	703	2.19	.728
Operates with a Quality Service Mindset	703	2.19	.664

The confirmatory factor analysis for the trait-based model returned an inadmissible solution, suggesting that the trait-based dimensions were not driving the variance of the Overall Assessment Ratings. The analysis for the one-dimension model showed a poor fit, $\chi^2(27) = 944.941$; RMSEA = .22; AIC = 980.941; CFI = .65; $p < .001$. The analysis of the task-based dimension model showed a close fit with the data, $\chi^2(24) = 95.65$; RMSEA = .065; AIC = 137.65; CFI = .973; $p < .001$. Specific results are displayed in Table 2.

Table 2

CFA Results Summary

Model	χ^2	*df*	RMSEA	AIC	CFI
Trait-based Dimension Model	941.35	24	.233	983.345	.651
Task-based Dimension Model	95.65	24	.065	137.650	.973
One Dimension Model	944.94	27	.220	980.941	.650

** p* < .05, *** p* < .001, **** p* < .0001

Note. RMSEA = root mean-square error of approximation, AIC = Akeike information criterion, CFI = comparative fit index.

This data strongly suggests that the Overall Assessment Ratings are being driven by the tasks, as opposed to the traits, for which the tasks were chosen.

CHAPTER IV

DISCUSSION

Lance (2008) suggested that dimensional analysis could be dropped entirely from the scoring of ACs. If the preponderance of evidence shows that trait-based ACs are construct invalid, then that would be logical. However, there are some other possible reasons that exercise effect models or task-based dimension models, such as the one in this study, tend to beat the original trait-based models. Meriac, Hoffman, and Woehr (2014) suggest that a lack of consistency in how trait-based dimensions are determined could be driving the lack of construct validity for trait-based assessment centers. While testing methods have specific methods and protocols that are consistent across different Assessment Centers, the dimensions that exist in two different organizations' Competency Models could have different definitions. Indeed, they found that by reducing the number of specific trait-based dimension factors -- Administrative Skills, Relational Skills, and Drive -- they found much better model fit.

Another suggestion (Collins, Schmidt, Sanchez-Ku, Thomas, McDaniel, & Le, 2003) is that the task variance is actually measuring personality traits and general cognitive ability. Collins et al. found that scores on specific tasks -- In-Basket Exercises and Leaderless Group Discussions -- were highly correlated with cognitive ability scores and the big-5 trait of extraversion, respectively. This might suggest that measures of personality and cognitive ability are driving the variance in the OARs in the guise of exercise effects. Hoffman, Kennedy, LoPilato, Monohan, and Lance (2015) similarly found convergence between commonly used tasks and measures of intelligence and

personality traits. While this may help to establish a nomological network for the task-based variables, the authors also noted that the lack of convergence with the big-5 traits of Conscientiousness and Agreeableness suggests that the task-based variables may be lacking in content validity.

Hoffman et al. (2015) recently used a meta-analysis to determine the criterion and incremental validity of using task variables, and found a significant reduction in the validity coefficient compared to trait-based AC ratings. When participants were rated by task-based dimensions rather than trait-based dimensions, the ratings were not as useful for predicting future success. The authors admitted, however, that the tasks were designed to elicit trait-based dimensional ratings and so some degree of fidelity was likely lost. Regardless, the study results indicated value in using tasks as variables, while at the same time showing the importance of trait-based dimensional ratings in ACs. If it is true that a combination of traits and tasks best explain the variance in ACs, then future AC designs could benefit from measuring an overall task-based rating, an overall trait-based rating, and then an interaction rating.

Other factors in the design of this Assessment Center could also play a role in diminishing the effect of the trait-based dimensions. Inconsistent with best AC practices (Thornton & Rupp, 2006), only one rater was assigned for each individual. With all the pressure on one rater, it is possible that the design puts too much cognitive load on the raters, such that the dimensions are conflated into a general overall performance rating (Lievens, 2002). Furthermore, raters in a call center talking on the phone may experience more fatigue than raters who are part of a more formalized Assessment Center.

The results of our study, as well as the large number of AC participants, provide strong evidence for a task-based factor structure. Bowler and Woehr (2006) suggested that a single large sample, such as in this study, is less susceptible to sampling error, redundancies, and differences in dimensions and exercises across different ACs. While the results of this study are not surprising, given the large number of studies with similar results, it does have some important implications. The largest implication is that the design of the AC from which the data was drawn is flawed, because it is designed for a trait-based factor structure. It is clear that the AC in question, and most ACs, are failing to measure what they purport to measure. Instead of sampling jobs for traits, it is appropriate to sample the tasks most important for successful job performance. ACs should then be designed to provide high fidelity simulations of tasks inherent to the job.

Additionally, while many ACs are designed for administrative or selection purposes, many others are used for training and developmental purposes. If an AC is being used to determine competencies or dimensions of an individual for purposes of training, the results may provide irrelevant or even counterproductive information. This doesn't mean ACs cannot be used for needs assessment purposes, it just suggests a different focus. That is, ACs should be used to assess strengths and areas for improvement in the kinds of tasks that are required for various positions within the organization. Furthermore, traits are less conducive to training than tasks, in that some traits may be fixed parts of personality – conscientiousness, extraversion, and agreeableness, while tasks are necessarily learned.

Finally, there is an obvious need for more studies like Lievens, Dilchert, & Ones (2009) to examine the predictive validity of task-based based ACs, rather than just critique the internal structure of current dimension based ACs. Some researchers (Hoffman et al., 2015) have made an effort to develop a taxonomy of AC tasks, which can be used as a basis for determining the predictive validity of different tasks across different ACs. Perhaps further studies can find a way to deconstruct tasks into specific behaviors that are predictive of overall performance within the task.

Limitations

There are several areas where this study could be improved. The most important is that the AC from which we collected data did not utilize multiple raters, which may make it more difficult to generalize to other ACs. Since one rater had to rate the participant on multiple dimensions during a single task, this could have created a halo effect that might explain the lack of dimension-based variance. Another limitation was our inability to account for the other dimensions on which the participants were rated. Overall, there were up to 7 dimensions than an individual rater may have had to consider for each participant, but only 3 of which that were consistently measured in our sample. Without the inclusion of those other variables, we may have missed potentially confounding variables. For example, if a rater had to consider 7 variables in a single interview, she or he may have subconsciously grouped the dimensions to ease her or his cognitive load. Finally, because the scope of the research paper was to compare three models, the models chosen were theoretically uncomplicated – task-based dimensions explain the variance, trait-based dimensions explain the variance, or a single general

dimension explains the variance. Much of the current literature suggests that some combination of dimensions and tasks would provide a more complete model, which we did not attempt to test in this research. A suggestion for future research is for researchers to work with organizations that administer and collect data on a large number of tasks.

Conclusion

The strength of the task-based based factor structure in this study fits with the landmark Sackett and Dreher (1982) study that has led researchers to reevaluate how we think about the design of Assessment Centers. While it is necessary to continue to determine the importance of tasks in ACs, it also behooves us to consider the practical implications of these findings. The results suggest that future AC designers should consider the key tasks that are best predictors of job performance, and design the AC exercises to best simulate said tasks. It may still be relevant for raters to consider specific trait-based dimensions, but the focus should be on overall performance on the task. In the future, it may also serve us better to describe someone by what they are capable of doing, rather than by using abstract dimensions of personality.

REFERENCES

Arthur W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology*, *56*(1), 125-154.

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology*, *91*(5), 1114-1124. doi:10.1037/0021-9010.91.5.1114

Campbell, D. T. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105. doi:10.1037/h0046016

Chen, H. C. (2006). Assessment center: A critical mechanism for assessing HRD effectiveness and accountability. *Advances in Developing Human Resources*, *8*, 247-264.

Collins, J. M., Schmidt, F. L., Sanchez-Ku, M., Thomas, L., McDaniel, M. A., & Le, H. (2003). Can basic individual differences shed light on the construct meaning of assessment center evaluations? *International Journal of Selection and Assessment*, *11*(1), 17-29. doi:10.1111/1468-2389.00223

Cronbach, L. J., Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52* (4): 281–302. doi:10.1037/h0040957

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, B. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, *72*, 493-511.

Hoffman, B. J., Kennedy, C. L., LoPilato, A. C., Monahan, E. L., & Lance, C. E. (2015). A review of the content, criterion-related, and construct-related validity of assessment center exercises. *Journal of Applied Psychology*, *100*(4), 1143-1168. doi:10.1037/a0038707

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*(1), 72-98. doi:10.1037/0033-2909.96.1.72

International Task Force on Assessment Center Guidelines. (2014). Guidelines and ethical considerations for assessment center operations. *Journal of Management*, *41*(4), 1244-1273.

Jackson, D. R., Stillman, J. A., & Atkins, S. G. (2005). Rating tasks versus dimensions in assessment centers: A psychometric comparison. *Human Performance*, *18*(3), 213-241. doi:10.1207/s15327043hup1803_2

Jansen, P. W., & Stoop, B. M. (2001). The dynamics of assessment center validity: Results of a 7-year study. *Journal of Applied Psychology*, *86*(4), 741-753.

Jansen, P. G., & Vinkenburg, C. J. (2006). Predicting management career success from assessment center data: A longitudinal study. *Journal of Vocational Behavior*, *68*, 253-266. doi:10.1016/j.jvb.2005.07.004

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183-202.

Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology*, *40*, 243-260.

Kuncel, N. R., & Sackett, P. R. (2014). Resolving the assessment center construct validity problem (as we know it). *Journal of Applied Psychology*, *99*(1), 38-47. doi:10.1037/a0034147

Lance, C. E., Lambert, T. A., & Gewin, A. G. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology*, *89*(2), 377-385.

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*(1), 84-97. doi:10.1111/j.1754-9434.2007.00017.x

Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. *Journal of Applied Psychology*, *87*, 675–686.

Lievens, F., Dilchert, S., & Ones, D. (2009). The importance of exercise and dimension factors in assessment centers: Simultaneous examinations of construct-related and criterion-related validity. *Human Performance*, *22*(5), 375-390.

Macan, T., Mehner, K., Havill, L., Meriac, J. P., Roberts, L., & Heft, L. (2011). Two for the price of one: Assessment center training to focus on behaviors can transfer to performance appraisals. *Human Performance*, *24*(5), 443-457.

Meriac, J. P., Hoffman, B. H., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology*, *93*, 1042-1052.

Meriac, J. P., Hoffman, B. J., & Woehr, D. J. (2014). A conceptual and empirical review of the structure of assessment center dimensions. *Journal of Management*, *40*(5). 1269-1296.

Petrides, K. V., Weinstein, Y., Chou, J., Furnham, A., & Swami, V. (2010). An investigation into assessment centre validity, fairness, and selection drivers. *Australian Journal of Psychology*, *62*(4), 227-235. doi:10.1080/00049531003667380

Robertson, I. T., Iles, P. A., & Sharpley, D. (1991). The impact of personnel selection and assessment methods on candidates. *Human Relations*, *44*, 963-982.

Robie, C., Osburn, H. G., Morris, M. A., Etchegaray, J. M., & Adams, K. A. (2000). Effects of the rating process on the construct validity of assessment center dimension evaluations. *Human Performance*, *13*(4), 355-370.

Rupp, D. E., Gibbons, A. M., Runnels, T., Anderson, L., & Thornton, G. C., III. (2003, August) *What should developmental assessment centers be assessing?* Paper presented at the 63rd annual meeting of the Academy of Management, Seattle, Washington.

Sackett, P. R., & Dreher, G. F. 1982. Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology*, *67*, 401-410.

Thornton, G. C., III, & Rupp, D. E. (2006). *Assessment centers in human resource management*. Mahwah, NJ: Lawrence Erlbaum.

APPENDICES

APPENDIX A

TASK AND TRAIT MATRIX

Tasks	**Competencies**		
	Drives Towards Success	**Demonstrates Personal Impact**	**Operates with a Quality Service Mindset**
Accountability Meeting and Debrief Interview	Exercise 1 x Dimension 1	Exercise 1 x Dimension 2	Exercise 1 x Dimension 3
Phone Call with a New Prospect Role Play	Exercise 2 x Dimension 1	Exercise 2 x Dimension 2	Exercise 2 x Dimension 3
Sales Presentation Role Play	Exercise 3 x Dimension 1	Exercise 3 x Dimension 2	Exercise 3 x Dimension 3

APPENDIX B

MTSU IRB APPROVAL LETTER

MIDDLE TENNESSEE STATE UNIVERSITY

8/27/2014

Investigator(s): Brannan Mitchel-Slentz, Dr. Michael Hein, Dr. Mark Frame
Department: Psychology
Investigator(s) Email Address: bwm3b@mtmail.mtsu.edu

Protocol Title: Do Exercises or Dimensions explain the variance in OARS in a Call Center based Assessment Center?

Protocol Number: #15-035

Dear Investigator(s),

Your study has been designated to be exempt. The exemption is pursuant to 45 CFR 46.101(b)(4) Collection or Study of Existing Data.

We will contact you annually on the status of your project. If it is completed, we will close it out of our system. You do not need to complete a progress report and you will not need to complete a final report. It is important to note that your study is approved for the life of the project and does not have an expiration date.

The following changes must be reported to the Office of Compliance before they are initiated:

- Adding new subject population
- Adding a new investigator
- Adding new procedures (e.g., new survey; new questions to your survey)
- A change in funding source
- Any change that makes the study no longer eligible for exemption.

The following changes do not need to be reported to the Office of Compliance:

- Editorial or administrative revisions to the consent or other study documents
- Increasing or decreasing the number of subjects from your proposed population

If you encounter any serious unanticipated problems to participants, or if you have any questions as you conduct your research, please do not hesitate to contact us.

Sincerely,

Lauren K. Qualls, Graduate Assistant
Office of Compliance
615-494-8918