Exploring the Construct Validity of Principal Ratings as a Measure of Teacher Performance Using Meta-Analysis

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Kristi Kay Logan

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Nathan R. Kuncel, Adviser

December, 2014

Acknowledgements

- First, I would like to thank my advisor, Nathan Kuncel. Without him, this project would never have happened. He changed my life by asking a simple question. I can never thank him enough for all his assistance.
- I would like to thank my committee member, John Campbell, for signing off on this second chance and for his support over the course of my graduate career. He has been the best role model of what an I/O Psychologist can and should be.
- I owe thanks to my other committee members, Mark Davison and Aaron Schmidt, for their support.
- I owe a debt of gratitude to all of my many graduate colleagues from the University of Minnesota. We spent many long hours and late nights studying which was manageable with all of the laughter, silliness, eating out, and the occasional fondue.
- Special thanks goes to Eryn O'Brien and Sarah Hezlett for all of the many study nights and many mugs of tea and hot chocolate. They have continued to support me over the years and have been instrumental in my graduate career.
- During my time at PDRI, there were many colleagues who helped train me to be a better psychologist, especially Jerry Hedge, Cheryl Paullin, Janis Houston, and Mary Ann Hanson. I am so thankful for their guidance and support.
- I am grateful to my parents, Keith and Sandra Logan, and the rest of my family for all of their support over the years.
- I must send a special thank you to all of my friends in Auburn. They don't know what an I/O Psychologist is, but they have been instrumental in supporting the

other roles in my life. I couldn't have survived without all of the playdates, lunches, and afternoons on the corner. WAR EAGLE!

- My children, Cecilia and Henry, were not even a thought when this journey started. They inspire me and make me smile every single day. Maybe one day they will appreciate this accomplishment and it will teach them that it is never too late to reach for a dream.
- Finally, I would like to thank my husband, Mathew. He has supported me over the years no matter which path I chose to follow. He believed in me even when I stopped believing in myself. Without his love and support, I could never have been brave and said yes to this project.

Dedicated to

my grandparents, Howard and Agnes Harris,

who taught me through their words and actions how important education is.

Abstract

Multiple meta-analyses have been conducted exploring the relationships between subjective and objective performance, multi-source performance ratings, and ratings and personality. Rarely has this work included the occupation of teachers. This meta-analytic research explores three areas of teacher performance: (1) What is the relationship between principal ratings and student achievement scores? (2) What is the relationship between principal ratings and ratings completed by students, peers, parents, and other classroom observers, and the relationships between these ratings with one another? (3) What is the relationship between principal ratings and personality ratings completed either by the teacher (self-ratings) or others?

Data was gathered from published and unpublished sources and analyzed using the Hunter and Schmidt (2004) psychometric meta-analytic method.

The correlation between principal ratings and student gain was .17, similar to what has been found in the past (Medley and Coker, 1987). Value-added scores produced a greater relationship with principal ratings (r = .23) than when all gain results were used. Arithmetic tests (r = .24) exhibited the largest relationship with principal ratings.

Multiple source rater data results found that principal ratings had moderate relationships between peers (r = .57), students (r = .31) and other classroom observers (r = .45). Student ratings exhibited their largest relationship with parent ratings, (r = .53). Self-ratings had low relationships with all other rating groups.

Personality ratings generated either by teacher self-ratings or ratings by others using an overall prosocial personality factor correlated .28 and .45, respectively with principal ratings. Results using other ratings of personality showed that

Conscientiousness and Emotional Stability are the most important for predicting high levels of performance (r = .23, for both). The results exploring the relationships between principal ratings and self-ratings of personality found low correlations for the global Big Five dimensions, ranging from -.11 to .06.

Overall, this research provides new information about the relationships that principal ratings have with other criteria and predictors. It raises many questions that can be explored in future research.

Table of Contents

List of Tables

Chapter I

The merits of what makes a good teacher have probably been discussed since a teacher first set foot in a classroom. Research has been published in journals such as *The Elementary School Teacher* and *The Elementary School Journal* since the early 1900's. Researchers have consistently expressed how difficult it is to measure teacher performance. Advances in statistics are now available to help researchers study teacher performance in ways that had previously not been possible. Through the years, there have been different definitions of teacher performance and how it should be assessed (Goe, 2007). One method of measuring teacher performance that has remained constant over the years is the use of principal ratings. A century of research can now be evaluated using meta-analysis to broadly examine the construct validity of principal ratings as a legitimate measure of teacher performance.

Most of the previous research has been conducted by colleagues in education and educational psychology. Even though industrial-organizational psychologists study job performance, they have typically not done much research on teacher performance in primary and secondary schools. Both fields may benefit from research that integrates multiple perspectives. Therefore, the two objectives of this thesis are to make both a theoretical and empirical contribution to the definition and operationalization of teacher performance. First, I will present a synthesis of job performance theory from I-O psychology with research and theory from education with the goal of providing a new and hopefully useful perspective to a specification of teacher performance. Second, I will present a quantitative research synthesis of predictors and performance correlates of principal ratings of teacher performance. Specifically, this meta-analysis focuses on the measure of teacher performance that is most closely aligned with a large literature in I-O psychology (the principal as a source of supervisory ratings) and will examine its

relationship with personality traits as well as correlations between supervisory ratings and other measures of teacher performance including, value added scores, self-ratings, student ratings, peer ratings, and classroom observer ratings. These analyses will distill over 100 years of research and help establish how principal ratings overlap with other performance measures as well as examine its determinants.

What is Performance?

According to Campbell (1990), "Performance is behavior. Performance is *not* the consequence(s) or result(s) of action; it is the action itself (p. 704)." Performance is under the control of the person. If we are discussing job performance in most organizations, then the behaviors performed must be relevant to the goals of the organization (more on stakeholders in education later). Campbell (1990) proposes that there are three direct determinants of job performance: declarative knowledge, procedural skill, and motivation. That is, individual performance in jobs is the direct result of the knowledge and skills required to perform and the choice to perform each of the major elements of the job. Different jobs would be expected to have different combinations of direct determinants (that is, not perfectly overlapping knowledge and skill requirements), but all performance will be based on the same three direct determinants.

Declarative knowledge is the information or facts about the task to be performed. A teacher must understand how to subtract two numbers and the rules surrounding discipline in a school. Procedural knowledge and skill are the skills needed to perform one or more dimensions of the job. Procedural knowledge in teaching might be how to best present the concept of subtraction as well as troubleshoot a child's confusion. Motivation is characterized by three volitional choices: choice to perform, level of effort, and persistence of effort. A person may possess much declarative knowledge and skill,

but if he or she has little to no motivation, then there will be little to no performance (that is, the person will not act).

Further Campbell (1990) states, “performance is to be distinguished from *effectiveness* and *productivity*. *Effectiveness* refers to the evaluation of the results of performance (p. 705).” A measure of teacher effectiveness would include evaluating teaching success that results from the behaviors performed by the teacher plus other factors contributing to that result, plus error. A commonly used criterion to assess teacher effectiveness is student test score gains. Since a teacher does not have complete control over how well his/her students perform on standardized tests, it is debatable if such measures of performance are appropriate given that other factors affect students gains. Stated another way, can we assert that teacher performance (behaviors) exerts a direct and causal effect on test scores changes? There are multiple value-added models which attempt to enhance our ability to attribute score gains to teacher behavior (which is often unobserved by most stakeholders save for the teacher, students, and occasional observers). A separate issue that arises by using student test scores is how we measure the effectiveness of those teachers who teach subjects which aren’t assessed through standardized tests, such as music or art teachers. The effectiveness of these models is still being discussed, but it is important to note that much of the debate around teacher performance assessment comes down to the same basic issues in performance measurement discussed in Campbell (1990).

In more recent years, researchers have developed models of teacher performance, but have used different names such as teacher quality. Three of these models are discussed and synthesized in the following paragraphs. While these models or frameworks do not directly cite Campbell’s (1990) model, Campbell’s work can be used as a framework for the other models. Table 1 lists the three direct determinants

proposed by Campbell (1990), followed by the operationalizations used in the three other models proposed by Danielson (1996), Goe (2007), and Lai, Auchter, and Wolfe (2012). These different efforts to model teacher performance will be the basic sources of the unified model presented in this thesis.

Modeling Teacher Performance

Danielson (1996) used the criteria developed as part of her work on the PRAXIS III (the national teacher certification exam) to develop her framework for teaching. Danielson (1996) summarized that her framework "seeks to define what teachers should know and be able to do in the exercise of their profession (p. 1)." As one can see from this definition, the two parts of her framework match very closely with the concepts of declarative knowledge and procedural skill.

The framework consists of four domains which are broken down into twenty-two components. Each of the four domains is matched with similar dimensions in Campbell's (1990) model in Table 1. According to Danielson (1996), her framework is based on past empirical and theoretical research. Each of these performance components should improve student learning. The first domain is planning and preparation which consists of six components: demonstrating knowledge of content, demonstrating knowledge of students, selecting instructional goals, demonstrating knowledge of resources, designing coherent instruction, and assessing student learning. From looking at the components, one can see that this domain contains much of the teacher's knowledge about learning content, students and resources. It deals with how a teacher prepares for teaching in the classroom. The second domain deals more exclusively with how the teacher interacts with the students in the classroom and is labeled the classroom environment: creating an environment of respect and rapport, establishing a culture for learning, managing classroom procedures, managing student behavior, and organizing physical space. This

domain pertains to whether a teacher can create the right atmosphere for the students to be able to learn.

The third domain is instruction and is what most would consider “teaching.” It has five components: communicating clearly and accurately, using questioning and discussion techniques, engaging students in learning, providing feedback to students, demonstrating flexibility and responsiveness. This is where a teacher must impart his/her knowledge to the students and be skillful in how the material is presented so the students will learn most effectively. The final domain deals with those activities that a teacher performs outside the classroom and is called professional responsibilities: reflecting on teaching, maintaining accurate records, communicating with families, contributing to the school and district, growing and developing professionally, and showing professionalism. New teachers may have the most difficulty with the components in this domain as it is not the main focus of teacher instruction. Over time, their competence should grow.

In this framework, raters are presented with a behaviorally-anchored rating scale for each of the components. This allows flexibility in adapting the framework for different grade levels or subject areas by changing the associated anchors. Each component has four levels of performance: unsatisfactory, basic, proficient, and distinguished.

Goe (2007) conducted a research synthesis on teacher quality. She distinguished between teach***er*** quality and teach***ing*** quality. Teacher quality deals with the set of inputs that a teacher brings into the classroom, such as college degrees, test scores or certification (a mix of direct and indirect performance determinants in Campbell’s model). Teaching quality consists of what a teacher does in the classroom – his/her behavior. Note that this framework fits nicely with the earlier model that differentiates between job performance (teaching quality) and the determinants of

teaching quality (teacher quality). This synthesis led to the development of a framework for teacher quality. It is composed of two sets of inputs (teacher qualifications and characteristics), processes (teacher practices) and outcomes (teacher effectiveness). Once again, we see reference to the knowledge that a teacher has paired with how a teacher shares that knowledge with his/her students. The parts of Goe's (2007) model can be found in Table 1.

Lai, Auchter and Wolfe (2012) conducted a confirmatory factor analysis to explore if teacher quality can be thought of as a two-factor construct built of teaching skill and content knowledge. They used assessment scores that teachers completed as part of their National Board Certification process. The assessment consisted of ten standards-based components. A teacher was asked to submit four portfolio entries which represented work in the classroom. Two entries were videos of the teacher conducting a lesson: one leading the whole class in a discussion and the other leading a small group. The third entry was an example of at least two students' work, each of which represented a different learning profile. The fourth entry highlighted the teacher's work with colleagues, parents and the community. The other six components were generated as part of work at an assessment center. These exercises measured a teachers' subject-matter knowledge.

Results supported a two factor model comprised of teaching skill and content knowledge. These two factors have been incorporated into Table 1. The model was tested in four different certification areas: adolescent/young adult English Language Arts, early adolescent Math, early childhood Generalist, and middle childhood generalist. A similar factor structure was found across the four groups. There were two differences in the factor structure worth discussing. First, the subject-specific groups exhibited a weaker relationship between teaching skill and professional collaboration. Lai, Auchter,

& Wolfe (2012) suggest that teachers who teach in the younger grades and teach more subjects may require more professional collaboration than teachers of older students. A second difference found that the content knowledge factor explained a larger proportion of the variance with the subject-specific groups. On the other hand, the opposite was found for the generalist group. The teaching skill factor explained a larger proportion of the variance in the teaching portfolio.

Overall, if we compare the results of this study with Campbell's (1990) model of performance, the similarities are striking. Teacher quality is just another name for teacher performance. The factor of teaching skill matches well with procedural knowledge and skill, whereas, the content knowledge factor aligns well with declarative knowledge. From looking at these models, there is a consistent opinion that there are two common determinants of teacher performance. It is again important to note that the literature does not consistently differentiate between performance (behavior exhibited on the job) and the direct determinants of performance (individual characteristics that are necessary for performance such as skill and knowledge).

Where is the third determinant of performance, motivation? There seems to be a lack of the motivation construct in the teacher performance models. Lai, Auchter and Wolfe (2012) acknowledge that they did not include a motivational element in their test model because it was assumed that teachers were highly motivated due to their participation in the optional certification process. This statement tacitly refers to the concept of maximal versus typical performance (e.g., DuBois, Sackett, Zedeck & Fogli, 1993) which is important when considering different operationalizations of teaching performance. However for theory building the exclusion of motivation from many models is puzzling. Surely there are some motivational differences across individuals for the job.

So much so, that I think it is safe to argue that any model of teacher performance that does not acknowledge motivation is deficient.

It is unclear whether researchers believe it is not relevant or are making the assumption that if a teacher shows up for work, then there will be at least a minimal level of motivation. In addition, it may also be assumed that motivation is always high or, if not, little can be done to influence it. Prick (1989) points out that Sarason (1977) had stated that teaching may be considered more of a vocation, like a clergyman or a doctor, instead of a job. Thus people who work in that context are going to be happy with their work no matter what. While it does make sense that when working with children, a teacher would have to perform at least a minimum number of duties to keep the classroom moving or the children would object or become out of control. It is not clear that this always occurs. Additional effort is also likely to result in better teaching for many tasks. Extra effort both within and outside of the classroom would be critical for a number of activities including providing good quality feedback to students, preparing lessons, and communicating with parents.

The lack of motivation in theories of teacher performance is a problem. There has been a history of researchers examining the relationship between teacher performance and the individual difference variable of personality (Chapter III will summarize some of this research). Personality is considered an indirect determinant of performance. It has been theorized that personality has a distal connection to performance through motivation. This is one area where I-O Psychologists are developing new theories which may contribute to the education literature. Johnson (2003) and Johnson and Hezlett (2008) have developed a more comprehensive model of performance which maps out how personality can mediate the relationship between

motivation and performance. These models give us new directions to help us place motivation within a model of teacher performance.

Looking at Table 1, we can see that there is some agreement between all of these models of performance. Whereas Campbell (1990) arranges his model hierarchically and delineates the direct determinants further into components, only one of the previous models of teacher performance (Danielson, 1996) does so. The next section will describe these components of performance.

Components of teacher performance. Campbell (1990) suggests that rather than looking for a single ultimate criterion of overall performance, eight performance components provide a superior specification: job-specific task proficiency, nonjob-specific task proficiency, written and oral communication tasks, demonstrating effort, maintaining personal discipline, facilitating peer and team performance, supervision, and management/administration. Each of these components is listed in Table 2. More recently he has organized the performance literature hierarchically (Campbell, 2013). Of course, not all of these components will be relevant for all jobs including teaching, but all will include job-specific task proficiency, demonstrating effort, and maintaining personal discipline (Campbell, 1990, 1996). I next map these dimensions onto teacher performance.

Teachers will vary on which core tasks make up their job-specific task proficiency due to drawing on different content knowledge. Primary teachers will teach a range of subjects, such as math, reading, science, social studies, etc. As the age of the students increases, teachers typically specialize in the subjects that they teach. Thus a middle school or high school teacher will most likely teach only one of these subjects. There will also be a subset of core skills required by teachers for teaching students in any of the above required subjects. These would be the types of skills learned during their teacher

preparation programs. Teachers will take classes to learn how to lead their students in small groups, how to structure a lesson plan, how to construct an exam, as well as many other teaching skills required to be a successful teacher. There will also be non-job-specific components. All teachers must grade their students' work, discipline students, take part in meetings with other school personnel and parents, as well many other behaviors. Danielson's (1996) framework attempts to capture relevant in class and out of class teacher behaviors such as these.

It is difficult to find studies which try and capture the full range of behaviors a teacher will perform. It is instructive to consult O*NET and look at how the job of teacher is defined. There are three categories which are most relevant to this study: elementary teachers, middle school teachers, and secondary teachers. While there are differences among the three categories, there are still many similarities. Core tasks for all three jobs consist of such tasks as instructing students, adapting teaching and materials to the needs of the students, establishing clear objectives, establishing and enforcing rules, meeting with parents to discuss progress or behavior, and preparing materials for class. I incorporated components from Danielson's (1996) framework and the O*NET with Campbell's (1990) components in Table 2.

It is obvious that oral and written communication skills are a very important component of teaching. If a teacher cannot effectively communicate what he or she wants the students to learn, then there will be little learning. This is supported by the fact that speaking is one of the top three skills listed for all three types of teachers in O*NET. Writing isn't rated as being as important, but it still falls within the top eleven for all three types of teachers. The importance of communication is also found in Danielson's (1996) framework. Three of her components deal with effective communication.

Demonstrating effort will also be critical given the demands of teaching. Some teachers will leave the building as soon as possible, while others will spend much more time in the preparation of lessons, communicating with parents, and providing additional student feedback. As discussed previously, many researchers don't assess any type of motivational component of teacher performance. O*NET does find the following work styles important for teachers: achievement/effort, initiative, and persistence.

A teacher must maintain personal discipline. If a teacher were to abuse alcohol or drugs, it may stand out more in the school setting. The consequences may be greater as well, especially if a teacher is working with small children. A teacher must have lessons prepared for each day or risk students losing interest. Tests and assignments need to be graded and returned in a timely fashion or students or parents may make a fuss. Maintaining personal discipline may also include exhibiting productive reactions to students when they are frustrating, inattentive, or even belligerent. O*NET includes the following work styles in its definition of teachers: dependability, integrity, self-control, and stress tolerance. All would have some impact on the personal discipline of a teacher. This is included in Danielson's (1996) framework with the showing professionalism component.

The next two components are where primary and secondary teachers may differ from other occupations. Teachers have peers within and between schools. There are multiple ways that one can think of a teacher as part of a team. The clearest example would be, of course, team taught courses. But less formal teams can exist as well. If a school has more than one teacher teaching a specific grade, then there would be a team of teachers from each grade. They would most likely work together to make sure that their lessons are comparable to one another. An entire school could also be thought of as a team. The place where things may become a bit more confusing is whether we

think of a teacher as being a team member with their students or as a supervisor. It is most likely that it is a bit of both. Teachers certainly set goals for their students, monitor their performance, and discipline them when necessary. Teachers also play a management/administrative role because they must be advocates for their students in terms of goals and resources in the school.

The work activities listed for teachers in O*NET provides support for these behaviors. Facilitating peer and team performance has several work activities which correspond such as communicating with supervisors, peers or subordinates, establishing and maintaining interpersonal relationships, coordinating the work and activities of others, and resolving conflicts and negotiating with others. There are some corresponding work activities for supervision: training and teaching others, coaching and developing others, guiding, directing, and motivating subordinates. As for the management/administrative component, a knowledge listed for teachers is administration and management. It can also be found listed under work activities with performing administrative activities.

Danielson (1996) has several components which fall under these categories. Components that have to do with establishing an appropriate environment are associated with facilitating peer and team performance: creating an environment of respect and rapport and establishing a culture for learning. There are no specific components which line up with supervision. There are three components which match well with the concepts of management/administration: managing classroom procedures, managing student behavior, and contributing to the school and district.

There is one specific instance when a teacher takes on a more traditional supervisor role, which is when serving as a supervising teacher to a student teacher. Periodically, a teacher is asked to let a student teacher participate in the teaching of the

class for a semester. There has been some research to determine which behaviors comprise this role for the supervising teacher. For the sake of completeness, I review the structure of the supervisory dimension next.

Teacher Supervisory and Mentoring Performance

Roth (1961) collected critical incidents from seventeen elementary student teachers. There were a total of 142 useable critical incidents, of which 101 were classified as effective and 41 were classified as ineffective. These incidents were grouped into nineteen behavioral criteria which are listed in the first column of Table 3.

Copas (1984) also used elementary student teachers and collected data from a sample of 476, who attended 31 different institutions of higher education. The process generated 1125 useable critical incidents. The incidents were sorted into two general categories: critical requirements of cooperating teachers that affect student teachers and critical requirements of cooperating teachers that affect children. The requirements that affect student teachers was the category most closely aligned with Roth's (1961) study and included the following dimensions of behaviors: orienting, inducting, guiding, reflecting, cooperating, and supporting. Examples are sorted in the second column of Table 3 which most closely aligned with Roth's.

A third study was done by Farbstein (1965). In this study, there were 300 student teachers who generated 703 examples of effective and ineffective behaviors of cooperating teachers. These were classified into five areas: provides supervision, provides opportunities for growth in classroom instruction, demonstrates superior teaching ability, exhibits commendable personal traits, and exhibits commendable social traits. Once again, incidents which were found by Farbstein are sorted in the third column of Table 3 which matched those found by the previous two studies.

All three studies provide useful information when thinking about and describing the supervisory aspect of teacher performance. Looking at Table 3, there is some consistency to the behaviors that the student teachers believed were most helpful in their student teaching experience. Student teachers want their supervising teacher to treat them equally and provide them with goals and opportunities for teaching experience. Providing feedback was also important for the supervising teachers to give their student teachers. This research offers a starting point for those who want to delineate this aspect of teacher performance more clearly.

Each of these models breaks down performance into its important components. All of these models are focused on the successful completion of tasks. More recently, I-O psychologists have separated performance into task and contextual components. The following section discusses this research more fully.

Task and Contextual Performance

The previous sections discussed ways that performance has been described. This section discusses how task and contextual performance have been defined and their relationships with other predictors. Borman and Motowidlo (1993) discussed four ways that contextual performance differed from task performance: (1)"support the organizational, social, psychological environment in which the technical core must function," (2) "activities are common to all jobs," (3)"behaviors are probably better predicted by volitional variables related to individual differences in motivational characteristics and predispositional variables represented by personality characteristics", and (4)" generally are not role-prescribed (p. 74)." They further discussed how contextual performance is similar and distinct from organizational citizenship behavior and prosocial organizational behavior. Since this literature is not relevant to this dissertation, readers are encouraged to read other discussions about these concepts if

they are interested and would like further information (Borman and Motowidlo, 1993 and 1997).

Motowidlo and Van Scotter (1994) began to explore whether task and contextual performance were two distinct constructs. One way to show this distinction was to see whether each correlated in different ways with individual difference variables such as ability and personality. Supervisor ratings of Air Force mechanics were used to determine whether task and contextual performance each contributed unique variance to ratings of overall performance. Results confirmed that this was true. Further, it was found that task performance had a stronger relationship with experience. Contextual performance showed stronger relationships with several of the personality dimensions. Ability was predicted to have a stronger relationship with task performance, but in this case, contextual performance had the significant relationship. Results began to show that task and contextual performance both contributed to overall performance and had distinct relationships with other variables.

Borman and Motowidlo (1997) presented a taxonomy of contextual performance after reviewing the research on task and contextual performance. They found that contextual performance is weighted equally with task performance when supervisors are evaluating overall job performance. Overall, results have shown that measures of ability are more strongly related to measures of task performance and measures of personality are more strongly related to measures of contextual performance.

After the introduction of the concept of contextual performance, much research was conducted. Individual researchers developed their own models and defined dimensions that they found in their studies. Coleman and Borman (2000) synthesized much of this research. Their purpose was to organize the relevant models and conclude which constructs make up contextual performance. By looking over the previous

literature, twenty-seven dimensions were identified. Each of these dimensions and its definition were sorted by a group of forty-four I-O psychologists. Similarity data was calculated across all of the judges producing a correlation matrix. This matrix was analyzed using exploratory factor analysis and MDS with a cluster analysis. Both sets of analyses produced similar results. Three categories of contextual performance were identified: (1) interpersonal citizenship behavior – "behaviors benefiting organization members," (2) organizational citizenship behavior – "behaviors benefiting the organization," and (3) job/task conscientiousness – "behaviors benefiting the job/task (p. 41)."

From this brief summary, we can see that the field has expanded its definition of job performance to include both task and contextual performance. Research has shown that both are relevant when a supervisor is making overall performance ratings of subordinates. Since researchers are still trying to determine the dimensions of teacher performance, this past research may inform them on what some of the possible components may be.

Summary

From this discussion, it is clear that teacher performance is composed of many different components. Although the importance of these components may vary depending on the subject taught and/or the grade level, there is some consensus on the basic dimensions of teacher performance. There is one point on which all can agree: teacher performance is multi-dimensional. With so many different components, how should we measure all of them? In the next chapter, a summary of different ways to assess teacher performance will be presented.

Chapter II

Unfortunately, assessing a teacher's performance has been a problem for researchers for many years. The long standing lack of an appropriate criterion was expressed, over forty years ago, by Gough, Durflinger & Hill (1968): "The prediction of performance in student and/or professional teaching is one of the long-standing, unsolved, and perhaps (some would say) unsolvable problems of educational psychology (p.119)." They further state that there are only three criteria that one can use to assess the effectiveness of teachers: subjective ratings by a supervisor after observing the teacher in the classroom, student achievement scores, or ratings of the teacher by the students themselves. This dissertation will use all three of these criteria and will explore the relationships among them. The ways in which each of the criteria is used has evolved over the years due to advances in statistical methodology and research. The use of student achievement scores has changed most dramatically. Gough et al. (1968) didn't mention the use of ratings by others who may also offer opinions on the performance of teachers, such as peers or parents. These types of ratings were being collected at the time of their paper and previously to it. Each of these criteria will be discussed in the following sections, as well as a discussion of the relationships among these different criteria.

Prior Research on Supervisory Ratings

In most occupations, ratings are used to assess performance. An estimate was given by Bernardin & Beatty (1984) that over 90% of the ratings reported in the literature were supervisor ratings. Thus the use of supervisor ratings is prevalent. With the popularity of 360-degree feedback and the use of more work teams, other types of ratings (peer and subordinate) are being used more frequently. Research on teacher performance loosely fits the multi-rater model with supervisor (principal), peer, and

subordinate (student) categories. However, the job of a teacher differs from many work settings in multiple, important ways. First, the opportunity to observe performance differs. Second, the nature of the relationships between the different perspectives and the teacher are not exactly the same as in other work settings. Third, the number of stake holders with teacher performance is larger than what is seen in typical work settings.

In the field of teaching, a teacher's supervisor is the school principal. The use of a principal's rating of a teacher's performance has not been met with a great deal of enthusiasm. During the early years of research, some findings showed that the correlation between principal ratings and a direct measure of teacher effectiveness were near zero (Medley & Coker, 1987). According to Medley & Coker (1987), these ratings "are only slightly more accurate than they would be if they were based on pure chance (p. 243)." In the past, researchers have asked principals to make their teacher ratings using poorly constructed rating forms. A principal gave one single overall performance rating on a numerical scale without any reference points or behavioral anchors. A problem with this type of rating, particularly in the absence of rater training, is that there will be little consistency between raters or between multiple ratings given by the same rater. Each will use his or her own definition of what effective teacher performance is at that time to make the rating. Some researchers have found better results after providing training to their raters (Mannatt & Daniels, 1990; Heneman, Milanowski, Kimball, and Odden, 2006).

In order to determine what was being assessed by a principal's ratings, some researchers have collected ratings from other sources: students, parents, and teacher self-ratings. A principal's ratings were then correlated with the ratings of others. The main reason for gathering ratings from different sources is based on the assumption that

different raters have a unique perspective on performance. By using different rater sources, these unique perspectives can be captured to provide a fuller perspective on job performance. Thus we would *not* expect the ratings from different sources to be in complete agreement (Borman, 1974). Recently, two studies (Mount, Judge, Scullen, Sytsma, & Hezlett, 1998; Scullen, Mount, & Goff, 2000) have suggested that there may not be variance attributable to the rater source, but it should be characterized as comprising all idiosyncratic rater variance. Hoffman, Lance, Bynum, & Gentry (2010) reanalyzed the data from the two previous studies to determine whether rater source does account for variance in multi-source performance rating data. They hypothesized that a different model would fit the data better, which did account for source variation. Their results found that a model which specified three performance dimensions, seven idiosyncratic rater factors, and three rater source second order factors fit the data most parsimoniously. Another model which also included a general performance factor also fit the model, but had little practical significance. Hoffman et al. (2010) concluded that rater source does account for a larger proportion of variance than was previously reported. In their study, it amounted to an average of 22% of the variance. Thus different rating sources are contributing unique information.

Relationships between rater sources across jobs. Researchers have conducted meta-analyses using data from other jobs on multi-source performance ratings: subordinate, supervisor, peer and self (Harris & Schaubroeck, 1988; Conway & Huffcutt, 1997). Overall, Harris & Schaubroeck (1988) found higher correlations than the Conway & Huffcutt (1997) meta-analysis. Both studies found their highest correlations for ratings between supervisors and peers [r =.62 (Harris & Schaubroeck, 1988) and r = .34 (Conway & Huffcutt, 1997)]. Correlations between supervisor and self were low (r = .22) as well as the correlations of peer with self (r = .19) according to the results found

by Conway & Huffcutt (1997). On the other hand, Harris & Schaubroeck (1988) found more moderate correlations for self-supervisor (r = .35) and self-peer (r = .36). Only the Conway & Huffcutt (1997) study used subordinate ratings. Subordinate rating correlations with other sources were low (supervisor r = .22; peer r = .22; and self r = .14). Both studies looked at whether job type was a moderator. Both found higher correlations for non-managerial jobs as compared to managerial jobs using the two self-rating comparisons with supervisors and peers, but Harris & Schaubroeck didn't find a moderator effect for the peer and supervisor relationship.

Relationship between principal ratings and other rating sources. There has been multi-source performance rating research conducted in the educational setting. In these studies, researchers have explored the relationships between principal ratings and other sources of ratings such as self, peer, student, and parent.

Ostrander (1996) conducted a study using multiple raters with teachers ranging from third grade all the way through high school. A questionnaire was designed to assess six categories of teaching: classroom environment, homework, grading, communication, instruction and interpersonal relationships. An overall mean rating of the teachers was also calculated. Teachers were rated using a four-point scale ranging from strongly disagree to strongly agree. Results showed that the strongest correlation among the four groups using the overall total score was between parents and students (r = .544). There was a moderate correlation between the principal's ratings and the student's ratings (r = .345). No relationship was found between the principal's ratings and the parent's (r = .012) or the teacher's self-ratings (r = .079).

Wilkerson, Manatt, Rogers and Maughan (2000) conducted a study which used a 360-degree feedback approach. Similar instruments were developed for each of the three groups. The student feedback questionnaires consisted of twenty items. The lower

grades (K-2) used a three-point scale, while the remaining grades (3-12) used a five-point scale. The teacher self-feedback instrument and the principal feedback instrument both consisted of twenty items using a five-point scale. Each principal also completed the district's summative evaluation instrument which consisted of fifteen items using a four-point scale. There were fairly high positive correlations between principal ratings and student feedback (r = .52) and teacher self-ratings (r = .30), but they were not reported as being significant at the .01 level. When the principal summative evaluation ratings were used, none of the correlations were significant at the .01 level for student feedback (r = .72), teacher self-ratings (r = .62) and principal ratings (r = .62), even though the values were fairly high.

It makes sense that there are not extremely high correlations among the different sets of ratings. Each group may view the teacher from a slightly different perspective. Each of these perspectives is important and adds a unique piece to the teacher performance puzzle. It is encouraging that there was a relationship between the principal's ratings and the student's ratings. Students by far spend the most time with the teacher and see the teacher's performance on the widest range of activities. Since a principal does observe the teacher within the classroom setting, those ratings should have a relationship with what the students are basing their ratings on as well.

Comparison of rating data. At this point, we should compare the results found in the multi-rater meta-analysis with the results found in the studies using multiple rating sources in a teacher population. There were only two areas which had rating data which could be compared: supervisor with self and supervisor with subordinate (student). The correlations between principal ratings and teachers' self-ratings were at all different levels, whereas supervisors and self-ratings were in the moderate range. The

correlations between principal ratings and student ratings were higher than those found for supervisor and subordinate ratings.

The comparison shows that the correlations between types of ratings may differ in the teaching setting from other work settings. There are several reasons why this may occur. In some ways, teachers are an occupation by themselves, not falling neatly in the managerial or blue-collar categories. Teachers most likely would be classified as closer to a managerial job than a blue-collar job. Most principals were teachers before being promoted. Thus their perspective as a supervisor may give them greater knowledge about the job. The same may be said for peers. Since teachers all essentially perform the same job, one might expect the supervisor, self, and peer ratings to be slightly higher than in other occupations. Although that may be tempered by the fact that peer teachers may only observe each other under a limited set of circumstances. Student ratings will also vary from subordinate ratings. Children may exhibit different types of rating errors than adults. A review done by Follman (1992) found that secondary students' ratings of teacher effectiveness were prone to the same types of errors (leniency and halo) as college students and other groups.

The teacher studies included another rater source (parents) that has no match in data using other jobs. This provides a unique perspective looking at the relationships of parent ratings with their own children as well as with the principal. We should expect a fairly high correlation between student and parent ratings, especially as the grade of the student increases. In these cases, parents are less likely to visit the classroom and more likely to obtain their information from their children (Ostrander, 1996). The level of agreement between the principal and the parents is another area that may be explored when using a teacher population.

The teacher profession has another category of outside observers/raters besides parents which will be explored in this meta-analysis. These are ratings provided by other administrators. The most comparable group in other occupations would be a supervisor up one level or more from the direct supervisor. One reason there is no comparison data in the Conway and Huffcutt (1997) study was that they did not use any ratings from supervisors which came from more than two levels above the ratee. In the teaching profession, these raters can range from the school superintendent to a county supervisor. This type of rating was more common in the past when schools were smaller (one or two room schoolhouses) and the superintendent was tasked with hiring staff for all of his/her schools and was thought of as more of a direct supervisor. Ratings by this group would be expected to be similar to the principal's ratings. It is likely that a superintendent has some knowledge of the performance level of the teachers and would be capable of making ratings.

There is another group of outside observers providing ratings of a teacher's performance: researchers. Many of the researchers conducting these studies went in the classrooms and observed the teachers in action. Afterwards, the researcher would make his or her own ratings. Further, many studies involved a group of outside observers who were trained to make ratings. Several observers would be sent to the different classrooms over a period of time to provide ratings. It is anticipated that this group of raters has a different perspective from administrators who have observed the teachers over time in multiple situations inside and outside of the classroom. These outside observers may only have their experience of watching the teacher conduct the class for an hour or so on which to base their ratings. It is projected that the outside observer ratings will show a smaller relationship with principal ratings.

For the meta-analysis in this dissertation comparing principal ratings with the ratings of other groups (students, parents, peers, self), I expect low to moderate correlations of a principal's ratings of a teacher's performance with ratings done by other groups. The correlation of student ratings with principal ratings should be moderate and may be higher than subordinate-supervisor ratings from other jobs. Students should observe the same behaviors as principals when the principal visits the classroom and observes behavior in that setting. The degree of the relationship between principal ratings and student ratings may provide some insight into the task work of the teacher. On the other hand, if a principal makes ratings also using those behaviors observed outside of the classroom (this could include both contextual behaviors as well as those concerned with development or interaction with other personnel), it will impact the level of the relationship.

The correlations between peer ratings and principal ratings will be expected to be in the moderate range. Peers are able to observe many of these same behaviors, so the relationship of their ratings should be higher with the principal as compared with supervisor-peer ratings in other jobs.

I would expect the relationships between principal ratings and teacher self-ratings to be lower than what has been found across other jobs. Teachers have a more unique perspective on how they are performing which may not be in direct agreement with the principal. Some research has shown that teachers rate themselves more highly than their principals (Payne and Hulme, 1988), which is similar to other occupations. Thus this correlation would be low.

The relationship of the principal ratings with ratings by other sources should provide differing perspectives on the performance of the teacher in and out of the

classroom. Now let us turn our attention to how a teacher's performance impacts how well students perform on standardized tests.

Student Achievement Scores

Student achievement test scores have been considered the direct measure of teacher performance. Therefore, Medley and Coker (1987) compared their principal rating data with student achievement scores. Since they declared the principal rating data lacking, the assumption must be made that they considered the student achievement test scores to be the "true" criterion. There are numerous ways that student achievement scores can be used. Morsh and Wilder (1954) divided student change into five classes: "raw gain (posttest minus pretest scores); achievement or accomplishment quotient (ratio of pupil's educational age to his/her mental age); miscellaneous measures; corrected raw gain (raw gain corrected for initial intelligence, grade, or other variables); and residual gain (actual gain minus predicted gain) (p. 51)." Researchers calculate the relationship between test scores (or gain scores) and the principals' ratings. Thus, these student achievement measures will vary with researcher and model used to calculate a student achievement score. As statistical methodology has advanced, more elaborate methods have developed to isolate improvements on student test scores that are attributable to a teacher's performance.

Value–added. One method of assessing teacher effectiveness that has been growing in popularity is value-added models. Although value-added model estimates are not the focus of this thesis, data relying on value-added analyses will be used to explore its relationship with principal ratings. These models estimate individual teacher's contributions to student achievement by examining student test scores over time. Many researchers believe that this is an improvement over more traditional methods of examining student's test scores. Sanders (2000) states that the new models are better

because they measure a student's academic progress, not a designated achievement level at a specific time. Thus, each student can make academic progress at individual rates. In the past, students were expected to progress to a certain academic level at a certain time. Value-added researchers believe that by using individual academic progress of students, then they will be able estimate these rates without any confounding due to socio-economic factors. Researchers make decisions as to how they want to specify the model. So there may be some differences in results depending on the model. Sanders (2000) believed that a teacher has "primary control of the rate of academic progress of their students (p. 331)." Since value-added models measure the productivity of a teacher, it is not possible for a teacher to have complete control over the performance of his or her students. Teachers also may face barriers outside of the classroom which will impact the performance of their students.

Not all people are convinced that determining a teacher's performance using value-added models is the best course to take. Goe (2007) pointed out that there are three reasons why it is difficult to measure teacher quality (performance) using standardized achievement test scores. First, these tests were designed to measure the performance of the students and not the teachers. We begin at an inferential disadvantage. Second, it is challenging to distinguish between those effects attributable to teachers and those attributable to the classroom. Finally, it is difficult to align a student's test data with the appropriate teacher. It is only with the advances in technology that the last problem has become more manageable. There is an ongoing discussion as to the proper way to use these statistical models to determine the effect of the teacher on student learning.

Since value-added models are becoming more popular and school districts are using the results for high stakes decisions, more research is being conducted on the

accuracy of such models. A report by Schochet and Chiang (2010) explored what the likely error rates would be using value-added models with different amounts of data. Their results found that with three years of student achievement data the likelihood of making a Type I or Type II error was twenty-five percent. If only one year of data was available, the error rate rose to thirty-five percent. This means that if a teacher had an average level of performance, one out of every four teachers would be misclassified as either being very good or very poor.

Papay (2011) found that the student achievement test used provided different results for ranking teachers on the value-added of their teaching effectiveness of their students. It is disturbing to think that consistent results are not found if different outcome measures are used. Results showed that the differences were not a result of test content, scaling, or the sample used. Papay also found that what time of year the tests were administered had an appreciable impact on the results. The Spearman rank correlations were higher when comparisons were made between tests given at the same time of year as compared with tests given at different times of the year. Papay hypothesized that these differences may be attributable to different summer's learning loss. The good news is that those teachers who had students perform well on one outcome measure, generally performed well on another outcome measure. It still begs the question whether we should make the assumption that the value-added scores are in fact the "true scores" if the results differ based on which test is used.

Achievement tests measure different constructs. Depending on what the content is that the teacher teaches will make a difference on how the student scores. For example, a teacher may use teaching methods that help the students learn the material at a higher level. On the other hand, if the test measures a student's knowledge at the most basic level, the higher level learning will be unknown and not assessed. According

to Papay (2011), a teacher may also introduce bias into the scores as well. If a teacher structures the content of the class to match the content of the test, student scores may be higher.

Hanushek & Rivkin (2006) point out that achievement tests will rank the effectiveness of teachers differently depending on the previous level of achievement of the students and/or the knowledge required by the test. An example given by Hanushek & Rivkin (2006) is the case where a test asks questions that tap knowledge that was gained prior to the current school year. In this case, it will be difficult to identify differences in teacher effectiveness. Another example deals with a test that doesn't cover information taught by the teachers. Gains would be higher for lower achieving students than initially higher achieving students. Researchers make the assumption that tests are measuring knowledge at the ratio or interval level when many times that assumption is not met.

All of these are legitimate reasons why we should use caution when making decisions based on value-added measures of teacher performance. It may not be advisable to make high stakes decisions until some of these issue are explored and resolved. The fact remains that researchers continue to compare results found about teacher performance based on principal ratings and student achievement scores. The next section discusses this debate more fully.

Principal Ratings Versus Value-Added

Many researchers have compared principal ratings with the achievement test scores as a test of the legitimacy of ratings. When the relationship between the two is not high, ratings are found to be lacking. This may not be the appropriate conclusion. We should not expect the two measures to be the same. They are separate ways to measure a teacher's performance and may not be influenced by the same behaviors.

Comparing one against the other as if one is the true score is certainly premature and probably a mistake. If both are properly measured, we should think of them as two pieces of the puzzle. Both give information that is helpful in measuring performance. Ratings may be the only method to obtain measurement of some behaviors. The principal may be one of the people in the best position to provide those ratings.

It seems logical that student achievement data, all else equal, captures part of what a teacher does in the classroom. More specifically, how well the teacher teaches the relevant content knowledge needed for the student achievement test. Looking at Table 2, this would refer to those things found under job specific task proficiency. It is not as clear how some of these other components in Table 2 would directly impact student achievement scores. On the other hand, ratings would be able to capture more of the other components in Table 2 as well as job specific task proficiency. A principal would be better able to assess how well a teacher communicates, supervises, manages, and facilitates team performance. Other groups of raters would also have the opportunity to witness how a teacher performs these tasks as well. Further, ratings may be a better method for assessing contextual performance. These behaviors may not be assessed at all when using student achievement scores.

Ratings have been labeled as a subjective measure of performance by researchers for many years (Heneman, 1986). Since student test performance is used as a criterion for teacher performance, how should we classify it? Many researchers call it an objective measure of teacher performance. Yet, it can only estimate what a student's gains are in a particular subject, but drawing causal inferences about the extent to which change is due to the teacher is highly problematic. There are many potential sources of systematic error which can affect it: students are not randomly assigned to teachers, class/school characteristics, student characteristics, test

characteristics, parent effects, etc. In fact, Berk (1988) discusses at least 50 factors which can have an impact on a teacher's effectiveness which are out of the teacher's control. Many of these are listed in Table 4. Thus it is questionable what specifically the teacher's impact has been on the student's performance. Additionally, it is only focused on gains on academic topics (which are undeniably important). Arguably, teachers are at least as important in training other skills that fall outside of the academic curricula. Further, there are many teachers who teach subjects or skills which are not assessed by student achievement tests: art, music, computer skills, etc. We must have some way to assess their performance if we can't rely on a student achievement test.

It might be helpful at this juncture to look at what the research has found when comparing subjective and objective measures of job performance. In a meta-analysis conducted by Heneman (1986), he explored the relationship between supervisory ratings and results-oriented measures of performance. After correcting his data for sampling error and attenuation, Heneman found a weak relationship between ratings and results (corrected mean correlation of .27). Thus he concluded, "ratings and results cannot be treated as substitutes for one another" (p. 818). Here is evidence that we should not equate measures of performance with measures of effectiveness. Each criterion adds to our knowledge of performance, but they are not interchangeable with one another.

Bommer, Johnson, Rich, Podsakoff & Mackenzie (1995) followed up several years later and retested Heneman's hypotheses using a larger sample of studies. Their results found a corrected mean correlation between subjective and objective performance measures of .389. The confidence interval did not contain zero, so they concluded that the measures were significantly related. While they found a relationship

between the two, their results supported Heneman's (1986) conclusion that subjective and objective performance measures are not interchangeable.

Throughout all of this research, the researchers are classifying the principal's ratings as subjective and the gain scores calculated from student's performance as objective. It seems to make a very large assumption that the gain scores are measuring the true score of a teacher's performance. Are we going to make the assumption that a gain score metric is a true score? If we compare a principal's rating against this criterion, why should we assume if the correlation is low that the principal has made inaccurate ratings? The important question may not be which of these two methods are "correct." The more important question may be trying to determine what each of these methods is measuring of teacher performance. Next, let us look at this research which has explored the relationship between principal ratings and student achievement scores.

Relationship of Principal Ratings with Student Achievement Scores

Early research found that the correlation between a principal's ratings and student achievement scores was near zero. Medley & Coker (1987) state that their results concurred with the previous research that they reviewed (Anderson, 1954; Barr, Torgerson, Johnson, Lyon, & Walvoord, 1935; Brookover, 1945; Gotham, 1945; Hellfritsch, 1945; Hill, 1921; Jayne, 1945; Jones, 1946; LaDuke, 1945; Lins, 1946; Medley & Mitzel, 1959). In their study, they argued that principal's ratings were not accurate for assessing the performance of those teachers they supervised (N = 46 principals and 322 teachers). They state that the results can't be blamed on any limitations in instrumentation. Although looking at the criterion measure they used, it could be considered psychometrically poor by today's standards. It consisted of a single principal's rating for each teacher on a scale of 1 to 20 on three different roles performed by the teacher.

A follow up study was conducted by Manatt & Daniels (1990) who tried to determine whether principals were able to rate their teacher's performance as compared to student achievement (N = 19 8th grade teachers and N = 34 4th grade teachers). This study attempted to account for some of the weaknesses in Medley & Coker's (1987) study such as providing an accurate conception of teaching, training the raters how to use the rating form, type of evaluation, using principals from more than one school, to name a few. The principals rated each teacher on a Teacher Performance Evaluation Instrument which contained 25 criteria using a scale of 1 (low performance) to 7 (high performance). Two types of achievement tests were administered during the year: a norm-referenced test and a criterion-referenced test developed by the school district which was aligned with the criteria of the evaluation instrument. A stepwise multiple regression technique was used to determine if there was a relationship between the ratings and posttest scores after the effects of the pretest had been removed. Thus the posttest score was used as the dependent variable and not a gain score. Results showed that there was no relationship between the principal's ratings and student achievement on the norm-referenced tests. On the other hand, ratings did account for variation in the posttest on the criterion-referenced tests. Different criteria were significant depending on grade level and subject. In this study, 4th graders were assessed in math and reading and 8th graders were assessed in math. One criterion was shared by all grades and subjects: effective interpersonal relations. Manatt & Daniels (1990) concluded that principals could accurately evaluate the performance of their teachers.

Wilkerson, Manatt, Rogers, and Maughan (2000) also used a criterion-referenced test to assess student's achievement (N = 35 teachers). Results found a significant correlation between a principal rating and the language arts test (r = .46), but no

significant relationships with math (r = .17) or reading (r = .09). Using the principal summative evaluations, they found significant correlations with math (r = .51) and language arts (r = .73), but not with reading (r = .34).

Jacob & Lefgren (2005) compared principal's ratings of teachers with student achievement gains based on a value-added measure (N = 202 teachers). The unadjusted correlation for reading was .20 and for math was .28. The principal's rating was not based on an overall rating of the teacher, but how well the teacher was at "raising student math (reading) achievement." Results showed that the principals were best able to identify those teachers who had students make small or large standardized achievement gains. They were less accurate rating those teachers in the middle of the distribution. In their study principal ratings were able to predict future student performance significantly better than when compared with using a teacher's experience, education, or level of compensation.

Heneman, Milanowski, Kimball & Odden (2006) used a similar approach to Jacob and Lefgren. They used an overall teacher evaluation rating and correlated it with average student achievement in reading and math. They used data from four different school districts over a three year time span. Each district varied somewhat in whom completed the teacher evaluation ratings. Most districts used the principal or assistant principal, but some also included peer raters or district personnel. Their results varied across the four different sites and were averaged across three years. In Cincinnati (N = 2500 teachers), results were higher with an average correlation of .35 in reading and .32 in math. Vaughn (N = 40 teachers) also showed higher correlations: .37 in reading and .26 in math. The other two districts had results closer to what other studies had found. Washoe (N = 3300 teachers) found an average correlation of .22 in reading and .21 in math, while Coventry (N = 475 teachers) was .23 for reading and .11 for math. The

authors concluded that the differences in the correlations could be attributed to the fact that some districts used multiple evaluators and gave their evaluators more training.

Holtzapple's (2003) study used some of the same data from the Cincinnati school district, but focused on earlier years. Correlations were calculated between composite evaluation ratings and the gain score residuals. The gain scores were calculated by taking the residual from the regression of a student's test score from one year (N = 166 teachers) on the test scores from the previous year (N = 80 teachers). Results were a bit higher than other studies have found and they had results for science and social studies as well. The following are the correlations they found for each of the two years: reading = .27 both years, math = .38 both years, science = .27 and .26, and social studies = .28 and .31.

Medley & Coker (1987) stated that there was no correlation between a principal's rating and student achievement. Looking at this group of studies conducted since then, there seems to be a fairly consistent pattern of modest correlations between a principal's ratings and student gain scores. The researchers in these studies stated that there was a relationship between principal's ratings and student achievement. Researchers have measured student achievement in different ways in the studies as well as have used different principal rating forms. Because of this, results have been different. A meta-analysis would be helpful to determine what the true correlation is between principal performance ratings and student achievement scores. Research has been carried out since the 1910s. No one has studied the research over time. A meta-analysis will combine all of these different years of data which will be instructive to help determine an answer as to what is the nature of this relationship. But as stated before, it won't be perfect. It will most likely be in the low to moderate range. Although low, a correlation of that size can still be practically significant.

Chapter III

As discussed in Chapter I, most theoretical work on teacher performance has focused more heavily on skill and knowledge determinants while minimizing or even ignoring possible motivational determinants. This chapter will present a review of personality predictors with the hope of improving our understanding of volitional choice behavior in teaching.

Personality Research

No attempt will be made to summarize all of the research surrounding the development of the field of personality psychology. It is vast. There are several strands of this research which are relevant for this dissertation which will be discussed.

Digman (1990) presented a short history of personality research. He declared that researchers had come to some agreement on five factors of personality. At that time, he thought there was more disagreement between researchers as to the meaning of each of the dimensions. The following labels were suggested for the five factors: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience. He stated that all of the work by Cattell, Guilford, and Eysenck fit into the five factor model. This model had been replicated using self-report inventories, ratings by others, and inventories in other languages. Digman (1990) expressed that the important lessons to take away from the research were that personality dimensions can be measured with reliability and validity and that they provide "a good answer to the question of personality structure (p. 436)." Once there was some level of agreement, personality researchers explored different directions. We look briefly at how researchers investigated broad personality dimensions and more specific dimensions of personality.

Higher order personality dimensions. In an effort to determine more about "why" personality develops, Digman (1997) explored the idea of higher-order factors of

personality. He reanalyzed correlation matrices from varying subjects and using various personality inventories. Using exploratory factor analysis, he found two factors. He labeled these factors alpha and beta. Factor alpha generally is comprised of Agreeablenness, Emotional Stability, and Conscientiousness. Factor beta is comprised of Extraversion and Openness to Experience or Intellect. He also performed confirmatory factor analyses and found that a two factor model had a good fit. He described Factor alpha as a social desirability factor. It contains those qualities that many might describe as socially desirable. Thus we might accurately describe it in lay terms as someone who has a "good personality." Hogan and Holland (2003) label Factor alpha as getting along and Factor beta as getting ahead.

No judgment is to be made as to the "correctness" of using two higher-order personality factors here. Researchers will continue the debate for a long time. This concept is useful for this dissertation. Many of the studies used in this meta-analysis are from the early years of research on teachers. Researchers in the 1940's and 50's had not heard of the Big Five. Thus the personality inventories they used didn't use more specific dimensions. Many of the inventories employed relied on an overall judgment of what a good personality was and were constructed by the researchers for their research. Since these studies are older, there is no way to see what most of these personality inventories looked like. Thus it would not be possible to attempt to classify which dimensions of personality the inventories were trying to measure. These early inventories might be good representations of what Digman (1997) labeled Factor alpha. In order to capture the information used in these studies, a global personality factor labeled, alpha, was used in the meta-analyses. It relies on the work by Digman (1997) that there might be the concept of a "good personality," but it is not considered a higher-order factor as described by him.

Facets of personality. Industrial-organizational psychologists interested in personality decided that research needed to be more specific. Personality was a better predictor of performance if it was tied more specifically to a criterion. Hough, Ones, and Viswesvaran (2001) explored the relationships between lower level facets of personality and more specific criteria. Hough (1992) found during Project A that if Extraversion was broken down into Affiliation and Potency and Conscientiousness was broken down into Dependability and Achievement, these scales showed different relationships with different criteria. If the larger variables Extraversion and Conscientiousness had been used, then these relationships may have been obscured.

Hough (1997) has advocated that the field develop a taxonomy for personality variables based on their nomological nets. She believes we should look at the relationships between personality variables and other criteria. Variables that show similar relationships with other variables should be placed in the same taxon or construct. She believes we need to break down the Big Five into smaller facets if we want to find better predictors.

Hough and Ones (2001) have developed a working taxonomy that researchers can use to begin to compile their data. They took most of the existing personality inventories and coded their scales according to these taxons. They had found that many researchers had similar items, but they were placed in different facets or a scale had items which were heterogeneous in content. They hope that these taxons will be used by researchers to determine the relationships that the taxons have with different criteria. From this information, the taxons can be merged or elaborated. Thus the taxonomy will be based on accumulated research with accompanying data to distinguish how variables are related to one another.

Research of personality and job performance. The early research conducted with personality inventories and job performance measures correlated all personality variables with all criteria that were available. Correlations between the two tended to be low. The conclusion that personality wasn't useful for predicting performance became the norm (Barrick, Mount, and Judge, 2001).

Many meta-analyses were performed using personality and work performance data. One of the first and most cited was done by Barrick and Mount (1991) using three different types of criteria (job proficiency, training proficiency, and personnel data) for five different job groups (professionals, police, managers, sales, and skilled/semi-skilled). Teachers were included in the professionals group, but professionals only made up five percent of the studies included in the meta-analysis. The other jobs that were also included in the professionals group were engineers, architects, attorneys, accountants, doctors, and ministers. Looking at this group of jobs together, one can be somewhat doubtful that they would require a similar personality profile to be successful. Results will be presented for Professionals, but their applicability to teachers as a group should not be weighted too heavily.

Barrick and Mount (1991) found Conscientiousness was consistently related to all three types of criteria for all five of the job groups. The other four personality dimensions had varying relationships depending upon the type of criteria and job. Due to these meta-analytic findings, researchers began to have more confidence that personality could be used as an effective predictor of work performance.

There are two sets of results from Barrick and Mount (1991) that are of interest to this dissertation. First are those results using subjective ratings collapsed across job groups: Extraversion $\rho = .14$, Emotional Stability $\rho = .09$, Agreeableness $\rho = .09$, Conscientiousness $\rho = .26$, and Openness to Experience $\rho = .04$. Since all of the studies

used in this dissertation will rely on subjective ratings, it is important to have some level of baseline that other researchers have found when using ratings as the criterion to compare personality against. Second, it is instructive to look at the results for the Professionals group. Results using Professionals collapsed across all three criterion types: Extraversion ρ = -.09, Emotional Stability ρ = -.13, Agreeableness ρ = .02, Conscientiousness ρ = .20, and Openness to Experience ρ = -.08. When an average of all the job groups collapsed across all three criterion types was calculated, the overall results were very similar to the ones found for ratings. The results are presented in the top of Table 5.

After comparing these two sets of results, one can see that except for Conscientiousness, Professionals have very different relationships with each of the other four personality dimensions than all of the groups combined. As stated previously, teachers don't fit well with many of the other occupations in this group, so I would suspect that the results from the meta-analysis completed in this dissertation will not agree with those of the Professional group. A much better result to compare the results from this meta-analysis will be those for all job groups combined. The job of teaching is different from many other types of jobs, so relationships between work performance and personality may exhibit unique characteristics from other professions.

Barrick, Mount, and Judge (2001) reviewed all of the meta-analyses of personality and performance that had been completed up to that time. A summary of the relevant results is presented in the lower portion of Table 5. The following are some of the conclusions they found about the Big Five personality dimensions and performance. Conscientiousness and Emotional Stability are positively correlated with job performance in most jobs. Conscientiousness has the stronger and more consistent relationship of the two. Both of these dimensions also show some relationship with teamwork and jobs that

have a large interpersonal component. Extraversion also tends to be important if the job involves a great deal of interaction with others. Agreeableness can be a good predictor of job performance for those jobs that rely on workers to cooperate, help, and nurture others. Openness to Experience has a smaller relationship with job performance and is more important if the criterion is training proficiency.

Teaching is a job that involves a great deal of interaction with students. In fact, there is only a small portion of a teacher's day spent without the students. Will the importance of Extraversion hold with teaching? No one can deny that there is a significant interpersonal component to teaching. Yet, all teachers are not extraverted. Barrick, Mount, and Judge (2001) also stated that Agreeableness was important for jobs that nurture others. Many would agree that an important component of a teacher's job is to nurture students. Once again, it will be interesting to see how the results of the meta-analysis in this dissertation compare.

Because Barrick, Mount, and Judge (2001) completed a somewhat definitive meta-analysis on the Big Five personality dimensions and job performance, they declared a moratorium on doing any more meta-analyses in this area and called for other types of research. The present meta-analytic research attempts to fill in some of the gaps in the previous research reported by Barrick, Mount, and Judge (2001). The first gap that this dissertation explores deals with the specific occupation studied: teachers. The teaching profession doesn't fit neatly in any of the occupational categories used by Barrick, Mount, and Judge (2001), even though Barrick and Mount (1991) included teachers in their professional category. A teacher is both a supervisor of his/her pupils and the pupils are customers that a teacher must satisfy. Because of the unique qualities of the teaching profession, it does seem instructive to complete a meta-analysis using this profession.

Second, Barrick, Mount, and Judge (2001) called for research using other means than self-report personality inventories to collect the personality data. They discuss how observer ratings may show a larger relationship to job performance than self-ratings. Hogan (2005) breaks down personality into two parts: identity and reputation. Self-report personality inventories tell us how we view ourselves, which is our identity. On the other hand, observers need to be relied upon to tell us what our reputation is. According to him, the Big Five model is a representation of a person's reputation. If we go a step further, he states that reputation describes our past behavior. There is no better predictor of future behavior than past behavior. Thus, using a person's reputation is a good predictor. So we should rely on observers to summarize a person's personality.

All of the meta-analyses discussed above rely on self-ratings for the personality data. There have been a few research studies which have explored the use of other ratings of personality and their relationship with ratings of job performance. Mount, Barrick, and Strauss (1994) studied the relationship between coworker and customer ratings of personality with supervisor ratings of performance with the job of sales representatives. Their results showed that other ratings were a more valid predictor of job performance that self-ratings.

Connelly and Ones (2010) conducted a meta-analysis which explored the relationship of other ratings of personality using the Big Five dimensions with job performance and their results can be seen in Table 6. They only used studies which had independent ratings of the other personality ratings and the performance ratings. Only a small number of studies was used and relied upon unknown occupations. Once again, the results using other personality ratings found greater relationships between the Big Five dimensions and performance than self-ratings. Connelly (2008) concluded that other ratings may be "more powerful in predicting job performance (p.145)."

This dissertation has a small subset of studies which relied on others to provide the personality ratings for the teachers. Thus, this research will provide more information about the relationship between other ratings of personality and ratings. It will be interesting to compare results using the teacher population in this study with the other occupations used by Connelly and Ones (2010). It is expected that the results for teachers will vary from those found with other occupations. This may be especially true for those Big Five dimensions which tend to vary more depending on the job. A comparison will also be able to be made between the other ratings of personality with the self-ratings of personality. Since all of the data will come from the same job of teachers, it will be informative to look at the similarities and differences in the relationships with the Big Five dimensions.

Chapter IV

Methodology

In order to find relevant data for the meta-analysis, several search techniques were employed. Searches were conducted of the databases PsychINFO (1887 – 2012), Google Scholar, ERIC (1966 – 2012), and *Dissertations Abstracts International* (1861 – 2012). The following keywords were searched: principal rating, teacher performance, personality, gain score, and value added. The abstract and title of each article were evaluated for relevance and those that appeared promising were downloaded. The reference lists of all these articles, dissertations, and technical reports were searched for any additional studies which might be related. If a related meta-analysis was located, then a search was conducted of the studies used in the meta-analysis. Each study was reviewed and coded by the author. The information recorded from each article included type of principal ratings, type of other rating, type of student achievement score, type of personality rating, correlation coefficient, and sample sizes.

In some cases, choices needed to be made as to what information to include from a study. If a study contained more than one useable correlation for one meta-analysis and they were not independent from one another, then those correlations were averaged and the mean correlation was used to represent the study. For example, if correlations between principal ratings and two different peers were presented, then the two correlations were averaged to obtain an estimate of the correlation between principal ratings and teacher peer ratings. There were some instances when more than one document contained the same data (e.g., a dissertation and a published study). In those cases, the larger or more complete data set was included in the meta-analysis.

Data from some studies was not used in the meta-analyses. Many of the studies, especially the value-added ones, used procedures which produced an unusable statistic,

such as a multiple regression. In these cases, if the statistic could not be converted to a usable one, the study was dropped from further consideration. If a study used only student teachers, then the study was not included. Some studies were dropped because the way the ratings scores were combined was suspect. For example, in one case actual scores were not used, but ranges. Some studies reported only statistically significant results without sufficient information to estimate the missing values. In those cases, the study was omitted.

Druva and Anderson (1983) conducted a meta-analysis that was focused solely on science teachers that was instrumental for the idea for including personality variables in this dissertation. Upon closer inspection of the studies included in that meta-analysis, it was discovered that none of the studies used the criterion of principal ratings. In the future, it would be interesting to replicate and add current studies to the work done by Druva and Anderson (1983). Since the criterion of interest in this dissertation is principal ratings of teacher performance, the studies used by Druva and Anderson (1983) were not included in the meta-analysis.

Data were aggregated using the Hunter and Schmidt (2004) psychometric meta-analytic method. The same set of statistics were calculated for each of the meta-analyses: the average sample size, weighted correlation across all studies (r_{obs}), the standard deviation of observed correlations (SD_{obs}), the standard deviation for the true validities (SD_{ρ}), and the 90% credibility interval. The observed standard deviation is the standard deviation of the correlations examined in each analysis. The standard deviation for true validities is the standard deviation of study effects subtracting the variability that would be expected due to sampling error and estimates the amount of true variability in study effects. It should be noted that other sources of variability can contribute to the estimates presented here including variability in the reliability of measures used across

studies. Insufficient information was presented to address these other sources of study design variability.

Chapter V

Results of Principal Ratings and Student Achievement

The final database for this section included 40 correlations based on 2,490 teachers from 28 independent samples. A meta-analysis of the relationships between principal ratings of teacher performance and multiple measures of student achievement was conducted. Student academic performance was measured in three different ways: value-added scores, student gain, and residual gain. All of these different measures of student performance were combined into one analysis first. Results are presented in Table 7. The correlations for student achievement (N = 2,490, k = 40) were similar to what has been found in the past (Medley and Coker, 1987) with a correlation of .17. The standard deviation of true correlations was small (SD_ρ = .02). Because of this, the 90% credibility interval is very small as well and does not contain zero (.15 to .19). When comparing these results with the meta-analytic results of the subjective versus objective performance data, the correlation is smaller than what Heneman (1986) and Bommer et al. (1995) found.

Separate meta-analyses were conducted by breaking down the data several different ways. First, meta-analyses were conducted based on the method used to calculate student achievement. A second way to segment the data was by the subject of the student achievement test. Finally there were several studies which used the same rating form to collect the principal ratings. These moderator analyses were conducted to evaluate the possible existence of method and content effects.

Method of calculating student achievement. Researchers employed different methods for determining the growth of student learning using student achievement test scores. Separate meta-analyses were conducted based on the method for calculating student achievement and those results are also presented in Table 7. Most of the older

studies used some form of a simple gain score or predicted gain score. Simple and predicting gains were grouped together for an overall meta-analysis (N = 1,667, k = 28). This group of studies produced a correlation of .15 which was slightly lower than the correlation found when combining all student gain scores. The standard deviation of observed correlations was .12, which is similar to what was found for the overall student gain meta-analysis. More recent studies relied on the newer statistical methodology of value-added scores which frequently control for more potential confounds. Most of this group of studies (N = 823, k = 12) was produced by a set of researchers working and publishing together. The meta-analysis of these data produced a correlation of .23 which is higher than that found for overall student gain. Its 90% credibility interval lower bound value was similar to the overall results (.16 to .30), but had a wider, but still small interval due to its greater standard deviation of true score validities of .05. It is worth noting that across studies the SD_{ρ} was consistently small suggesting fairly homogenous effects after subtracting variability that would be expected due to sampling error. This conclusion needs to be tempered by a relatively small number of studies for some analyses.

Finally there were a few studies which provided the correlations between the initial or final test score and the principal ratings. It is important to note that these correlations were not used in the overall student gain meta-analysis, but are included in the spirit of providing complete information. Looking at the results using the initial test score (N = 135, k = 2), the correlation was just slightly negative at -.02, with a standard deviation of observed correlations of .01. Using the final test score (N = 161, k = 3) produced slightly better results. The correlation calculated was .10. The standard deviation of observed correlations was similar in value (SD_{obs} = .13) to what was

calculated with the overall student gain and the separate gain and value-added score meta-analyses.

Subject test used. Results are presented in the lower portion of Table 7 broken down into the subject test used. There was enough data to conduct meta-analyses for three subjects. Arithmetic (N = 1,183, k = 20) has the largest correlation at .24. This correlation is higher than the overall correlation found across the various achievement tests combined. Its 90% credibility interval doesn't contain zero. Arithmetic has more variation than what was found in the overall results. Thus it has a greater standard deviation of true validities of .12, causing the credibility interval to be wider than the overall results (.10 to .40).

The correlation for reading (N = 1,674, k = 20) is .19. This value is closer to what was found for the overall results. Although based on a moderate number of studies, reading appears to have a smaller amount of variation than arithmetic. Its standard deviation of true validities was .06. The 90% credibility interval was narrower than the one found for mathematics and didn't contain zero (.12 to .26).

Finally a meta-analysis was conducted for language arts tests (N = 122, k = 3) which produced a correlation of .10. There were fewer studies used in this meta-analysis, which had a much higher level of variation. The standard deviation of true validities was .30. The 90% credibility interval was quite wide and contained zero (-.29 to .49).

Principal rating scale. The previous set of results dealt with differences in the student achievement scores. In this section, the focus shifts to the way that the principal ratings were collected. This subset of studies was analyzed because they relied on the same principal rating scales. Most of these studies were also conducted by a set of researchers working at the same university around the same time, so they calculated

their gain scores using the same principal rating scale. Results are presented in Table 8. Sample sizes and the number of studies weren't high, but there was enough data to calculate separate meta-analyses. Examples of each of the first three ratings scales can be obtained from the author. Since each of these rating scales collected the ratings in different ways, I wanted to see if the rating scale itself might make a difference in the relationship. For instance, one scale focused more on skills, one used personality descriptors, while another asked for the behaviors used to deal with certain situations. Despite content differences, the correlations with gains were similar across all of the different rating scales. The Michigan and the Torgerson Rating Scales (N = 168, k = 4) came from the same four studies and used the same student performance measures. The Michigan Teacher Rating Scale had a slightly higher correlation of .17. It had the smallest amount of variation with a corresponding standard deviation of true score correlations of .02. The 90% credibility interval was fairly small (.14 to .20) and didn't contain zero. A Master's thesis was also included in the results using the Almy-Sorenson Rating scale (N = 208, k = 5). The Almy-Sorenson Rating Scale for Teachers and the Torgerson Diagnostic Teacher Rating Scale had correlations just slightly smaller of .16. Both of these scales had the same standard deviation of true score correlations of .08 with very similar 90% credibility intervals which did not contain zero (.06 to .26 and .05 to .27, respectively). A different set of studies (N = 116, k = 6) used the Wisconsin M-Blank as its principal rating scale. The correlation for this group of studies was .16 as well. The standard deviation of observed correlations was .08, which was about half the size as that found for the other three rating scales.

Discussion

Medley and Coker (1987) posited that the true correlation between principal ratings and student achievement was near .20. For them, it solidified their conclusion

that a principal is not a “good judge of teacher performance” (p. 245). From their perspective, they were attempting to measure teacher performance in two ways: student achievement scores and principal ratings. Since these two measures did not have a very high correlation, there were a limited number of conclusions that could be drawn. Measuring teacher performance using student achievement scores was not accurate. Measuring teacher performance using principal ratings was not accurate. Both measures of teacher performance were wrong. Medley and Coker (1987) concluded that principal ratings were wrong and student achievement scores were “the” way to measure teacher performance.

The problem with this conclusion is that it assumes that there is only ONE way to measure performance and it can only be described in ONE way. We know from the discussion back in Chapter I that performance is not one thing (Campbell, 1990) and that measures of teacher performance use many different dimensions (Danielson, 1996). In this meta-analysis, the overall results show that the correlation between principal ratings and student gain is 0.17. Medley and Coker (1987) were correct in their estimation. However, it is a mistake to conclude that this correlation supports their conclusion. Student’s standardized test scores and principal ratings both measure important aspects of teacher performance. If nothing else, previous research has suggested the existence of task and contextual performance (Borman and Motowidlo, 1997). Research has shown that most job performance ratings are influenced by both task and contextual effectiveness. In this case, student achievement can reasonably be considered an aspect of task performance while principals are likely influenced, possibly greatly influenced, by contextual performance. It is likely that both are two pieces of the puzzle when trying to define what effective teacher performance is. If the relationship between

them is not high, it is premature to declare that one measurement is better than the other.

If we go a step further, we should not equate student test performance as the ultimate or sole objective or results-oriented measure of teacher performance. Looking at the previous meta-analyses by Heneman (1986) and Bommer et al. (1995), they found correlations of .27 and .389, respectively, between supervisory ratings and objective measures of performance. In this study, the correlation was .17. There was an even smaller relationship between the measures in this study. We may conclude that it is not appropriate to label student test performance as a results-oriented measure of teacher performance.

If principal ratings are not a results-oriented measure of teacher performance, then future research needs to explore what it is measuring. In order to develop a model of teacher performance, we need to better understand how to measure each of these criteria and what each contributes to our understanding of teacher performance. Student test performance data may help us measure the actual lessons that are taught in the classroom, which may translate into measuring task performance. As mentioned earlier, principal ratings may be assessing some portions of contextual performance. Finding the common as well as the unique variance each of these criteria add to our picture of teacher performance will begin to help us build this model.

There is some variation in the results of the meta-analyses after accounting for the measure of gain. The data show the lowest correlations with ratings when only an initial test score or final test score is used. These are not going to give the best results. Luckily, current practice uses more sophisticated measures of gain. There is a difference in results between using the older technique of calculating a gain or growth score as compared to the newer technique of value-added scores. Since the meta-analysis using

only value-added measures had a higher correlation than the one with only gain scores or both types of scores combined, there may be promise that this relationship is higher. At this point, there have not been enough studies done calculating teacher performance using the value-added technique and comparing them with principal ratings. Most of the studies used in this meta-analysis came from the same group of researchers. There are also variations in the model used to calculate the value-added scores. Taking these differences into account may also affect the results. Many school systems are currently calculating value-added scores for the teachers based on student test scores as well as having principals provide performance ratings. Future research needs to analyze the relationship between these two criteria to provide more information as to the extent of overlap between the two.

Since differences were found in the level of the relationship between principal ratings and student achievement scores based on the type of achievement test, we may want to conclude that certain subject areas may be better suited to providing this information. The meta-analysis using principal ratings and math tests produced the highest correlation of all the meta-analyses in this category. Since the content in math classes is more straightforward than other subjects, principals may be able to assess how well a teacher is doing more easily than with other subjects. Math tests may be considered more objective than other subject tests. Usually there is only one correct answer to a question, whereas a reading test answer may be open to more than one interpretation.

The correlation obtained in the meta-analysis with reading tests was at a slightly higher level than the overall student gain result. The variation in that group of studies was smaller than what was found for the math tests. The number of studies included in the language arts meta-analysis shows that more research needs to be done to obtain a

more definitive answer as to its relationship with principal ratings. Overall, the different results found for subject tests show promise that we may be able to increase this relationship for certain subject areas with further research. If that is the case, then it may give us a better understanding of what student achievement test scores are measuring.

The results based on which rating scale was used had correlations at almost the same value. Levels of variation were different. These were all older rating forms which are not in use today, so how much information this gives us is questionable. It would be helpful if research using more current rating forms could be conducted. School systems are all developing new rating forms to assess their teaching staff. Many are based on the same model of performance, such as Danielson's (1996). As school systems are becoming more standardized in their assessment and rating processes, it provides the opportunity to compare schools within a school system as well with each other. Once again, there is potential for future research.

Taken as a whole, the numbers show that there is a small amount of overlap between student test performance scores on one hand and principal ratings and teacher observations on the other. It is my position that school districts should stop making high-stakes decisions based exclusively on student test performance data. No matter how good the value-added models used are they still don't adequately represent the full picture of teacher performance.

Chapter VI

Multiple Rater Results

It was possible to compare ratings by principals with ratings made by several other rating sources that may have different perspectives or experiences with the teacher. Traditionally, raters include peers, supervisors, and subordinates (for managers). Because of the unique nature of the job of teacher, there are also students, parents, and classroom observers with perspectives on teacher performance. This leads to more groups of raters than what has been used in other meta-analyses of multi-source rater data using other occupations. Results comparing all rater sources are presented in Table 9. The final databased consisted of 210 correlations from 12,562 teachers from 48 independent samples.

Principal ratings. A meta-analysis was performed comparing principal ratings with peer teacher ratings (N = 660, k = 15). Similar to the results found by Harris & Schaubroeck (1988) and Conway & Huffcutt (1997), the correlation of .57 was the highest of all of those found using principal ratings. Its value was also closer to that found by Harris & Schaubroeck (1988). The standard deviation of true score correlations was high at .25, showing that there was considerable variation across studies. Thus the 90% credibility interval was wide (.25 to .89), but didn't contain zero. Some of the variability may be due to rating content and reliability. Studies did not provide sufficient information to examine these possible moderators systematically.

The meta-analysis between principal ratings and teacher self-ratings (N = 1,603, k = 24) produced one of the lowest correlations, r_{obs} = .15, using principal ratings. The standard deviation of observed correlations is smaller than the other observed standard deviations in this category. This correlation was lower than the results found by Harris & Schaubroeck (1988) and Conway & Huffcutt (1997) between supervisor and self-ratings,

although is consistent with these studies in that self-ratings are typically the most weakly correlated performance rating perspective.

Comparing principal ratings with student ratings (N = 1,783, k = 31) generated a more moderate correlation of .31. It has a smaller amount of variation than the peer ratings and the 90% credibility interval did not contain zero and was smaller than that found for the peer ratings. This meta-analysis found a larger relationship than Conway & Huffcutt (1997) using supervisor and subordinate ratings, although the alignment between students and work subordinates is clearly not perfect.

Parents are a distinct group of raters that doesn't have a comparable group in the multi-source rating data. The meta-analysis using principal ratings and parent ratings (N = 704, k =5) found a small correlation of .10, with a standard deviation of true score correlations of the same magnitude. The 90% credibility interval did contain zero (-.02 to .22), suggesting consistently small to zero relationships across settings, measures, and samples. This meta-analysis also had fewer studies. The importance of collecting ratings from parents about their children's teachers may be an underutilized resource.

In the research for this dissertation, it was discovered that there was a large group of others who provided teacher ratings: outside agencies, outside observers, investigators, other supervisors. An initial meta-analysis was conducted grouping all of these other observer's ratings together (N = 1,798, k = 35). If there was more than one of these groups in a particular study, then their ratings were averaged. The correlation for this meta-analysis was the second largest found using principal ratings, r_{obs} = .45. The standard deviation of true score correlations was .12 and the 90% credibility interval did not contain zero (.31 to .59).

In order to try and better understand the relationship that each of these other groups of raters had with principal ratings, where possible separate meta-analyses were

conducted breaking each of these groups down into three smaller sets. The first meta-analysis consisted of ratings made by those who were outside the school (N = 477, k = 8). All of the studies in this meta-analysis that met this criterion came from the 1940's and 50's. At that time, there were outside educational agencies that provided ratings for the teachers. In some instances, the study would only state that the ratings were collected, but no further information could be provided about the ratings. Other ratings were provided by Universities following up on their graduates. This meta-analysis had the smallest correlation of .36 for all of the meta-analyses using other ratings. This may be due to the fact that many of these raters may have only visited the classroom the one time that the rating was given. It is unknown how familiar the rater was with the teacher. It is assumed that these ratings were not shared with the principal. The standard deviation of true score correlations was higher than the overall other rater meta-analysis. It also had a wider 90% credibility interval.

The second meta-analysis consisted of raters which were in many cases part of the research team conducting the study (N = 1,219, k = 21). Ratings were provided by the investigator himself or by raters selected and trained by the investigator for the study. These raters visited a teacher's classroom various numbers of times using rating forms provided by the investigator. These ratings were taken by the investigator or given to the investigator without sharing them with the principal. Thus the assumption should be made that the principal had no knowledge of these ratings. The correlation found for this group, .46, was slightly larger than that found in the overall other meta-analysis. It did have a slightly smaller standard deviation of true score correlations and the 90% credibility interval was similar (.34 to .58).

The final group of raters for this set of meta-analyses consisted of those people who were considered to be higher level supervisors or coworkers of the teachers (N =

464, N = 13), who were assigned the task of observing classroom teaching and providing an evaluation. This group was comprised of superintendents, assistant principals, and supervising teachers. The meta-analysis examining this group of higher level supervisors had the highest correlation of .48. It had one of the smaller standard deviations of true score correlations of .10, which produced the smallest 90% credibility interval of .35 to .51. Obviously, this group has the most in common with the principals. There was no comparable data in the previous meta-analyses because Harris & Schaubroeck (1988) and Conway & Huffcutt (1997) only used ratings by the direct supervisor and not by any higher level supervisor.

The rest of the meta-analyses in this section do not deal with principal ratings. There was enough data to run meta-analyses comparing many of these rating groups with one another. These meta-analyses provide some comparisons for the results found by Harris & Schaubroeck (1988) and Conway & Huffcutt (1997). While in other cases, they provide a unique perspective for the job of teacher, due to the ratings by distinctive groups.

Student Ratings. The first group of meta-analyses compares ratings from students within a teacher's classroom with the other four groups: peer, self, other, and parent. First, the relationship between student ratings and peer ratings (N = 259, k = 8) are reviewed. The correlation is .39 which is similar, but slightly larger, than that found between principal ratings and student ratings. The standard deviation of true score correlations is .23, with a corresponding 90% credibility interval of .10 to .68. These values are also similar to the values found for principal ratings and student ratings. The correlation is slightly lower than the value found by Conway & Huffcutt (1997) for the relationship between subordinate ratings and peer ratings.

Next, student ratings were compared with the teacher's self-ratings (N = 546, k = 10). This meta-analysis had the smallest correlation using student ratings, r_{obs} = .13. The correlation found in this meta-analysis is almost identical to the one found by Conway & Huffcutt (1997) for subordinate and self-ratings. The standard deviation of true score correlations is .07 with a 90% credibility interval of .04 to .22.

The relationship between student ratings and other ratings (N = 698, k = 11) was explored in the next meta-analysis. The correlation for this relationship was moderate at .29, and was somewhat in the middle of all the meta-analyses involving student ratings. The standard deviation of true score correlations is .05 with a 90% credibility interval of .22 to .36.

There were enough studies that the other ratings could be broken down into two groups: outside agency and observer ratings. Some of these studies contained both types of ratings. The relationship between student ratings and outside agency ratings (N = 114, k = 3) had a correlation which was slightly smaller than what was found with all the other ratings combined. It had the smallest observed standard deviation of correlations. The meta-analysis between student ratings and observer ratings (N = 677, k = 10) produced a correlation that was the same magnitude as the one calculated for all the other ratings combined. The standard deviation of true score correlations was similar to the overall other ratings and the range of the 90% credibility interval was slightly wider.

Finally a meta-analysis between student ratings and parent ratings (N = 389, k = 3) was calculated. The correlation for this meta-analysis was the highest for all of the student ratings at .53. The standard deviation of true score correlations was .05, with a corresponding 90% credibility interval of .48 to .58. It makes sense that there would be a stronger relationship between student ratings and parent ratings. Parents most likely use

information given to them by their children in order to make ratings about their child's teachers, especially when the child is older and the parent doesn't visit the classroom often.

Self-Ratings. The next group of meta-analyses compares self-ratings with other types of ratings. First self-ratings are compared with peer ratings (N = 357, k = 4). The correlation is .13. The observed standard deviation of correlations was .10, which was the largest of the three in this group of meta-analyses. Conway & Huffcutt (1997) found a slightly larger relationship between peer with self (r = .19), while Harris & Schaubroeck (1988) found a higher correlation for self-peer (r = .36). Meta-analyses could also be conducted between self with other raters (N = 127, k = 3) and self with parent raters (N = 322, k = 2). There was variation between the correlations. The self-other meta-analysis found a relationship of .21, which was the largest of any of the self-ratings. On the other hand, the relationship between self-parent ratings was lower at .07. The observed standard deviation of correlations between self and parent ratings was the smallest at .02, while it was slightly higher for self and other ratings at .06.

Peer Ratings. Finally, the relationships between peer and other ratings (N = 83, k = 2) and peer and parent ratings (N = 282, k = 2) were explored even though the number of studies was small. The correlations are .28 and .16, respectively. The standard deviations of observed correlations for each study are .09 and .03, respectively.

Discussion

There were many meta-analyses calculated between all the different rating sources. The strongest relationship found was between principal and peer ratings. From this we might conclude that a principal and peer teachers are seeing similar behaviors in their day to day relationship with fellow teachers which they use to make their ratings. It

adds credibility to the conclusion that principal ratings do have value. This conclusion is furthered bolstered by the fact that there was a moderately large relationship between principal ratings and ratings made by those outside of the school. This group of raters ranged from higher level supervisors to investigators for a specific study. It would be difficult to assume that this group of raters had a similar agenda to the principal when making the ratings. The moderate degree of similarity shows that both groups are seeing commonalities in the teacher's performance. Finally, there is a moderate relationship between the principal's ratings and the student ratings. Since students spend the most time with the teacher, their perspective will be unique to others. Once again, though, it shows a degree of agreement with the principals. They are all seeing some of the same behaviors. This study shows that we have a convergence of observer ratings.

This research study indicates that we shouldn't discount principal ratings as being inaccurate. The levels of agreement found between the different groups of raters in this study are comparable to those found by other meta-analyses [Harris & Schaubroeck (1988) and Conway & Huffcutt (1997)] using a variety of occupations. Some results from this study were very similar to Harris & Schaubroeck (1988) and Conway & Huffcutt (1997), such as the relationship between principal and peer ratings. Other results found values which were higher or lower than what was found by Harris & Schaubroeck (1988) and Conway & Huffcutt (1997), but the relationships were in the same direction. For example, this study had a lower relationship between supervisor and self-ratings, but a higher one using supervisor and subordinate ratings.

Because of some of these differences, it might be helpful to separate out the different dimensions of teacher performance. Unfortunately, it was not possible to run analyses based on the different dimensions of performance in this study, but it could be a potential moderator in the future. As previously noted, research has shown that raters

rely on contextual and task performance dimensions when making their overall ratings (Borman, White, and Dorsey, 1995). Johnson (2001) proposed the question of how raters combine these different performance dimensions when making their overall performance ratings. Newer statistical methods may soon be available (Thomas, Zumbo, Kwan, and Schweitzer, 2014) to help us determine the breakdown of relative weights for different performance dimensions using principal ratings. Taking this a step further, determining the breakdown for all the other types of ratings would be very interesting. This would help us determine where the overlap in the ratings may be and where we may find more unique aspects of different group's ratings. Conway and Huffcutt (1997) discuss how subordinates may pay more attention to interpersonal behaviors, whereas peers don't place as much weight on those behaviors. If we had the weights different groups used in their ratings, we would be better able to determine if this was a true statement or not.

Conway and Huffcutt (1997) stated that task oriented behaviors can be seen more easily by supervisors and peers than by subordinates. This effect may be reversed with teachers. Students should be able to assess both task oriented and interpersonal behaviors. Once again, it would be interesting to do a breakdown of student's ratings to determine what proportion of task oriented behaviors and contextual behaviors students use to make their ratings.

Conway and Huffcutt (1997) explained how we don't want different rater sources to have too high of a relationship because the ratings may be redundant and won't add to our incremental validity. They also stated we should expect lower correlations between sources when the jobs are more complex. The relationships found between different sources of teacher ratings support both of those statements. There are relationships between the different sources of ratings, but they aren't so high that we

believe that they don't contribute unique portions of variance. Further, the job of teacher would most likely be rated on the higher end of the job complexity scale.

Future research should explore how each of the different rating groups adds to our knowledge of teacher performance. By studying the incremental validity that each group adds, we may be able to better develop a model of how these different ratings add together for teacher performance. As discussed above with the other criteria, how do different rating groups contribute information about task and contextual performance? With these types of ratings, we may be able to determine which type of rating we should collect to provide a certain type of information. For example, we may find that principals or peers contribute information about how the teacher performs tasks outside the classroom, but students can provide a fuller picture of what happens inside the classroom. Parents may help give information on criteria that have nothing to do with knowledge, such as higher motivation levels or more confidence in the student. Each type of rater contributes different information. As we add these pieces together, our model of teacher performance continues to expand.

In this research, the Principal – Parent relationship is low, but we shouldn't discount it. It offers a unique perspective. As suggested, it may help us assess more of the qualitative changes in a student's performance and behavior. Parents are those individuals who are most likely to see positive attitude or motivation changes. These are very difficult to assess using a standardized test. Since there wasn't much data, it was difficult to determine whether the age of the student may have an impact on the results. Parents may also be focused on some combination of their child's reports and reactions at home as well as their own experiences with the limited contact points with teachers (e.g., responding to questions, parent teacher meetings). Further research into this area

will help us establish exactly what the parent ratings are adding to our knowledge of teacher performance.

Table 10 presents all of the peer ratings compared to the other groups. We see a similar order of peer ratings as that found for the principal ratings. One main difference is that the student-peer relationship is greater than the peer-other relationship. The peer teachers have more in common with the students than with the outside raters. Since the outside rater group may include other supervisors, it makes sense that the principals would have a stronger relationship with this group. It is interesting that the peer teachers have a stronger relationship with the student ratings (r = .39) than the principals have with the student ratings (r = .32). Some of these peer teachers may have more contact with their fellow teachers than the principals. Another possibility is that the peer teachers may witness behaviors exhibited by their fellow teachers that are not observed by the principals.

Self-ratings didn't show a great amount of relationship with any of the groups as can be seen in Table 11 listing all of the self-ratings. Atwater, Ostroff, Yammarino, and Fleenor (1998) concluded that we shouldn't worry about the lack of agreement of self-ratings with other ratings. According to their research, the important variable to take into consideration is whether the self-ratings over or underestimate the other ratings. They found that how successful an employee is varies with whether the person over or underestimates his or her performance. Previous meta-analyses using other occupations found small relationships between self-ratings and other groups, but they were higher than what was found in this study. The only relationship that was of a comparable level was that found between subordinates and self (r = .14) and student ratings with self-ratings (r = .13). The interesting relationship in the group of self-ratings is the relationship with ratings given by outside raters. It had the highest correlation, but was based on a

small sample size and few studies. Further work in this area needs to be done to see if the effect holds up. If so, we need to do further exploration as to why raters outside the school are more likely to agree with a teachers' self-ratings, especially since they also exhibited good agreement with the principals. We need to better understand how this group of outside raters may bridge the gap between principal ratings and teacher self-ratings.

Next, we look at Table 12 to see the relationships between student raters and the other rating groups. There is a strong relationship between the ratings of students and parents. The nature of this relationship needs to be explored more. Parents may rely on information provided by their children to make the ratings. Both parent and student ratings may be impacted by how well the student is performing in the classroom. This may account for the stronger relationship. There is a similar level of relationship between student and peer ratings as principal and student ratings, although the peer group is higher. There may be some day to day behaviors that students and fellow teachers see that principals are unaware of. The relationship between student and other ratings is similar to the level of student and principal ratings which makes sense. The smallest relationship is between student and self-ratings. It is interesting that the two groups which spend the most time together do not have a higher level of relationship. In the future, it would be instructive to determine whether age of the student is a moderator in this relationship. Younger students spend almost their entire school day with the same teacher. Thus they would witness many more of the teacher's behaviors. On the other hand, older students may only spend approximately one hour with a teacher, so would have a smaller sampling of behaviors on which to base their ratings.

Finally in Table 13, we look at the relationships that the outside raters had with the other rater groups. When thinking about this group of raters, we might be skeptical

that ratings given by a person who observes a classroom once can give us more information than other rating groups. We may doubt how reliable or valid they may be. From looking at the level of the relationships with the other groups of raters we can see that it is possible for those ratings to have meaning. It could be possible that those ratings may be capturing unique variance from the other rating groups. This is an area where there is not a comparable group in the workplace. Employees may have others observe them while they work, but it is a very rare occurrence that those others would be asked to make ratings on their level of work performance. This is an area that needs to be studied more fully to determine exactly what those rating are capturing.

Chapter VII

Classification of Personality Ratings

Hough and Ones (2001) classified many of the popular personality inventories into one of the Big Five dimensions and corresponding facets. For each study used in this meta-analysis, the personality measure used by the researchers was coded based on the Hough and Ones (2001) taxonomy. If the inventory was included in the taxonomy, it was coded first in its Big Five dimension. Some measures only could be classified by a global Big Five dimension. If the measure did fall under one of the corresponding facets, then that information was coded as well. In all cases, the scoring direction of the scale was checked. If the scale was positive, then no changes were made. If the scale was negative, then the sign of the correlation from that study was reversed. Not all of the personality measures used in the meta-analyses were included in the Hough and Ones (2001) taxonomy. If the measure was not included, the author classified the personality measure in one of the Big Five dimensions, using any other relevant classification information available. If there was a corresponding facet that seemed appropriate, then the measure was categorized into that facet as well. As discussed in Chapter III, in some cases a personality inventory was used to assess overall positive pro-social personality, which has been labeled “Alpha” in this study. This Alpha is not considered to be a representation of the higher order factor labeled Alpha by Digman (1997). For these instances, no attempt was made to classify the inventory according to the Big Five dimensions or into facets. The correlation coefficient for the entire inventory was used as the value for the meta-analysis.

All of the personality studies were coded into one file. After that, all of the studies were divided into two files based on the source of the personality rating: self-rating or rating generated by someone else. This step was completed in order to determine

whether there was enough data to run meta-analyses separately based on rating source. There was a sub-group of studies that relied on other sources to generate the personality data. Most of these studies dated from the early years of personality research: three from the 1910's, three from the 1920's, two from the 1930's, two from the 1940's, one from the 1960's, and two from the 1990's. Even though there are only a small number of studies, this is a perspective that has rarely been explored in previous reviews of teacher effectiveness, but has been studied for other occupations (Connelly and Ones, 2010).

Results of Principal Ratings and "Other" Ratings of Personality

Other ratings of personality were generated by different sources. These sources have been classified into two separate groups. First are those personality ratings that were provided by raters who are different from those raters that supplied the performance ratings. The second group is composed of principals or supervisors who provided both the personality ratings as well as the performance ratings. Having the same raters make both sets of ratings may artificially inflate the results. Since rater was considered to be a moderator, separate meta-analyses were run for each of the two rater groups. Table 14 lists each of the studies used in the different rater other ratings meta-analyses, year published, who provided the ratings and how the ratings were coded according to the Big Five dimensions.

The final database for this section included 6,911 teachers with 41 correlations from 9 independent samples. For some of the studies using different raters, ratings were gathered during the teacher's senior year in high school or during their time in college. There were two separate groups providing ratings for Lins (1946) during the teacher's time as a student. First, the high school principal provided ratings about the teacher's personality when he or she attended high school. During college, two interviewers

provided personality ratings which were combined into overall ratings. Simun and Asher (1964) had faculty members from the teacher's senior year in college rate several areas on a five-point graphic scale. Jones (1923) collected personality ratings on a five-point scale from faculty members during the teacher's senior year in college the first year of his study. In subsequent years, he also collected ratings from prominent seniors on the same scales and combined them with the faculty ratings to provide overall personality ratings. Somers (1923) asked college teachers to make ratings about eight personality traits at the end of the teacher's first semester of college. Approximately three to seven ratings were collected and averaged to determine a composite score on the traits.

In the other three studies, other ratings of personality were given by those who were familiar with the teacher's performance or observed their classroom performance. Odenweller (1936) collected personality ratings from three other teachers in the school building who were familiar with the teacher. Shectman (1992) was able to obtain ratings during a short assessment center evaluation of the teachers while they were students at a teaching college. The ratings were provided by the assessors of the assessment center and fellow participants. Finally, Ambady and Rosenthal (1993) videotaped one class each of thirteen high school teachers with their permission. These videotapes were watched by eight college undergraduates who were paid to rate the teachers on the provided scales.

Table 15 shows those studies that used the same rater for the personality ratings and the performance ratings. Each dimension of personality rated is listed as well as how it was classified into its Big Five dimension. In some cases, more than one person could have provided each of the ratings, but no information was provided at the individual level as to who the specific rater was for each set of ratings. Due to this, it was assumed that the same rater made the ratings.

The results between other ratings of personality and principal ratings of teacher performance for both the same and different raters are presented in Table 16. Since the results using the same raters may be inflated, only the different rater results will be discussed here. First, the results for Alpha (N = 997, k = 8) showed the highest of any of the correlations calculated using different raters. There was more variation in these results than the Big Five dimensions as shown with its higher standard deviation of true validities of .16. The 90% credibility interval was wide, but its upper limit approached the range of the results using the same raters.

The rest of the results in this section are for the global Big Five dimensions. Conscientiousness (N = 216, k = 4) and Emotional Stability (N = 162, k = 3) have the same correlation which was the highest correlation of all the Big Five dimensions (r_{obs} = .23). The observed standard deviation for Conscientiousness fell in the middle of the range for the other Big Five dimensions at .09, while Emotional Stability showed slightly less variation.

Agreeableness (N = 162, k = 3) has a smaller relationship with other ratings than Conscientiousness and Emotional Stability, but was still noteworthy at .19. It did exhibit the largest observed standard deviation, but was the only Big Five dimension which a standard deviation of true validities and 90% credibility interval could be calculated. Its level of variation was less than what was found using the same raters.

There were only four studies with data for Openness to Experience, and only one of them used different raters. Thus, no meta-analysis could be calculated using different raters. The results for the one study which used different raters (Lins, 1946) reported a correlation of .17. We can use this one data point as our estimate. It is the same correlation as was found for Extraversion (N = 214, k = 4). The observed standard

deviation is .12 which is a greater amount of variation shown than some of the other Big Five dimensions.

Results of Principal Ratings and Self-Ratings of Personality

After filtering other ratings of personality, only self-ratings of personality were left. This data file was used for the next set of analyses. The final database for this section was comprised of 77 correlations based on data from 4,979 teachers from 22 independent samples. Certain self-ratings of personality were classified as Alpha, same as the analyses using other ratings of personality. Table 17 lists those inventories which were classified as Alpha and the corresponding studies which used them. Results for self-ratings of personality and principal ratings are shown in Tables 18 and 19. In only a few instances was the correlation substantially larger than zero and in many cases the estimate of the standard deviation of true validities or 90% credibility interval could not be calculated (due to zero or negative estimates for the variance attributable to sampling error). Most of the 90% credibility intervals contained zero showing that the results are not sizable and probably do not generalize.

In Table 18, the overall personality dimension of Alpha (N = 850, k = 12) once again had the largest relationship of any calculated using self-ratings with a correlation of .28. Its level of variation was similar, if not slightly higher, than what was found for the Big Five dimensions and its facets. The standard deviation of true validities was .08 with a 90% credibility interval of .18 to .38. There is a moderate relationship between how a teacher rates his/her overall personality and how the principal perceives the teacher performing on the job.

Since there were several studies that used the same personality inventory to collect their ratings, four separate meta-analyses were able to be calculated based on each of these inventories. The Minnesota Teacher Attitude Inventory (N = 543, k = 6)

has the largest correlation (r_{obs} = .35) in this group of meta-analyses. Its level of variation was smaller than the overall alpha results, so had a much narrower 90% credibility interval of .33 to .37. The Washburne Social Adjustment Inventory (N = 71, k =2) has the next largest correlation of .27, but had the largest standard deviation of observed correlations. The Morris Trait Index L (N = 96, k = 3) had a smaller relationship with a correlation of .07, but had the least amount of variation. Finally, a negative relationship was found for the Bell Adjustment Inventory (N = 73, k =2) with a correlation of -.13 and a standard deviation of observed correlations of .13. These results show that the personality inventory used is definitely a moderator.

The next result listed is for Intelligence (N = 179, k = 4). This dimension is not what we would normally label intelligence. On the 16 PF, Cattell had Factor B which was labeled as general intelligence. Three of the studies used in this analysis had correlations between the 16PF Factor B and principal ratings. The fourth study used Cattell's Creative Effort Test and a Two-Digit Numbers test. Since there was enough data to perform a meta-analysis, one was conducted and the results are presented here. The correlation was 0.18 showing a small, but positive relationship between intelligence, as measured by Cattell, and principal ratings. The standard deviation of true validities was 0.15 with a 90% credibility interval of -.01 to .37.

Results pertaining to the Big Five personality dimensions can be found in Table 19. At the bottom of Table 19, Agreeableness is listed. There are no corresponding results because there was no data for this dimension. Thus, data is only presented for the other four personality dimensions. As can be seen, there was enough data for some analyses to be run at the facet level. If there were at least two studies which had data at the facet level, then a meta-analysis was conducted. The global Big Five dimensions were those personality scales categorized by Hough and Ones (2001) as assessing the

overall personality dimension as well as any facet scales which couldn't be used for a facet level meta-analysis.

Conscientiousness is the Big Five dimension which generally has the largest relationship with job performance. That was not the case in this study. In the meta-analysis using global Conscientiousness (N = 258, k = 4), the correlation was .03 with a standard deviation of observed correlations of .11. Two facet level analyses could be conducted. The correlation using Cautiousness (N = 223, k = 3) was in effect zero with a value of .004. Achievement (N = 147, k = 2) had a more promising result with a correlation of .13, although Achievement showed more variation than Cautiousness. Unfortunately the 90% credibility interval contained zero. Since the credibility interval contains zero, it is difficult to determine whether these results would generalize. They are based on a few studies, so it is unclear how well the result would replicate.

Emotional Stability (N = 336, k = 6) had a slightly larger relationship with principal ratings with a correlation of .06. The first facet level meta-analysis using Optimism (N = 258, k = 4) was the only one that had a positive correlation which is .02. The other two facet level meta-analyses using Self-Esteem (N = 93, k= 3) and Low Anxiety (N = 168, k =2) produced the same correlation of -.06. There were differences in the standard deviation of observed correlations. Self-Esteem has the highest level of variation, while Low Anxiety had the smallest and Optimism fell in between.

The global Big Five dimension with the largest correlation belonged to Openness to Experience (N = 467, k = 5) and has a negative value, r_{obs} = -.11. Its standard deviation of true validity correlations is .07. The 90% credibility interval has an upper value that was right at zero which spanned down to –.20. Even though the results were a bit promising, they may not generalize. A meta-analysis for the facet Traditionalism (N =

258, k = 4) was able to be conducted. The correlation is slightly negative, but would be considered zero with a value of -.004 and a 90% credibility interval that includes zero.

Finally, Extraversion (N = 496, k = 4) also has a negative correlation of -.03, but has the smallest amount of variation of all the global Big Five dimensions. There were many facet level meta-analyses for this Big Five dimension. Activity/Energy Level (N = 269, k = 4) is the only one that a 90% credibility interval (-.09 to .25) and standard deviation of true validity correlations (SD_ρ = .13) could be calculated. It has one of the highest correlations at .08. Dominance (N = 350, k = 7) has the highest correlation of .09, but has less variation than Activity/Energy Level. Warmth (N = 258, k = 4) and Reflective (N = 97, k = 2) had the same correlation of .05, with similar standard deviation of observed correlations of .06 and .04, respectively. Sociability (N = 172, k = 5) has the smallest correlation of .02 with the largest standard deviation of observed correlations of .15. The only facet with a negative correlation is Autonomy (N = 97, k = 2), but it has the smallest amount of variation of all the facets of Extraversion.

Discussion

Principal ratings and other ratings of personality. The Alpha dimension using different raters has the largest relationship with principal ratings. Its correlation was much greater than those found for the Big Five dimensions. It is unclear why this rating of pro-social personality is doing a better job of predicting who will perform well on the job. Not much information is given about what is being rated as Alpha in the original studies. Many of them are just labeled personality. Since this term can be widely construed by different people, there is no way to equate these terms even among the raters within the same study. The best we may be able to do is state that Alpha in these studies is an overall general impression about the teacher. Thus we can conclude that those teachers who are thought of as having a “good personality” are more successful

on the job than those who are not. More research will need to be done to determine what this result means.

There are two comparisons groups for the other ratings of personality using the Big Five dimensions in this research. The most similar is the research done by Connelly and Ones (2010) who also used other ratings of personality. They likewise compared their results with those found by Barrick, Mount, and Judge (2001). The best course is to compare the current meta-analytic results found in this study with the results found by both of the above researchers. It makes most sense to use the values found when having different raters provide the two sets of ratings, especially since Connelly and Ones (2010) didn't use any studies which used the same groups to provide both sets of ratings.

The results found in this study are very similar to what Connelly and Ones (2010) found between ratings and job performance. Conscientiousness had the largest relationship and Extraversion the smallest which is the same pattern for the current results. The ranking of the middle three found by Connelly and Ones (2010) varied slightly from the results in this meta-analysis.

Since no corrections were made to the data for these analyses, these results will be compared to the mean correlation found by Connelly and Ones (2010). The same correlation of .23 was found for Conscientiousness in this study as well as by Connelly and Ones (2010). Teachers are similar to other occupations for how well Conscientiousness can be used to predict job performance. In this study, the same correlation of .23 was also found for Emotional Stability. This was higher that what Connelly and Ones (2010) found ($r = .14$). This may indicate that the emotional well-being of a teacher may be an important factor in how well the teacher is able to perform.

Having a separate group provide those personality ratings may offer us more information as to the well-being of the teacher than what the teacher provides himself/herself.

Since there was not enough data to provide any results for Openness to Experience, no comparison can be made. Connelly and Ones (2010) found that it was the Big Five dimension that had the second highest relationship with job performance. In this study, the correlation for Agreeableness is .19, which is slightly higher than the correlation of .13 found by Connelly and Ones (2010). It does have the same relative position in comparison to the other Big Five dimensions in both studies. The value of the correlation for Extraversion of .17 is almost twice what was found by Connelly and Ones (2010) at .08. Both studies found that this was the Big Five dimension that had the smallest relationship with job performance. Since the occupations studied by Connelly and Ones (2010) are unknown, it is difficult to know what this means for their data. In this study, it shows that how outgoing a teacher happens to be is not the best indicator of that teacher's job performance.

Next will be the comparison of the rank order of these results with those found in other occupations. Conscientiousness had the largest relationship with principal ratings. This finding is typical in many other occupations. Barrick, Mount, and Judge (2001) found that most previous meta-analyses had their strongest relationship between conscientiousness and performance. It looks like the same can be said for the occupation of teachers. As described by past researchers, it makes sense that those workers who are rated as working hard are also the ones rated by their supervisors as performing well on the job. Thus we can conclude that teachers are similar to other occupations in this respect.

The second highest relationship was found between emotional stability and principal ratings. Once again, this replicates the findings of other meta-analyses using

other occupations. Others (Barrick, Mount, and Judge, 2001) have suggested that it makes sense that those workers who are more emotionally stable make better performers. Teachers seem to fall in that category as well.

Where these results begin to differ with other occupations is the ordering of the next three dimensions. Typically with other occupations, extraversion would rank third. In this case, extraversion has the smallest relationship with principal ratings. For teachers, agreeableness has the third largest relationships with principal ratings. Based on the correlation found in one study, we may conclude that Openness to Experience had the fourth largest relationship with principal ratings. Thus the rank order of the relationships between the Big Five dimensions and other ratings of personality does differ from other occupations. Since the number of studies in this research was small, it would be helpful to have more data for each of the Big Five dimensions in the future for teachers. Thus we would be better able to determine how robust these results are and how teachers compare to other occupations.

The results using the same raters for both the personality ratings and the performance ratings can be consulted to provide more information on the rank ordering of the dimensions. It is difficult to know how much confidence to put in the results using the same raters for both sets of data. The correlations are much higher when the same raters made both ratings as compared to having different raters make each set of ratings. The rank order of the relationships between job performance and the Big Five dimensions is exactly the same as when using different raters. Thus we may conclude that there is some level of agreement among all groups of other raters when using personality predictors for teacher performance. It helps to bolster our confidence somewhat in the results found using different raters. This is helpful information because

it shows that the teacher population may differ from other occupations in the importance of different personality dimensions in predicting performance.

In general, the results of this study found a greater degree of relationship between other ratings of personality and job performance than what has been found in the past using self-ratings. As concluded by Connelly and Ones (2010), this result points to the fact that there is value in collecting other ratings of personality. They may provide more information or possibly different information than self-ratings.

In the future, it may be beneficial to separate the results based on the type of teacher (elementary, middle school, high school) or possibly even by subject taught. Because of the small number of studies, these divisions could not be made in the current study. These further delineations could help researchers draw conclusions about whether results could be generalized across the occupation of teachers or whether effects might be more specific to a certain type of teacher. Whether the levels of these relationships replicate in future research is a question yet to be answered.

Principal ratings and self-ratings of personality. The results exploring the relationship between principal ratings and self-ratings of personality do not look very promising. Similar to the results using other ratings of personality, the largest correlation was between the Alpha dimension and principal ratings. Thus we may be able to draw the conclusion that those teachers who rate themselves as having positive personality traits are also rated by their principals as performing well on the job. There may be some usefulness for an overall personality trait.

One specific inventory, the Minnesota Teacher Attitude Inventory, had an even greater relationship with principal ratings with a correlation of .35. Most of this group of studies was conducted by the same researcher with the same population over time which may account for the consistency of the results. This inventory deals more

specifically with a person's attitudes towards students. Thus it makes sense why it would be a good predictor of how a teacher performs on the job. On the other hand, with this logic we would expect the Morris Trait Index L to also show a strong relationship with principal ratings since it deals with how a teacher views and behaves in different classroom situations. It exhibited a much smaller relationship with principal ratings with a correlation of .07. The Washburne Social Adjustment Inventory also showed a stronger relationship with principal ratings with a correlation of .27. This inventory asks questions about how a person behaves or reacts to personal situations. Why this inventory has a larger relationship is unknown. The sample sizes and the number of studies for most of these analyses are small. Thus it is difficult to determine how robust these relationships are or whether the effects would replicate. More research needs to be done to determine why this result is found.

There is not a strong correlation using Cattell's idea of intelligence from the 16PF. It was interesting to run the analysis to see what the results would tell us. It did have a fairly large correlation in comparison to the ones found with the Big Five dimensions. Previous research has established the B scale from the 16 PF as a reasonable measure of general cognitive ability. The finding that the reasoning scale is positively but weakly correlated with teacher effectiveness is consistent with a large body of research showing similar relationships for other cognitive ability measures.

With all of the personality research that has been done more recently, it was disappointing there was not more data to analyze. The sample sizes and number of studies was often even smaller using self-ratings than what was available for other ratings. Even the usual stalwart of Conscientiousness didn't exhibit much of a relationship with principal ratings. The one global dimension that was most interesting was the negative relationship Openness to Experience had with principal ratings,

although again the results were based on small amounts of evidence. Improved prediction of teacher performance may require the construction of personality scales that are both tailored to the job of teachers and factorially complex. That is, obtaining even reasonable prediction of the personality of effective teachers may require the creation of compound trait measures based on multiple facets of the Big Five. Since Openness to Experience is usually linked to the idea of creativeness, it is interesting that those teachers who rated themselves higher on Openness to Experience were rated as performing more poorly on the job. More research needs to be done to explore why Openness to Experience has a negative relationship with job performance. This is an area where it may be important to know what level the teacher is teaching or what the subject matter of the class is. One might hypothesize that this effect would not hold up if the teacher were teaching classes such as art or band. More data would need to be collected in order to make more sense of this result.

There is some evidence to support Hough and Ones (2001) theory that personality facets exhibit different relationships with a criterion than other facets or the global personality dimension. A good example is Conscientiousness with its two facets: Cautiousness and Achievement. Conscientiousness had a small correlation and Cautiousness was in effect zero. Achievement had the largest correlation of all the meta-analyses using the Big Five dimensions and facets. Although it was only based on two studies, it is as promising as any other traditional personality measure. Another good example deals with Emotional Stability and its facets. Emotional Stability and Optimism have positive relationships with principal ratings while Self-Esteem and Low Anxiety have negative relationships with principal ratings.

Hough and Ones (2001) have suggested that using the facet of Achievement may help us predict better than the global dimension of Conscientiousness. With the

profession of teachers, this may be true. The effects of Cautiousness and Achievement may be balancing each other out. We need to explore whether those teachers who are higher achievers are also the ones who are performing better on the job. Two other facets that may be helpful predicting strong teaching performance are Activity/Energy Level and Dominance. Their correlations with principal ratings (r = .08 and .09, respectively) were some of the higher values found in this study. Instead of asking whether a teacher is outgoing, a better question may be to ask whether the teacher has a high level of energy or is dominant in the classroom. These questions may help us determine which qualities a teacher may possess which are more predictive of higher levels of performance. More research needs to be done.

Comparison of results using other ratings of personality and self-ratings of personality. The most consistent effect across both types of ratings was that the Alpha concept had a moderate relationship with principal ratings. Collecting overall ratings of a teacher's personality by the teacher or by others can tell us something about performance. This collection of scales, as displayed in Tables 14-15 and 17, is admittedly something of a mess of ideas. However, among the personality scales (with the exception of achievement) it is the most promising lead based on this review of research going back almost 100 years. The goal of future research should be establishing what precisely in the Alpha category appears to work. How much this result gains for us in the long run is yet to be determined.

The rank order of the relationships was different for both types of ratings with the Big Five dimensions. It may be instructive in the future to collect both types of personality ratings. By understanding the relationships between each type of personality rating and the Big Five dimensions, we may understand teacher performance better. Thus, we may rely on different types of ratings to provide different information. This was

especially true for Openness to Experience. Other ratings had a positive relationship with teacher performance, while there was a negative relationship with self-ratings. It would be instructive to determine what the differences are that other raters are seeing as opposed to self-ratings. This dimension may be helpful, but how it helps us in our prediction of teacher performance may depend heavily on which type of ratings it is based upon.

There has not been much research collecting personality ratings of teachers by their students. There were no studies found for this meta-analysis which had student personality ratings and principal performance ratings. It would be interesting to collect personality ratings from students and correlate them with principal ratings. We would most likely have to rely on older students for this type of study. Since students spend so much time with their teachers, they would be in a good position to make ratings about personality. Students may see things that a teacher may not realize that he or she is doing. Once again it may provide us a new perspective on teacher performance that hasn't been explored previously.

Overall, the results from both sets of personality data contribute to our understanding of using personality inventories to predict teacher performance. Personality data should be collected when trying to predict teacher performance. Now we need to answer questions about who we should be collecting the personality ratings from to gather what specific information. There is still much that needs to be discovered.

Chapter VIII

Limitations

Two concerns cover all of the results in this study. The first is limited information about what the rating measures truly capture. For example, are principal ratings and peer ratings focused on the same aspects of behavior? Related to this is the second issue which is that most of the measures used are likely deficient measures of the multi-dimensional domain that is teacher performance (as shown in Table 2). This is to be expected but is important to keep in mind while examining the results. Critically, we are limited in what standardized achievement tests can assess. There are many teachers who teach subjects that are not assessed by these tests. Thus we have a gap in the data that is gathered to assess certain kinds of teachers. We need to rely on other criteria to assess these teachers. Since the value-added data in this study only came from teachers who taught those subjects assessed by standardized achievement tests, it is unclear how these results may generalize to the whole teacher population.

At this point in time, it is unclear which way the causation path flows between teacher performance and student achievement scores. Some may argue that better performing teachers may work with higher performing students causing the correlation to be high between teacher performance and student achievement tests. Until a more definitive answer as to which way the causation flows between these criteria, caution is warranted.

When analyzing a research base which has spanned over one hundred years, one would expect to have a multitude of studies. In the area of teacher performance, finding many useable data sets was not the case. There were many more studies of teacher performance that were unable to be used in this meta-analysis. Unfortunately, much of the data was lost or underutilized because researchers didn't present complete

data when publishing their results. There were a small group of studies that may have been useable that were never published in a journal. It proved too difficult to find a copy of these works which could be analyzed. As stated earlier, some data was presented in a manner that could not be translated into a useable statistic for the meta-analysis. With the prevalence of meta-analyses these days, we can be hopeful that future researchers will report their data in a manner that will be more functional for these types of analyses. Because of the small number of studies, there was not enough data to assess many interesting moderators. This small number of studies also makes the results less robust.

Future Research

There are many areas in which future research needs to be conducted. Many of these have been discussed throughout the dissertation. The following are a few ideas for future research.

We need to determine which dimensions of performance principals are using to make their ratings. In order to do this, we need to use rating scales which are multi-dimensional, instead of relying on an overall rating of performance. Using a rating scale which assesses both task and contextual components of performance would be a good place to start. Future researchers could analyze principal ratings and determine how much of their variance can be attributed to task and contextual performance. The same process could be conducted for other types of performance ratings as well: student, peer, parent, and observer. Once we have the breakdowns of what dimensions make up the ratings, then we can do comparisons of these breakdowns between the different rating sources. This research would begin to help us establish where the overlap is between sources of ratings as well as their unique contributions.

After we have task and contextual performance ratings by principal and other raters, we could explore their relationships with student achievement scores. At this

point, only an overall principal rating has been compared to value-added scores. Future research needs to compare value-added scores to dimensional ratings of performance, assessing task and contextual performance. Once we can look at these relationships, we may be able to better determine exactly what value-added scores are assessing. I would hypothesize that the value-added scores would exhibit a higher level of relationship with task dimensions as compared with contextual dimensions. Similar work could be done to compare value-added scores to other ratings of performance. Important moderators such as grade level and subject taught may also impact the results.

We know that there is a high degree of relationship between student and parent ratings. More research needs to be done to determine whether parents are relying on their own interactions and observations of the teacher or are influenced by their child's opinion. Ratings from both the students and the parents could be collected over the course of the school year. Comparisons of these ratings over time could be done to see what the levels of congruence are between them. Once again, using a rating scale that is multi-dimensional would be useful. Scales could be constructed that more specifically assess any interactions that the parent may have had over time with the teacher, such as conferences. Thus comparisons between the two sets of ratings could be done across dimensions. It could be that there are areas that the parent may rely more heavily on the opinions of their children. Age of the child will most likely be a moderator in this study. I would predict that the older the student is, then the more the parent will rely on the opinion of the student.

Future research should try and determine why those personality inventories which were labeled alpha in this study have a larger relationship with principal ratings. The first step may be to collect both self-ratings and other ratings on several different

measures of personality. Information comparing ratings using these inventories of alpha and more modern personality inventories may help us better understand what these alpha inventories are assessing. Once again, we should collect principal ratings of teacher performance using multi-dimensional scales. We could compare these more specific ratings of teacher performance with the alpha personality inventories to see if the relationships differ. Most likely, the contextual dimensions of performance would show a greater relationship with the personality inventories than the task ratings of performance.

We should continue to use personality inventories to assess teachers. Ratings should be generated by both the teachers themselves and using other raters who are familiar with the teachers. Finding new sources of these other ratings may provide new information that has not been obtained before. Determining which dimensions of personality are best measured using which type of rater will go far in helping us predict the successful performance of teachers.

Conclusion

After analyzing the past century of research, I conclude that principal ratings are a legitimate measure of teacher performance. With the results of this research, we are better equipped to develop a more comprehensive model of teacher performance. This data helps us to understand the relationships between several criteria (student test performance and multiple-rater ratings), as well as the relationship between principal ratings and personality predictors.

References

(1) Denotes student score improvement metrics meta-analysis

(2) Denotes multiple rater sources meta-analysis

(3) Denotes other ratings of personality meta-analysis

(4) Denotes self-ratings of personality meta-analysis

Ahmadnia, H. (2006). The relationship between teachers' performance ratings and the achievement of their students. *Applied H.R.M. Research, 11(1)*, 75-78. (1)

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluation from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64(3)*, 431-441. (3)

Anderson, H. M. (1954). A study of certain criteria of teaching effectiveness. *Journal of Experimental Education, 23*, 41-71. (1)(2)

Atwater, L. E., Ostroff, C., Yammarino, F. J., and Fleenor, J. W. (1998). Self-other agreement: Does it really matter? *Personnel Psychology, 51*, 577-598.

Bach, J. O. (1952). Practice teaching success in relation to other measures of teaching ability. *The Journal of Experimental Education, 21(1)*, 57-80. (2)

Baird, J., & Bates, G. (1929). The Basis of Teacher Rating. *Educational Administration and Supervision*, *15*, 175-83. (1) (3)

Ball, M. A. (1984). Relationships between ratings of teachers made by primary pupils and professional educators (Doctoral dissertation, University of Georgia). (2)

Barr, A. S., Torgerson, T. L., Johnson, C. E., Lyon, V. E., & Walvoord, A. C. (1935). The validity of certain instruments employed in the measurement of teaching ability. *The measurement of teaching efficiency*, 73-141. (1)

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

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: what do we know and where do we go next? *International Journal of Selection and Assessment*,*9*(1-2), 9-30.

Berk, R. A. (1988). Fifty reasons why student achievement gain does not mean teacher effectiveness. *Journal of Personnel Evaluation in Education, 1(4)*, 345-363.

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work.* Boston, MA: Kent Publishing Company.

Boardman, C. W. (1928). *Professional tests as measures of teaching efficiency in high school* (No. 327). Teachers college, Columbia University. (2)

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & Mackenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48*, 587-605.

Borman, W. C. (1974). The rating of individuals in organizations. An alternate approach. *Organizational Behavior and Human Performance, 12(1)*, 105-124.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In Schmitt, N., and Borman, W. C. (Eds.), *Personnel Selection in Organizations* (pp. 71-98). San Francisco, CA: Jossey-Bass.

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10(2)*, 99-109.

Borman, W. C., White, L. A., and Dorsey, D. W. (1995). Effects of rate task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology, 80(1)*, 168-177.

Boyce, A. C. (1912). Qualities of merit in secondary school teachers. *Journal of Educational Psychology*, *3(3)*, 144-157. (3)

Bradley, J. H. (1918). A study of the relative importance of the qualities of a teacher and her teaching in their relation to general merit. *Educational Administration and Supervision*, *4,* 358-363. (3)

Brandt, W. J. (1949). A follow-up of some earlier Wisconsin studies of teaching ability. *The Journal of Experimental Education, 18(1)*, 1-29. (1) (2)

Brauchle, P. E., McLarty, J. R., and Parker, J. (1989). A portfolio approach to using student performance data to measure teacher effectiveness. *Journal of Personnel Evaluation in Education, 3*, 17-30.

Brookover, W. B. (1940). Person-person interaction between teachers and pupils and teaching effectiveness. *The Journal of Educational Research, 34(4)*, 272-287. (2)

Bryan, R.C. (1937). *Pupil rating of secondary-school teachers* (No, 708). Teachers College, Columbia University. (2)

Burry, J. A., & Shaw, D. G. (1988). Teachers and administrators differ in assessing teacher effectiveness. *Journal of Personnel Evaluation in Education*, *2*(1), 33-41. (2)

Callis, R., Brown, K. B., Burgess, T. C. (1952). *Studies on the Effectiveness of Teaching*. University of Missouri – Columbia Agricultural Experiment Station. (4)

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (Vol. 1, 2nd ed.). Palo Alto, CA: Consulting Psychologists Press.

Campbell, J. P. (2013). Assessment in industrial and organizational psychology: An overview. In Geisinger, Kurt F. (Ed); Bracken, Bruce A. (Ed); Carlson, Janet F.

(Ed); Hansen, Jo-Ida C. (Ed); Kuncel, Nathan R. (Ed); Reise, Steven P. (Ed); Rodriguez, Michael C. (Ed), (2013). APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology. APA handbooks in psychology., (pp. 355-395). Washington, DC, US: American Psychological Association.

Campbell, J. P., Gasser, M. B., and Oswald, F. L. (1996). The substantive nature of job performance variability. In. K. R. Murphy (Ed.), *Individual Differences and Behavior in Organizations.* San Francisco, CA: Jossey-Bass.

Coleman, V. I., & Borman, W. C. (2000). Investigating the underlying structure of the citizenship performance domain. *Human Resource Management Review, 10(1)*, 25-44.

Connelly, B. S. (2008). The reliability, convergence, and predictive validity of personality ratings: An other perspective (Doctoral Dissertation, University of Minnesota).

Connelly, B. S., and Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136(6)*, 1092-1122.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10(4)*, 331-360.

Cook, W. W., & Leeds, C. H. (1947). Measuring the teaching personality. *Educational and Psychological Measurement*. (2) (4)

Cooke, D. H. (1937). How do teachers rate themselves? *Educational Administration and Supervision,23*, 473-476. (2)

Copas, E. M. (1984). Critical requirements for cooperating teachers. *Journal of Teacher Education, 35(6)*, 49-54.

Coxe, W. W., and Cornell, E. L. (1933). The Prognosis of Teaching Ability of Students in New York State Normal Schools. *University of the State of New York Bulletin*, *December*, No. 1033. (2)

Crabbs, L. M. (1925). *Measuring efficiency in supervision and teaching* (No. 175). Teachers college, Columbia University. (1)

Cutchin, G. C. (1998). Relationships between the big five personality factors and performance criteria for in-service high-school teachers (Doctoral dissertation, Purdue University). (2) (4)

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Davenport, K. (1944). An investigation into pupil rating of certain teaching practices. *Purdue University, Studies in Higher Education*, No, 49. (2)

Day, H. P. (1959). A study of predictive validity of the Minnesota Teacher Attitude Inventory. *The Journal of Educational Research, 53(1)*, 38-38. (4)

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417-440.

Digman, J. M. (1997). Higher-order factors of the big five. *Journal of Personality and Social Psychology, 73(6)*, 1246-1256.

Druva, C. A., and Anderson, R. D. (1983). Science teacher characteristics by teacher behavior and by student outcome: A meta-analysis of research. *Journal of Research in Science Teaching, 20(5)*, 467-479.

DuBois, C. L. Z., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and white-black differences. *Journal of Applied Psychology, 78(2)*, 205-211.

Epstein, J. L. (1985). A question of merit: Principals' and parents' evaluations of teachers. *Educational Researcher, 14(7)*, 3-10. (2)

Erickson, H. E. (1954). A factorial study of teaching ability. *The Journal of Experimental Education, 23(1)*, 1-39. (2) (4)

Farbstein, M. (1965). Critical requirements for cooperating teachers: A study of cooperating teachers as perceived by student teachers in the State of New Jersey. (Doctoral Dissertation, Rutgers University).

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. *The High School Journal, 75(3)*, 168-178.

Frutchey, F. P. (1953) Differential characteristics of the more effective and less effective teachers. Washington, D. C.: U. S. Department of Agriculture, Extension Service.

Gallagher, H. A. (2004). Vaughn Elementary's Innovative Teacher Evaluation System: Are Teacher Evaluation Scores Related to Growth in Student Achievement?. *Peabody Journal of Education*, *79*(4), 79-107. (1)

Geithman, B. W. (2009). Examining principal perceptions, and teacher and school effectiveness through a value-added accountability model (Doctoral Dissertation, University of Southern California). (1)

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis.* National Comprehensive Center for Teacher Quality.

Gotham, R. E. (1945). Personality and teaching efficiency. *Journal of Experimental Education, 14*, 157-165. (1) (4)

Gough, H.G., Durflinger, G.W., & Hill, R.E. (1968). Predicting performance in student teaching from the California Psychological Inventory, *Journal of Educational Psychology, 59(2),* 119-127.

Gowan, J. C. (1955). Prediction of teaching success: Rating of authority figures. *California Journal of Educational Research*, *6(4)*, 147-52. (2)

Hamrin, S. A. (1927). A comparative study of ratings of teachers-in-training and teachers-in-service. *The Elementary School Journal*, *28*(1), 39-44. (2)

Hanushek, E. A., and Rivkin, S. G. (2006). Teacher quality. *Handbook of the Economics of Education, 2*, 1051-1078.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*, 43-62.

Hellfritsch, A. G. (1945). A factor analysis of teacher abilities. *Journal of Experimental Education, 14*, 166-199.

Heneman, H. G., Milanowski, A., Kimball, S. M., & Odden, A. (2006). Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay. Consortium for Policy Research in Education. (1)

Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology, 39*, 811-826.

Hill, C. W. (1921). The efficiency ratings of teachers. *Elementary School Journal, 21*, 438-443. (1)

Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63(1)*, 119-151.

Hogan, J., and Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88(1)*, 100-112.

Hogan, R. (2005). In defense of personality measurement: New wine for old whiners. *Human Performance*, *18*(4), 331-341.

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17(3)*, 207-219.

Hoogeveen, K., & Gutkin, T. B. (1986). Collegial ratings among school personnel: An empirical examination of the merit pay concept. *American Educational Research Journal*, *23*(3), 375-381. (2)

Hough, L. M. (1992). The “Big Five” personality variables – construct confusion: Description versus prediction. *Human Performance, 5 (1 & 2),* 139-155.

Hough, L. M. (1997). The millennium for personality psychology: New horizons or good old daze. *Applied Psychology*, *47*(2), 233-261.

Hough, L. M., and Ones, D. S. (2001). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In N. R. Anderson, D. S. Ones, H. K. Sinangil, and C. Viswesvaran (Eds.), *Handbook of work psychology* (pp. 233-267). New York: Sage.

Hough, L. M., and Ones, D. S., and Viswesvaran, C. (2001). Personality correlates of managerial performance constructs. In R. C. Page (Chair), *Personality determinants of managerial potential, performance, progression and ascendancy.* Symposium conducted at the 13th Annual Convention of the Society for Industrial and Organizational Psychology, Dallas.

Hunter, J. E., and Schmidt, F. L. (Eds.) (2004). *Methods of Meta-Analysis (2nd Ed.).* Thousand Oaks, CA: SAGE.

Jacob, B. A., & Lefgren, L. (2005). Principals as agents: Subjective performance measurement in education. Working Paper (RWP05-040). (1)

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101-136.

Jayne, C. D. (1945). A study of the relationship between teaching procedures and educational outcomes. *Journal of Experimental Education, 14*, 101-134.

Johnson, J. W. (2001). The relative importance of task and contextual performance dimensions to supervisor judgments of overall performance. *Journal of Applied Psychology, 86(5)*, 984-996.

Johnson, J. W. (2003). Toward a better understanding of the relationship between personality and individual job performance. In M. R. Barrick and A. M. Ryan (Eds.), *Personality and work: Reconsidering the role of personality in organizations* (pp. 83-120). San Francisco, CA: Jossey-Bass.

Johnson, J. W., and Hezlett, S. A. (2008). Modeling the influence of personality on individuals at work: A review and research agenda. In S. Cartwright & C. L. Cooper (Eds.), *Oxford handbook of personnel psychology* (pp. 59-92). Oxford, England: Oxford University Press.

Jones, E. S. (1923). The prediction of teaching success for the college student. *School & Society, 18*, 685-690. (3)

Jones, M. L. (1956). Analysis of certain aspects of teaching ability. *The Journal of Experimental Education, 25(2)*, 153-180. (4)

Jones, R. D., and Barr, A. S. (1946). The prediction of teaching efficiency from objective measures. *The Journal of Experimental Education, 15(1)*, 85-100. (1) (2) (4)

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly*, *45*(1), 34-70. (1)

Knight, F. B. (1922). *Qualities related to success in teaching* (No. 120). Teachers college, Columbia University. (2)

LaDuke, C. V. (1945). The measurement of teaching ability: Study number three. *The Journal of Experimental Education, 14(1)*, 75-100. (1)

Lai, E. R., Auchter, J. E., & Wolfe, E. W. (2012). Confirmatory factor analysis of certification assessment scores from the national board for professional teaching standards. *The International Journal of Educational and Psychological Assessment, 9(2)*, 61-81.

Lamke, T. A. (1951). Personality and teaching success. *Journal of Experimental Education, 20*, 217–259. (2)

Leeds, C. H. (1950). A scale of measuring teacher-pupil attitudes and teacher-pupil rapport. *Psychological Monographs: General and Applied*, *64*(6), 1-24. (4)

Leeds, C. H. (1969). Predictive validity of the Minnesota Teacher Attitude Inventory. *Journal of Teacher Education, 20(1)*, 51-56. (4)

Leeds, C. H. (1972). The predictive validity of the "Minnesota Teacher Attitude Inventory." Final Report. (ERIC Document ED072111). (4)

Lins, L. J. (1946). The prediction of teaching efficiency. *The Journal of Experimental Education, 15(1)*, 2-60. (1) (2) (3)

Manatt, R. P., and Daniels, B. (1990). Relationships between principals' ratings of teacher performance and student achievement. *Journal of Personnel Evaluation in Education,4(2)*, 189-201.

McCall, W. A., & Krause, G. R. (1959). Measurement of teacher merit for salary purposes. *The Journal of Educational Research*, *53*(2), 73-75. (1)

McElvain, J. L., Fretwell, L. N., and Lewis, R. B. (1963). Relationships between creativity and teacher variability. *Psychological Reports, 13,* 186. (4)

Medley, D. M., and Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research, 80(4)*, 242-247. (1)

Medley, D. M., and Mitzel, H. E. (1959). Some behavioral correlates of teacher effectiveness. *The Journal of Educational Psychology, 50(6)*, 239-246. (1) (2)

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33-53. (1)

Montross, H. W. (1954). Temperament and teaching success. *The Journal of Experimental Education, 23(1)*, 73-97. (2) (4)

Morsh, J. E., and Wilder, E. W. (1954). Identifying the effective instructor: A review of the quantitative studies, 1900-1952. Air Force Personnel and Training Research Center, Chanute AFB, IL: AFTPRC-TR-54-44.

Motowidlo, S. J., and Van Scotter, J. R. (1994).Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology, 79(4)*, 475-480.

Mount, M. K., Barrick, M. R., and Strauss, J. (1994). Validity of observer ratings of the Big Five personality factors. *Journal of Applied Psychology, 79(2)*, 272-280.

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater, and level effects in 360-degree performance ratings. *Personnel Psychology, 51*, 557-576.

Nanninga, S. P. (1928). Estimates of teachers in service made by graduate students as compared with estimates made by principal and assistant principal. *The School Review, 36(8)*, 622-626. (2)

Nelson, K. G., Bicknell, J. E., & Hedlund, P. A. (1956). *Cooperative Study to Predict Effectiveness in Secondary School Teaching: Development and Refinement of Measures of Teaching Effectiveness; First Report*. University of the State of New York, State Education Department. (2)

O*NET OnLine. www.onetonline.org

Odenweller, A. (1936). Predicting the quality of teaching. *Contributions to Education* (No. 676). New York City: Teachers College, Columbia University. (3)

Orphanos, S. A. (2008). *Do Good Grades Make a Good Teacher? An Investigation of the Relationship Between Teachers' Academic Performance and Perceived Teacher Effectiveness in Cyprus*. Dissertation Stanford University. (2)

Ort, V. K. (1964). A study of some techniques used for predicting the success of teachers. *Journal of Teacher Education*, *15*(1), 67-71. (4)

Ostrander, L. R. (1996). Multiple judges of teacher effectiveness: Comparing teacher self-assessments with the perceptions of principals, students, and parents. Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12). (2)

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48(1)*, 163-193.

Payne, D. A., & Hulme, G. (1988). The development, pilot implementation, and formative evaluation of a grass-roots teacher evaluation system—or, the search for a better lawnmower. *Journal of Personnel Evaluation in Education*, *1*(4), 365-372.

Peterson, K. D. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal*, *24*(2), 311-317. (2)

Prick, L. G. M. (1989). Satisfaction and stress among teachers. *International Journal of Educational Research, 13(4)*, 363-377.

Reed, H. J. (1953). An investigation of the relationship between teaching effectiveness and teacher's attitude of acceptance. *Journal of Experimental Education, 21(4),* 277-325. (2)

Riesch, K. P. (1949). A study of some factors in pupil growth. *The Journal of Experimental Education, 18(1)*, 31-55. (1)

Riner, P. S. (1992). A comparison of the criterion validity of principals' judgments and teachers' self-ratings on a high-inference rating scale. *Journal of Curriculum and Supervision, 7(2)*, 149-169. (1) (2)

Rolfe, J. F. (1945). The measurement of teaching ability: Study number two. *The Journal of Experimental Education, 14(1)*, 52-74. (2) (4)

Rostker, L. E. (1945). The measurement of teaching ability: Study number one. *The Journal of Experimental Education, 14(1)*, 6-51. (1) (2)

Roth, L. H. (1961). Selecting supervising teachers. *Journal of Teacher Education, 12(4)*, 476-481.

Ruediger, W. C., & Strayer, G. D. (1910). The qualities of merit in teachers. *Journal of Educational Psychology*, *1*(5), 272-278. (3)

Sanders, W. L. (2000). Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14(4)*, 329-339.

Sarason, S. B. (1977). *Work, aging and social change*. New York: The Free Press.

Scates, D. E., & Hedlund, P. A. (1953). Cooperative Study to Predict Effectiveness in Secondary School Teaching. *Journal of Teacher Education*, *4*(3), 230-234. (2)

Schochet, P. Z., and Chiang, H. S. (2010). *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains* (NCEE 2010-4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Schwartz, A. N. (1950). A study of the discriminating efficiency of certain tests of the primary source personality traits of teachers. *The Journal of Experimental Education, 19(1)*, 63-93.

Scullen, S. E., Mount, M. K., Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*, 956-970.

Seagoe, M. V. (1946). Prediction of in-service success in teaching. *The Journal of Educational Research*, *39*(9), 658-663. (4)

Shaffer, J. G. (1990). A study of the relationship between student achievement and summative teacher evaluations (Doctoral dissertation, East Texas State University). (2)

Shechtman, Z. (1992). A group assessment procedure as a predictor of on-the-job performance of teachers. *Journal of Applied Psychology*, *77*(3), 383-387. (4)

Simmons, E. (1932). Correlation of administrators ratings of teacher and pupil achievement (Masters Thesis, George Peabody College for Teachers). (1) (2)

Simun, P. B., and Asher, J. W. (1964). The relationship of variables in undergraduate school and school administrators' ratings of first-year teachers. *Journal of Teacher Education, 15(3),* 293-302. (3)

Singer, A. (1954). Social competence and success in teaching. *The Journal of Experimental Education*, *23*(2), 99-131. (2)

Somers, G. T. (1923). *Pedagogical prognosis: predicting the success of prospective teachers* (No. 140). Teachers College, Columbia University. (4)

Spiggle, J. A. B. (2003). Relationship Between Teacher Performance and Student Growth Outcomes in a School District in North Carolina's Public Schools' Fifth Grades (Doctoral Dissertation, North Carolina State University). (1)

Start, K. B. (1966). The relation of teaching ability to measures of personality. *British Journal of Educational Psychology, 36*, 158–165. (4)

Stoelting, G. J. (1955). The selection of candidates for teacher education at the University of Wisconsin. *The Journal of Experimental Education, 24(2)*, 115-132. (2)

Strom, R. D., and Larimore, D. (1970). Predicting teacher success: The inner city. *The Journal of Experimental Education, 38(4)*, 69-77. (4)

Symonds, P. M. (1955). Characteristics of the effective teacher based on pupil evaluations. *The Journal of Experimental Education*, *23*(4), 289-310. (2)

Taylor, H. (1930). Teacher influence on class achievement: A study of the relationship of estimated teaching ability to pupil achievement in reading and arithmetic. *Genetic Psychology Monographs, 7(2)*, 81-174. (1) (2)

Thomas, D. R., Zumbo, B. D., Kwan, E., and Schweitzer, L. (2014). On Johnson's (2000) relative weight method for assessing variable importance: A reanalysis. *Multivariate Behavioral Research, 49(4)*, 329-338.

Tiegs, E. W. (1928). *An evaluation of some techniques of teacher selection*. Public school publishing company. (2)

Tolor, A. (1973). Evaluation of perceived teacher effectiveness. *Journal of Educational Psychology*, *64*(1), 98-104. (2)

Ullman, R. R. (1931). *The prognostic value of certain factors related to teaching success*. Ashland, OH: A. L. Garber Co. (2) (4)

Vecchio, R. P. (1993). Self- and supervisor ratings: A dyadic approach. *The International Journal of Organizational Analysis, 1(1)*, 73-83. (2)

Washington, A. D. (2011). *Formal evaluation of teachers: An examination of the relationship between teacher performance and student achievement* (Doctoral dissertation, University of South Carolina). (1)

Wilkerson, D. J., Manatt, R. P., Rogers, M. A., and Maughan, R. (2000). Validation of student, principal, and self-ratings in 360 degree feedback for teacher evaluation. *Journal of Personnel Evaluation in Education, 14(2)*, 179-192. (1) (2)

Table 1

Direct Determinants of Performance

Campbell (1990)	Declarative Knowledge	Procedural Skill	Motivation
Danielson (1996)	Planning and Preparation	Classroom Environment Instruction Professional Responsibilities	Not Present
Goe (2007)	Teacher Qualifications	Teacher Practices	Not Present
Lai, Auchter, & Wolfe (2012)	Content Knowledge	Teaching Skill	Not Present

Table 2

Components of Job Performance

Campbell (1990)	*Danielson (1996)*	*O*NET*
Oral & Written Communication	3a – Communicating clearly & accurately 3b – Using questioning & discussion techniques 4c – Communicating with families	Speaking (skills) Writing (skills) Oral expression & comprehension (abilities) Written expression & comprehension (abilities)
Demonstrating Effort		Achievement/effort (work styles) Initiative (work styles) Persistence (work styles)
Maintain Personal Discipline	4f – Showing professionalism	Dependability (work styles) Integrity (work styles) Self-control (work styles) Stress tolerance (work styles)
Job Specific task proficiency	1a – Demonstrating knowledge of content & pedagogy 1b – Demonstrating knowledge of students 1c – Selecting instructional goals 1d – Demonstrating knowledge of resources 1e – Designing coherent instruction 1f – Assessing student learning 2e – Organizing physical space 3c – Engaging students in learning 4a – Reflecting on teaching	Instructing students Adapting teaching and materials to the needs of the students Establishing clear objectives Preparing materials for class

Non-job specific task proficiency	3d – Providing feedback to students 3e – Demonstrating flexibility & responsiveness 4b – Maintaining accurate records 4e – Growing & developing professionally	Meeting with parents to discuss progress or behavior Enforce all administration policies and rules governing students Maintain accurate, complete, and correct student records as required by laws, district policies, and administrative regulations
Facilitating peer & team performance	2a – Creating an environment of respect & rapport 2b – Establishing a culture for learning	Communicating with supervisors, peers, or subordinates (work activities) Establishing & maintaining interpersonal relationships (work activities) Coordinating the work & activities of others (work activities) Resolving conflicts & negotiating with others (work activities)
Supervision		Training & teaching others (work activities) Coaching & developing others (work activities) Guiding, directing & motivating subordinates (work activities)
Management/Administration	2c – Managing classroom procedures 2d – Managing student behavior 4d – Contributing to the school and district	Administration & management (knowledge)

Table 3

Components of Teacher Supervisory Performance

Roth (1961)	Copas (1984)	Farbstein (1963)
Arranged for conferences.	Provided for interaction with the student teacher through conferences.	Provides for periodical or regular conferences with the student teacher.
Interrupted appropriately.	Interrupted the student teacher's lesson at appropriate times and in an appropriate manner.	Avoids interrupting or interfering with lessons being taught by the student teacher.
Maintained flexible scheduling		
Used practices worthy of imitation.		Demonstrates effective discipline and disciplinary techniques. Demonstrates ability to teach specific subjects effectively.
Studied children.	Provided opportunities for the student teacher to study children and their learning processes.	Permits the student teacher to get to know children.
Worked as a team with the student teacher.		Provides the student teacher with experiences in team teaching.
Provided full-time teaching experience.	Helped the student teacher develop skills in planning and evaluating learning experiences.	Leaves the classroom on occasion allowing the student teacher to teach without observation.
Inducted the student teacher gradually.	Structured responsibilities which gradually inducted the student teacher into full-time teaching.	Provides for gradual induction of the student teacher into the teaching program.
Alleviated frustrations.	Demonstrated sensitivity to the emotional needs of the student teacher in personal relationships.	Is pleasant toward the student teacher. Displays control of temper at all times.
Shared ideas.	Worked with the student teacher in developing skills of presentation.	Makes helpful suggestions regarding teaching procedures.
Encouraged the student teacher to use his own ideas.	Encouraged the student teacher to explore and develop unique teaching behaviors.	Permits the student teacher to use her own teaching methods or techniques.

Provided for the student teacher to reach his goals.	Assisted the student teacher in developing skills of discipline and control throughout the student teaching experience. Helped the student teacher locate resource materials, persons, and supplementary materials.	Provides opportunities for the student teacher to participate in school routine and clerical tasks, pupil guidance activities, scholarly activities, school activities, assign homework.
Gave the student teacher an awareness of his strengths and weaknesses.	Observed the student teacher and provided feedback as to the effectiveness of performance.	Keeps a record of the student teacher's progress and shares it with the student teacher.
Remained available.		
Treated the student teacher as a teacher.	Accepted the student teacher as a co-worker of equal status in guiding the learning process.	Refers to the student teacher as a teacher when introducing her to pupils.
Placed confidence in the student teacher.		
Gave praise with criticism.	Informed the student teacher of errors in a manner which protected the student teacher from embarrassment.	Provides constructive criticism or comments following observation.
Had faith in himself.		
Defined requirements clearly.	Provided the student teacher with information basic to adjustment to the class and school.	Defines and describes what is expected of the student teacher.

Table 4

Berk's (1988) Factors Which Can Impact Teacher Effectiveness

School Characteristics

1. School conditions
   - School library
   - Class size
   - Size of a type of class
   - Age of building
   - Size of school site
   - Size of school enrollment
   - Size of staff
   - Turnover of staff
   - Expenditures
   - Quality of instructional materials and equipment
   - Schoolwide learning climate
   - Instructional support
2. Instructional personnel
   - Education degree
   - Undergraduate education type
   - Teaching experience
   - Verbal achievement
   - Race
   - Gender
   - Teaching load
   - Time in discipline
   - Job satisfaction

Test Characteristics

1. Type of achievement test
2. Curricular and instructional validity
3. Test score metric

Pretest-Posttest Design Characteristics

1. History
2. Maturation
3. Statistical regression
4. Mortality
5. Interactions with selection
6. Multiple sources of invalidity

Student characteristics

1. Intelligence
2. Attitude
3. Socioeconomic level
4. Race/ethnicity
5. Gender
6. Age
7. Attendance

Table 5

Results of Past Researchers' Self-Report Personality Meta-Analyses

		Barrick & Mount (1991)		
Dimension		*Number of r's*	*Obs r*	*Population*
Extraversion				
Supervisor ratings		93	.08	.14
Professionals		4	-.05	-.09
Emotional Stability				
Supervisor ratings		95	.05	.09
Professionals		5	-.07	-.13
Agreeableness				
Supervisor ratings		83	.05	.09
Professionals		7	.01	.02
Conscientiousness				
Supervisor ratings		94	.15	.26
Professionals		6	.11	.20
Openness to Experience				
Supervisor ratings		62	.02	.04
Professionals		4	-.05	-.08

		Barrick, Mount, & Judge (2001)		
Dimension	*No. of MA's*	*k*	*Obs r*	*Population*
Extraversion				
Work performance	5	222	.06	.12
Supervisor ratings	4	164	.07	.11
Professionals	1	4	-.05	-.09
Emotional Stability				
Work performance	5	224	.06	.12
Supervisor ratings	4	167	.07	.12
Professionals	2	8	.04	.06
Agreeableness				
Work performance	5	206	.06	.10
Supervisor ratings	4	151	.06	.10
Professionals	2	10	.03	.05
Conscientiousness	5			
Work performance		239	.12	.23
Supervisor ratings	4	185	.15	.26
Professionals	1	6	.11	.20
Openness to Experience				
Work performance	5	143	.03	.05
Supervisor ratings	4	116	.03	.05
Professionals	1	4	-.05	-.08

Table 6

Connelly and Ones (2010) Other Ratings of Personality and Job Performance Meta-Analysis

	k	N	r	ρ_{ov}	ρ_{xy}
Emotional Stability	7	1,190	.14	.17	.37
Extraversion	6	1,135	.08	.11	.18
Openness	6	1,135	.18	.22	.45
Agreeableness	7	1,190	.13	.17	.31
Conscientiousness	7	1,190	.23	.29	.55

Table 7

Correlations Between Principal Ratings and Student Score Improvement Metrics

	N	*k*	r_{obs}	SD_{obs}	SD_{ρ}	90% cred.
Student Gain	2,490	40	.17	.13	.02	.15 to .19
Measure of Gain						
Gain or Growth of Scores	1,667	28	.15	.12	.00	.15 to .15
Value-Added	823	12	.23	.13	.05	.16 to .30
Initial Test Score	135	2	-.02	.01	.00	-.02 to -.02
Final Test Score	161	3	.10	.13	.00	.10 to .10
Type of Achievement Test						
Arithmetic	1,183	20	.24	.17	.12	.10 to .40
Reading	1,674	20	.19	.12	.06	.12 to .26
Language Arts	122	3	.10	.34	.30	-.29 to .49

Note. N = sample size, *k* = number of studies, r_{obs} = sample size weighted average correlation, SD_{obs} = standard deviation of observed correlations, SD_{ρ} = standard deviation of true score correlation, 90% cred. = 90% credibility interval.

Table 8

Correlations Between Different Principal Rating Scales and Student Gains

Rating Scale	**N**	***k***	r_{obs}	SD_{obs}	SD_{ρ}	**90% cred.**
Almy-Sorenson	208	5	.16	.17	.08	.06 to .26
Michigan	168	4	.17	.15	.02	.14 to .20
Torgerson	168	4	.16	.17	.08	.05 to .27
Wisconsin M-Blank	116	6	.16	.08	.00	.16 to .16

Note. N = sample size, *k* = number of studies, r_{obs} = sample size weighted average correlation, SD_{obs} = standard deviation of observed correlations, SD_{ρ} = standard deviation of true score correlations, 90% cred. = 90% credibility interval.

Table 9

Correlations Between Performance Ratings Across Multiple Rater Sources

	N	*k*	r_{obs}	SD_{obs}	SD_{ρ}	90% cred.
Principal - Peer	660	15	.57	.27	.25	.25 to .89
Principal - Self	1,603	24	.15	.10	.00	.15 to .15
Principal - Student	1,783	31	.31	.21	.18	.08 to .54
Principal – Parent	704	5	.10	.13	.10	-.02 to .22
Principal – Other	1,798	35	.45	.16	.12	.31 to .59
Principal – Outside Agency	477	8	.36	.20	.17	.14 to .58
Principal – Observers	1,219	21	.46	.14	.09	.34 to .58
Principal - Supervisor	464	13	.48	.16	.10	.35 to .51
Student - Peer	259	8	.39	.28	.23	.10 to .68
Student - Self	546	10	.13	.15	.07	.04 to .22
Student - Other	698	11	.29	.13	.05	.22 to .36
Student – Outside Agency	114	3	.21	.03	.00	.21 to .21
Student – Observers	677	10	.29	.14	.08	.18 to .40
Student - Parent	389	3	.53	.08	.05	.48 to .58
Self – Peer	357	4	.13	.10	.00	.13 to .13
Self – Other	127	3	.21	.06	.00	.21 to .21
Self – Parent	322	2	.07	.02	.00	.07 to .07
Peer – Other	83	2	.28	.09	.00	.28 to .28
Peer - Parent	282	2	.16	.03	.00	.16 to .16

Note. N = sample size, *k* = number of studies, r_{obs} = sample size weighted average correlation, SD_{obs} = standard deviation of observed correlations, SD_{ρ} = standard deviation of true score correlations, 90% cred. = 90% credibility interval.

Table 10

All Correlations of Peer Ratings

	N	*k*	r_{obs}	SD_{obs}	SD_{ρ}	90% cred.
Peer - Principal	660	15	.57	.27	.25	.25 to .89
Peer - Student	259	8	.39	.28	.23	.10 to .68
Peer – Other	83	2	.28	.09	.00	.28 to .28
Peer - Self	357	4	.13	.10	.00	.13 to .13

Table 11

All Correlations of Self-Ratings

	N	*k*	r_{obs}	SD_{obs}	SD_{ρ}	90% cred.
Self – Other	127	3	.21	.06	.00	.21 to .21
Self - Principal	1,603	24	.15	.10	.00	.15 to .15
Self - Student	546	10	.13	.15	.07	.04 to .22
Self – Peer	357	4	.13	.10	.00	.13 to .13
Self – Parent	322	2	.07	.02	.00	.07 to .07

Table 12

All Correlations of Student Ratings

	N	*k*	r_{obs}	SD_{obs}	SD_{ρ}	90% cred.
Student - Parent	389	3	.53	.08	.05	.48 to .58
Student - Peer	259	8	.39	.28	.23	.10 to .68
Student - Principal	1,783	31	.31	.21	.18	.08 to .54
Student - Other	698	11	.29	.13	.05	.22 to .36
Student – Observers	677	10	.29	.14	.08	.18 to .40
Student – Outside Agency	114	3	.21	.03	.00	.21 to .21
Student - Self	546	10	.13	.15	.07	.04 to .22

Table 13

All Correlations of Other Ratings

	N	*k*	r_{obs}	SD_{obs}	SD_{ρ}	90% cred.
Supervisor - Principal	464	13	.48	.16	.10	.35 to .51
Observers - Principal	1,219	21	.46	.14	.09	.34 to .58
Other - Principal	1,798	35	.45	.16	.12	.31 to .59
Outside Agency - Principal	477	8	.36	.20	.17	.14 to .58
Other - Student	698	11	.29	.13	.05	.22 to .36
Observers - Student	677	10	.29	.14	.08	.18 to .40
Other - Peer	83	2	.28	.09	.00	.28 to .28
Outside Agency - Student	114	3	.21	.03	.00	.21 to .21
Other - Self	127	3	.21	.06	.00	.21 to .21

Table 14

Sources of "Other" Ratings by Different Raters

Lins (1946)

High School Principal personality rating of

- Appearance and manner – Coded as Agreeableness
- Needs prodding – Coded as Conscientiousness
- Do others do what he wishes – Coded as Extraversion
- Control of emotions – Coded as Emotional Stability
- Finite purpose – Coded as Conscientiousness

Two interviewers during college rated the following

- Professional judgment – Coded as Conscientiousness
- Social – Coded as Extraversion
- Work habits – Coded as Conscientiousness
- Motivation and values – Coded as Conscientiousness
- Originality, creativity, and initiative – Coded as Openness to Experience
- General impressions of the student – Coded as Alpha

Simun and Asher (1964)

College Faculty from the Senior Year rated students on a 5-point graphic scale

- Appearance – Coded as Emotional Stability
- Poise – Coded as Emotional Stability
- Conversation – Coded as Extraversion
- Judgment – Coded as Conscientiousness
- Initiative – Coded as Conscientiousness
- Professional Competence – Coded as Conscientiousness
- Cooperation – Coded as Agreeableness
- Reliability – Coded as Conscientiousness
- Personality – Coded as Alpha

Jones (1923)

Combination of ratings by teachers and fellow students during senior year of college on a five point scale on the following qualities:

- Energy – Coded as Extraversion
- Sociability – Coded as Extraversion
- Reliability – Coded as Conscientiousness
- Personality – Coded as Alpha

Somers (1923)

- Personality – At the end of the first semester, college teachers were asked to estimate for individuals each of eight personality traits
- 3-7 judges were averaged for each trait which were weighted into a composite score
- Coded as Alpha

Odenweller (1936)

- Three teachers in the building gave ratings of the personality
- Coded as Alpha

Shechtman (1992)

- Assessment Center Rating – three clusters of behavior: oral communication – clarity and organization of thoughts, focus on essentials, logical presentation and sequential development of thoughts, verbal expressiveness, fluency of speech, precision and extensive vocabulary; human relationships – expression of warmth, friendliness, supportiveness, display of respect, sensitivity, and rapport; leadership – dynamism, alertness, initiative, enthusiasm, responsibility, self-assurance, and self-directiveness; overall rating – general fitness for the teaching profession
- 6-point scale; average score
- Nondirective group introduction, followed by a directive group interview that focused on attitudes and values, two controversial topics were discussed, leaderless group discussion in which members acted as a committee to solve a problem, provided feedback to one another and assessor
- Coded as Alpha

Ambady and Rosenthal (1993)

Undergraduates rated teachers on the dimensions of

- accepting – Coded as Agreeableness
- active – Coded as Extraversion
- attentive – Coded as Conscientiousness
- competent – Coded as Conscientiousness
- confident – Coded as Emotional Stability
- dominant – Coded as Extraversion
- empathetic – Coded as Agreeableness
- enthusiastic– Coded as Extraversion
- honest – Coded as Conscientiousness
- likeable – Coded as Agreeableness
- optimistic – Coded as Emotional Stability
- professional – Coded as Conscientiousness
- supportive – Coded as Agreeableness
- warm – Coded as Extraversion
- global – Coded as Alpha

Table 15

Sources of "Other" Ratings by the Same Raters

Boyce (1912)

Principal or superintendent rank of

- Energy and endurance – Coded as Extraversion
- Self-control – Coded as Emotional Stability
- Sympathy-tact – Coded as Extraversion
- Adaptability – Coded as Openness to Experience
- Sense of humor – Coded as Extraversion
- Fair mindedness – Coded as Agreeableness
- Initiative – Coded as Conscientiousness
- Executive capacity – Coded as Emotional Stability
- Co-operation – Coded as Agreeableness
- Intellectual capacity – Coded as Openness to Experience
- Instructional skill – Coded as Conscientiousness
- Governmental skill (discipline) – Coded as Conscientiousness
- Studiousness – Coded as Conscientiousness

Ruediger and Strayer (1910)

Principal or superintendent gave rankings of

- Initiative or originality – Coded as Openness to Experience
- Strength of personality – Coded as Alpha
- Teaching skill – Coded as Conscientiousness
- Order or ability to control – Coded as Conscientiousness
- Following suggestions – Coded as Conscientiousness
- Accord between teacher and pupil – Coded as Extraversion
- Studiousness or progressive scholarship – Coded as Conscientiousness
- Social factor outside of school – Coded as Extraversion

Baird and Bates (1929)

Principal rating of

- Personality – Coded as Alpha
- Social Intelligence – Coded as Alpha
- Professional Spirit – Coded as Emotional Stability
- Control over method of teaching – Coded as Conscientiousness
- Executive ability – Coded as Emotional Stability
- Adaptability – Coded as Openness to Experience

Bradley (1918)

Principal's rating of

- Physical efficiency – Coded as Extraversion
- Moral efficiency – Coded as Agreeableness
- Intellectual efficiency – Coded as Conscientiousness
- Directed efficiency – Coded as Emotional Stability
- Professional efficiency – Coded as Conscientiousness
- Social efficiency – Coded as Extraversion
- School Management – Coded as Conscientiousness
- Government, discipline – Coded as Conscientiousness

Odenweller (1936)

- Principal or supervisor gave ratings of personality
- Coded as Alpha

Rolfe (1945)

- Scale for Personal Fitness of Teachers - Scale consists of 33 teacher traits such as accuracy, health, loyalty, sociability, thrift, etc., each of which is rated on an 11-point scale. If the rater could think of no teacher who was better with respect to a given trait, then check 0 values, and proceeded with higher scores as more teachers were better. Score was the arithmetic average of the numbers checked for the 33 traits. Three raters rated each teacher on this scale: county superintendent of the schools, county supervisor, and the investigator; Coded as Alpha.
- Personality Rating Scale - Rater was asked to assign one of eleven intensity values to each of six terms (pleasing, forceful, wholesome, interesting, stimulating, and confidence-inspiring) which are descriptive of the total personality effect the teacher has upon others. Rater checked 11-point scale (0 - 10) on how many people could think of who were superior to the teacher being rated; Coded as Alpha.

Table 16

Correlations of Principal Ratings with Other's Ratings of Personality

	N	*k*	r_{obs}	SD_{obs}	SD_{ρ}	90% cred.
Alpha						
Different Raters	997	8	.45	.17	.16	.25 to .65
Same Raters	1,127	4	.71	.13	.12	.55 to .87
Conscientiousness						
Different Raters	216	4	.23	.09	.00	.23 to .23
Same Raters	1,095	4	.80	.08	.07	.70 to .90
Emotional Stability						
Different Raters	162	3	.23	.07	.00	.23 to .23
Same Raters	1,095	3	.72	.10	.09	.60 to .84
Agreeableness						
Different Raters	162	3	.19	.14	.05	.12 to .26
Same Raters	380	2	.69	.12	.11	.55 to .83
Openness to Experience						
Different Raters*	54	1	.17	-	-	
Same Raters	880	3	.60	.04	.01	.59 to .61
Extraversion						
Different Raters	214	4	.17	.12	.00	.17 to .17
Same Raters	583	3	.53	.09	.08	.43 to .63

Note. N = sample size, *k* = number of studies, r_{obs} = sample size weighted average correlation, SD_{obs} = standard deviation of observed correlations, SD_{ρ} = standard deviation of true validity correlations, 90% cred. = 90% credibility interval.
*There was only one study with different raters. The correlation was the one calculated by the researcher.

Table 17

Sources of Alpha Self-Ratings

Social Intelligence scale **- Ullman (1931)**
Intended to measure the subject's ability to get along with others with six sections.
1. Judgment of social situations - 30 questions – a social problem is stated and 4 possible solutions are given
2. Memory for names and faces - Handed a paper and names and faces of 12 people; given 4 minutes to learn, take section one; then given 25 pictures and have to pick original ones
3. Recognition of mental states from facial expressions - Correctly identify 12 mental states as illustrated by pictures included in the test
4. Observation of human behaviors – 30 statements about human behavior; decide whether true or false
5. Social information – 50 questions; true or false covering facts which should be known by those who are interested in matters which are significant from the social standpoint
6. Recognition of the mental state of the speaker – Indicate from the list of 20 mental states given which one most accurately describes the person making each of the 27 statements

Minnesota Teacher Attitude Inventory - **Cook and Leeds (1947), Leeds (1952, 1967, 1972), Callis (1952), Day (1959)**
It consists of 150 opinion statements concerning the nature and behavior of children in general and pupils in particular. It is supposed to tap the non-cognitive elements in the teaching personality which would seem to relate to the ability to establish and maintain rapport with children; 5-point scale
Example items:
1. Most pupils do not make an adequate effort to prepare their lessons.
2. Most children are obedient.
3. Shyness is preferable to boldness.
4. Most pupils lack productive imagination.
5. Children dress more sensibly nowadays.
6. Children "should be seen and not heard"

Bell Adjustment Inventory - **Seagoe (1946), Jones and Barr (1946)**
Circle Yes, No, or ?
Example items:
1. Do you take responsibility for introducing people at a party?
2. Do you day-dream frequently?
3. Do you get discouraged easily?

Humm-Wadsworth Temperment Scale- **Seagoe (1946)**
Normal component –providing rational balance and temperamental equilibrium; hystoid component – criminalism and self-preservation; cyloid component – the maniac and the depressed; schizoid component – heightened imagination: autism and paranoia; epileptoid component – physical symptoms of epilepsy

Thurstone Personality Schedule - **Seagoe (1946)**
Social – items dealing with subjects' reactions to human environment (Are you troubled with shyness?); Extrovert – items dealing with subjects' reaction to nonhuman environment (Are you systematic in caring for your personal property?); Fantasy – items dealing with subjects' inner experience (Are you frequently burdened with a sense of remorse?); Physical – items dealing with subjects' somatic phenomena (Did you ever have anemia badly?); Parental – items dealing with subjects' immediate family (Was your mother the dominant member of the family?); Sex – items dealing with subjects' sex attitudes (Have you ever been afraid that you are sexually inferior to other men or women?)

Washburne Social Adjustment Inventory – **Gotham (1945), Rostker (1945), Rolfe (1945)**
123 questions; Answer yes or no to each statement; Some questions ask for a different response, such as very happy, fairly happy, neither, etc.; there are place to write chief wishes and suppressed desires. A copy can be obtained from the author.
Example items:
1. Did you ever cry because someone hurt you?
2. Do you always report other people whom you see cheating?
3. Do your friends call you a tease?

Morris Trait Index L - **Rostker (1945), Rolfe (1945)**
5 sections which ask for answers in various formats. A copy can be obtained from the author.
1. Mark your feeling about a word or statement. 5-point scale – Like very much to dislike very much. 38 items.
   a. Studying.
   b. People between 7 and 11 years of age.
   c. Slow pupils.
2. A statement is made about something a teacher may say. Must indicate which type of student the student is most appropriate for: bright ones, dull ones, careless ones, lazy ones, "bluffers", and conscientious ones.
   a. In your case quality is more important than quantity.
   b. "Practice makes perfect," you know! Try going over thing you are to learn an extra time or two.
3. A situation is given. Must characterize the situation from the standpoint of the teacher using these options: amusing, embarrassing, necessitating firm control, interesting, and necessitating correction of mistakes.
   a. Pupils call attention to the fact that a word which the teacher has written on the board is misspelled.
   b. Pupil asks a question related to the work but beyond the teacher's knowledge.
4. Statements are given. Indicate how you feel about the statement using the following options: always true, usually true, sometimes true or sometimes false, rarely true, and never true.
   a. It is easy to arouse other people's interest and curiosity.
   b. It is interesting for the teacher to prepare lessons.
5. A situation is presented with six options for how a person may feel about the situation. Mark how you would feel.

a. If a teacher mentioned reviewing work every two weeks, which idea would occur to you:
b. The only way I know to review is just to go over the work again.
c. For review it is interesting to have members of the class question each other.
d. Repetition is essential for learning.
e. Usually I learn something new when I review.
f. I like frequent chances to go over work.
g. Repetition is tedious.

Rudisill Scale for the Measurement of Personality of Elementary School Teachers – **Gotham (1945)**
Measures 20 personal traits assumed to be related to teaching efficiency: accuracy, adaptability, considerateness, attitude toward children, enthusiasm, cooperation, fairness, industry, interest in work, judgment, forcefulness, loyalty, optimism, leadership, originality, open-mindedness, progressiveness, sociability, understanding of children, and refinement.

Table 18

Correlations of Principal Ratings with Self-Ratings of Personality

	N	*k*	r_{obs}	SD_{obs}	SD_{ρ}	90% cred.
Alpha	850	12	.28	.14	.08	.18 to .38
Minnesota Teacher Attitude Inventory	543	6	.35	.10	.02	.33 to .37
Washburne Social Adjustment Inventory	71	2	.27	.15	.00	.27 to .27
Morris Trait Index L	96	3	.07	.05	.00	.07 to .07
Bell Adjustment Inventory	73	2	-.13	.13	.00	-.13 to -.13
Intelligence	179	4	.18	.21	.15	-.01 to .37

Note. N = sample size, *k* = number of studies, r_{obs} = sample size weighted average correlation, SD_{obs} = standard deviation of observed correlations, SD_{ρ} = standard deviation of true validity correlations, 90% cred. = 90% credibility interval

Table 19

Correlations of Principal Ratings with Self-Ratings of Big Five Personality Dimensions

	N	*k*	r_{obs}	SD_{obs}	SD_{ρ}	90% cred.
Conscientiousness	258	4	.03	.11	.00	.03 to .03
Cautiousness	223	3	.004	.07	.00	.004 to .004
Achievement	147	2	.13	.18	.14	-.05 to .31
Emotional Stability	336	6	.06	.10	.00	.06 to .06
Optimism	258	4	.02	.08	.00	.02 to .02
Self-Esteem	96	3	-.06	.15	.00	-.06 to -.06
Low Anxiety	168	2	-.06	.06	.00	-.06 to -.06
Openness to Experience	467	5	-.11	.13	.07	-.20 to 0
Traditionalism	258	4	-.004	.16	.10	-.13 to .13
Extraversion	496	4	-.03	.04	.00	-.03 to -.03
Activity/Energy Level	269	4	.08	.18	.13	-.09 to .25
Dominance	350	7	.09	.07	.00	.09 to .09
Warmth	258	4	.05	.06	.00	.05 to .05
Sociability	172	5	.02	.15	.00	.02 to .02
Reflective	97	2	.05	.04	.00	.05 to .05
Autonomy	97	2	-.04	.01	.00	-.04 to -.04
Agreeableness	n/a	n/a	n/a	n/a	n/a	n/a

Note. N = sample size, *k* = number of studies, r_{obs} = sample size weighted average correlation, SD_{obs} = standard deviation of observed correlations, SD_{ρ} = standard deviation of true validity correlations, 90% cred. = 90% credibility interval