THE DEVELOPMENT AND VALIDATION OF A MEASURE OF ADMINISTRATOR DECISION-MAKING IN STUDENT DISCIPLINE

by

JOSHUA DAVID KAHN

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2018

DISSERTATION APPROVAL PAGE

Student: Joshua David Kahn

Title: The Development and Validation of a Measure of Administrator Decision-Making in Student Discipline

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Educational Methodology, Policy, and Leadership by:

Dr. Michael D. Bullis	Chairperson, Advisor
Dr. Gina Biancarosa	Core Member
Dr. Keith Hollenbeck	Core Member
Dr. Charles Martinez	Core Member
Dr. Kent McIntosh	Institutional Representative

and

Sara D. Hodges	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2018

DISSERTATION ABSTRACT

Joshua David Kahn

Doctor of Philosophy

Department of Educational Methodology, Policy, and Leadership

June 2018

Title: The Development and Validation of a Measure of Administrative Decision-Making in Student Discipline

The art and success of being a competent school administrator relies in large part on the ability to make decisions that address problems effectively, equitably, and efficiently. Despite the importance of this skill, there is a dearth of psychometrically-sound, quantitative measures that focus on school-based administrators (i.e., principals and asst. principals) and the decisions they make. To fill this gap, this study developed and validated a constructed response measure of Administrator Decision-Making in Student Discipline (ADMin-SD). ADMin-SD was developed and validated in three iterative phases: examining the content validity of the items, followed by pilot testing them, and concluding with a field test. The instrument demonstrates adequate reliability and moderate discriminant validity. Implications for researchers include having a tool to conduct future studies of administrator decision-making. As ADMin-SD collects qualitative data and transforms it into quantitative scores, both qualitative and quantitative studies can be conducted. Practitioners have a measurement tool that can help guide instructors of administrative licensure programs in their development of instructional units on decision-making skills. Further, districts and states can identify who is a strong decision-maker in student discipline situations and who needs further professional development.

CURRICULUM VITAE

NAME OF AUTHOR: Joshua David Kahn

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Mercy College, Bronx, NY
Trinity College, Hartford, CT

DEGREES AWARDED:

Doctor of Philosophy, Educational Leadership, 2018, University of Oregon
Master of Science, Elementary and Urban Education, 2005, Mercy College
Bachelor of Arts, History, 2003, Trinity College

AREAS OF SPECIAL INTEREST:

Educational Decision-making
Quantitative Research Methods

PROFESSIONAL EXPERIENCE:

Adjunct Faculty,
Moravian College
September, 2017 – December, 2017

Research Assistant,
Behavioral Research and Teaching,
September, 2015 – June, 2017

GRANTS, AWARDS, AND HONORS:

Betty Foster McCue Dissertation Research Award, University of Oregon, 2017

College of Education Scholarship, University of Oregon, 2017

College of Education Doctoral Research Award, University of Oregon, 2016

PUBLICATIONS:

Anderson, D. J., Kahn, J.D., Tindal, G. A. (2017). Exploring the robustness of a unidimensional item response theory model with empirically multidimensional data. *Applied Measurement in Education, 30*(3), 163-177.

Kahn, J. D. & Girvan, E. (2017). Applying rules and standards accurately: Indeterminacy and transfer among adult learners. In a special issue of *Human Resources Development Quarterly*, *28*, 87-112.

Andreou, T. E., McIntosh, K., Ross, S. W., & Kahn, J. D. (2015). Critical incidents in the sustainability of school-wide positive behavioral interventions and supports. *Journal of Special Education*, *49*(3), 157-167. doi: 10.1177/0022466914554298.

ACKNOWLEDGMENTS

I wish to express my sincere gratitude to Dr. Michael D. Bullis, without whom this study would not have been conceived. His guidance and mentoring have been invaluable; I will be forever grateful to him. In addition, special thanks are due to my committee members – Dr. Gina Biancarosa, Dr. Keith Hollenbeck, Dr. Charles Martinez, and Dr. Kent McIntosh – who each made critical contributions to this study. I also must acknowledge the judges that participated in this study. I thank and commend them for sticking with the project; their insight, time, and effort were instrumental in shaping the final product of this study. I would also like to acknowledge Dr. Lauren Kahn without whom this study could never have been conducted – for innumerable reasons.

The study was supported in part by Dr. Michael D. Bullis as well as grants from the University of Oregon including the Betty Foster McCue scholarship and the Doctoral Research Award.

I dedicate this work to my family, who continually push to me to be a better father, husband, son, brother, and scholar. I also dedicate this work to educators who must make difficult and complex decisions every day and to the students whose lives are affected by those decisions.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

Table Page

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

Making decisions is the "sine qua non of administration" (Hoy & Tarter, 2008, xiii) in that "…deciding *is* the quintessential administrative act" (Allison, 1996, p. 5). Despite the importance and prevalence of this skill, there is a dearth of psychometrically-sound, quantitative measures that focus on school-based administrators (i.e., principals and asst. principals) and the types of decisions they make. Administrator decision-making has generally been measured observationally or through interviews, recorded qualitatively, and occasionally transformed into quantitative data. Although these approaches have helped generate theory, they are inadequate for testing it and inadequate for providing an objective measure of the skill. To fill this gap, I have developed, revised, and conducted an initial validation study of a measure of Administrator Decision-Making in Student Discipline (ADMin-SD). In this chapter, I lay the groundwork for the study by making a case for its importance as well as explicating the theoretical frameworks I used in this study. Following this introduction, I review the empirical studies of administrator decision-making and the common approaches to assessing the construct.

Before beginning, let me clarify how I use some terminology in this document. Though the terms *decision-making* and *problem-solving* are sometimes used interchangeably in social science research, I take the view that problem-solving is a specific type of decision-making that is constrained to generating solutions to problem situations. Problems are defined as negative circumstances for which a solution is not immediately obvious (D'Zurilla & Goldfried, 1969; Mayer, 1992; OECD, 2010; O'Neill & Schacter, 1997). However, the demarcation between problem-solving and decision-

making is somewhat arbitrary. For instance, solving math word problems is universally considered problem-solving, but firemen determining how to rush into a burning house to put out the fire is typically considered decision-making, and a principal who must shuffle classrooms to avoid recent storm damage could be construed as either problem-solving or decision-making. In my view, all problem-solving requires decisions to be made, but not all decisions require problems to solve. Therefore, I use the general term of decision-making because administrators face all types of situations routinely. Additionally, I refer to administrators generally and principals specifically throughout the manuscript. The inclusion of assistant principals is implied when principals are mentioned because ADMin-SD is intended to assess them as well because they may be responsible for student discipline in larger schools.

Importance and Statement of the Problem

School administrators have a substantial impact on the lives of students, particularly regarding their academic achievement (Branch, Hanushek, & Rivkin, 2013; Spillane, Halverson, & Diamond, 2004; Waters, Marzano, & McNulty, 2003). The art and success of being a competent school administrator relies in large part on the ability to make decisions that address problems efficiently, but most administrators struggle to make decisions effectively, equitably, and with few errors (Glasman, 1995; Hoy & Tarter, 2008; Leithwood & Steinbach, 1995). This struggle results from various causes including, but not limited to, unclear goals, complex social situations, and a lack of time or other resources. Additionally, administrators may waste resources and stall student learning by rushing to fix a problem rather than testing their assumptions about the problem's causes and potential solutions (Newmann, Smith, Allensworth, & Bryk, 2001;

Robinson, Meyer, Le Fevre, & Sinnema, 2015). Despite the importance of decision-making, I could not locate any technically-adequate, peer-reviewed performance measures of administrator decision-making ability. Thus, developing a way to assess and measure the performance of this skill is a crucial first step in building a program of research aimed at improving administrators' decision-making skills.

Due to the absence of standardized measures of this skill, the studies I review in the following pages generally constructed their own measure(s) for their respective studies but did not discuss the development or psychometric properties of the measures, which is problematic for several reasons. First, each research team developed the content of their items by themselves, without discussing pilot testing or content validation studies, bringing into question whether the items represent the content domain adequately. Second, results of studies cannot be compared directly because of differences in the measures that were used (e.g., one study used instructional vignettes, another used what they called "strategic" and "human relations" vignettes, and a third used "general administrative" vignettes). These studies made inferences about administrator decision-making in general even though expertise and the ability to make effective decisions is domain-specific (Chi, Glaser, & Farr, 1988). In other words, a principal, who is adept at solving student discipline problems, may not solve budget problems equally well. The inferences generated by these findings, then, should be localized to the domains reflected in the measures.

To address these validity concerns, I developed a measure of school administrator decision-making that focuses only on the way school administrators make decisions in response to student discipline situations. Principals spend a lot of time engaged in solving

these problems despite poor outcomes of their time. For example, Chan and Pool (2002) reported that principals spend the biggest portion of their time on student interactions and discipline. Despite this time spent, student discipline problems at the elementary, middle, and high school level result in disproportionate outcomes between students who are Black and White in terms of rates of suspension and expulsion (Skiba, Michael, Nardo, & Peterson, 2002; Skiba et al., 2011). To be precise, based on a sample of 364 elementary and middle schools drawn from a national database, school principals were sent office discipline referrals from teachers at a disproportional rate; Black students in middle school were 3.78 times more likely than White students to be referred to the office for disciplinary issues (Skiba et al., 2011). Principals usually endorsed the referrals, perpetuating the disproportionality between Black and White students, leading to a reduction of opportunity and instructional time for Black students.

Overall, improving administrators' decision-making skills should make schools function more efficiently and improve outcomes for students and teachers (Brenninkmeyer & Spillane, 2008; Leithwood & Steinbach, 1995). Specifically, improving administrators' deliberate decision-making skills may reduce the influence of implicit bias when making disciplinary decisions (Godsil, Tropp, Goff, & Powell, 2014; Kahneman, 2011). Further, time spent on disciplinary problems should decrease and decisions regarding student disciplinary situations should be made more equitably because, as expertise in a domain increases, the number of errors made and the time spent on solving problems tends to decrease (Chi, Glaser, & Farr, 1988). Lastly, improving school-wide student discipline, and reducing the amount of time spent managing it, will allow administrators to spend more time on instruction (Scott & Barret, 2004).

Construction of ADMin-SD is an initial step aimed at yielding these distal benefits, as a reliable and valid measure of this skill would contribute to the field of educational administration in at least four specific ways. First, ADMin-SD helped this line of research move from theory-building to hypothesis-testing. Of 12 studies in the literature pool, nine were qualitative. A quantitative measure allows the field to test some of the hypotheses generated by those qualitative studies. Second, a reliable measure improves comparability across studies and allow for systematic replication of intervention studies. Third, from a practitioner standpoint, schools, districts, and states could include a valid and reliable tool in their evaluations of school administrators. Fourth, administrator preparatory programs could use the measure for instructional purposes, including the training and assessment of future and current administrators.

School administrators' impact on students originates largely through the decisions they make. Yet they struggle to make those decisions effectively and equitably. Coincident with these struggles, and as I will discuss below, the literature base does not contain a quantitative measure of this ability. This study then contributes a quantitative measure that will help practitioners and researchers assess this skill for various purposes.

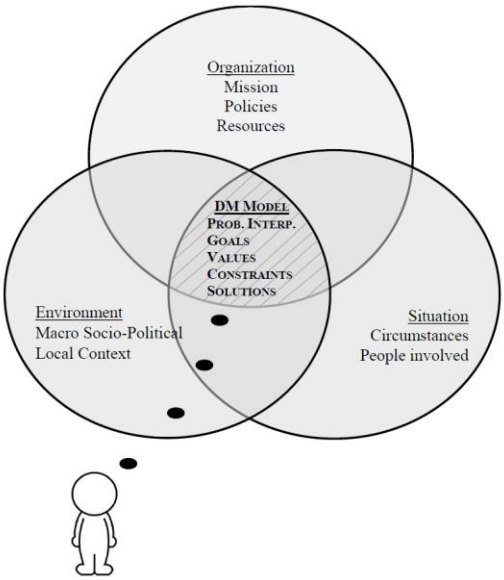
To construct this measure, I rely on the contingency theory of administration and the cognitive approach to school leadership. The following sections explicate these theoretical orientations and how they guide the construction of ADMin – SD.

Theoretical Grounding

Two theoretical frames guide the conceptual development of ADMin-SD and the methods of this study: Contingency Theory of Administration (Derr & Gabarro, 1972; Donaldson, 2001; Hoy & Miskel, 1987) and the Cognitive Approach to School

Leadership (Hallinger, Leithwood, & Murphy, 1993). Contingency theory posits that administrators must understand how the environment, the organization, and the situation work together to respond optimally to problems. The Cognitive Approach to School Leadership (CASL) specifically examines principals' decision-making processes and the problems they face. The theories are compatible and complementary because, together, they address how principals make decisions within a contingent environment. Contingency theory provides the macro view of the context in which administrators make decisions with a rationale for the structure and constructed response format. CASL provides the micro view of the mechanics of how administrators make those decisions and thus provides the substance and content of ADMin-SD. Figure 1 presents how I have integrated these theories; it depicts an administrator using the decision-making model, as determined by the Cognitive Approach to School Leadership, to consider the interaction of the situation, the environment, and the organization to respond optimally.

Figure 1. The Cognitive Approach to School Leadership Within a Contingency Framework



Note. The administrator uses the decision-making (DM) model to respond to contingent events. Prob. Interp. = Problem Interpretation.

Contingency theory of educational administration. According to Contingency Theory, there is no single best way to make decisions. In fact, contingency theory posits there is no single best way to lead, manage, make decisions, and solve problems (Reyes, 2006; Hoy & Tarter, 2008; Tarter & Hoy, 1998). Rather, because of the unique features of local context and the changing nature of organizations and people over time, the best solution is the one that best fits the situation, environment, and organization. For example, a principal may rely on a personal relationship with a parent to address a student discipline problem in one instance; whereas, in another instance, all other things being equal, the principal may have to follow formal procedures if he or she does not have a personal relationship with the student or parent. Because of the idiosyncratic aspect of whether a personal relationship exists or not, the principal must act contingently – based on the situation, environment, and organization.

Contingency theory evolved to address weaknesses in previous administrative theories. Historically, administrative theory focused on three main strands: organizational, managerial, and bureaucratic (Barbour, 2006). These strands weave together the notions that organizations have specific goals and functions, that people's efforts should be coordinated to meet those goals and functions, and that a bureaucratic division of labor and authority is the most efficient method for coordinating those efforts (Barbour, 2006). For example, schools have the goal and function of educating all school-aged children. To accomplish this goal, multiple groups of people need to coordinate efforts - teachers, principals, teachers' aides, school support staff, and other groups of people are all needed to make a school function. A bureaucratic division of labor and

authority allows those different groups to understand their roles and responsibilities by allocating decision-making authority and tasks to different participants.

Over time, those three strands (organizational, managerial, and bureaucratic) have shifted in terms of how people and their relationship to the organization are conceived. Historically, organizations were viewed mechanistically, where people were interchangeable. For example, managers thought that members of an assembly-line could be substituted without any effects on the organization or final product. More recently, however, aligned with social learning theory, organizations and people are seen as influencing each other, and Contingency Theory has developed to address the interaction of the organization, the environment, and the situation. For example, the organization, the environment, and the situation must all be considered when deciding what to do when a school safety officer thrashes a student of color and the video is posted on the internet. The organization's culture, procedures, and policies influenced what took place; the macro and local socio-political environment influenced what took place, and the individual circumstances of the situation influenced what took place. Thus, the administrator should consider the event through the lens of the three factors.

Contingency theory informs this study in a very fundamental way. Because there is no single best way to make decisions (Hoy & Tarter, 2008), assessing responses as correct or incorrect belies the complexity of the administrative reality. Further, suggesting there is a single best answer implies that researchers can pre-determine that best answer. This suggestion then ignores the fact that best answers vary across contexts. The best response to a given situation in urban Portland may not be the best response to the same situation in rural Eastern Oregon. Instead of developing numerous versions of

ADMin-SD for urban, suburbran, and rural contexts across the regions of the United States, ADMin-SD employs a constructed-response format, which requires the use of expert judges to assess performance. Although recruiting from a local population is more selective and challenging than recruiting from a national pool, I view the use of expert judges from local contexts to assess local responses as a strength rather than simply as a limitation. With local judges, responses will always be scored against the local standards, which is how judges can estimate how effective or feasible the response would be, for example; these are the same standards that the administrators are judged against in their every day practice; thus, the local judges help maintain the ecological validity of the measure across contexts. As Contingency Theory helps define the context and structure of ADMin-SD, the Cognitive Approach to School Leadership provides the theory and content that defines what should be measured, which is discussed next.

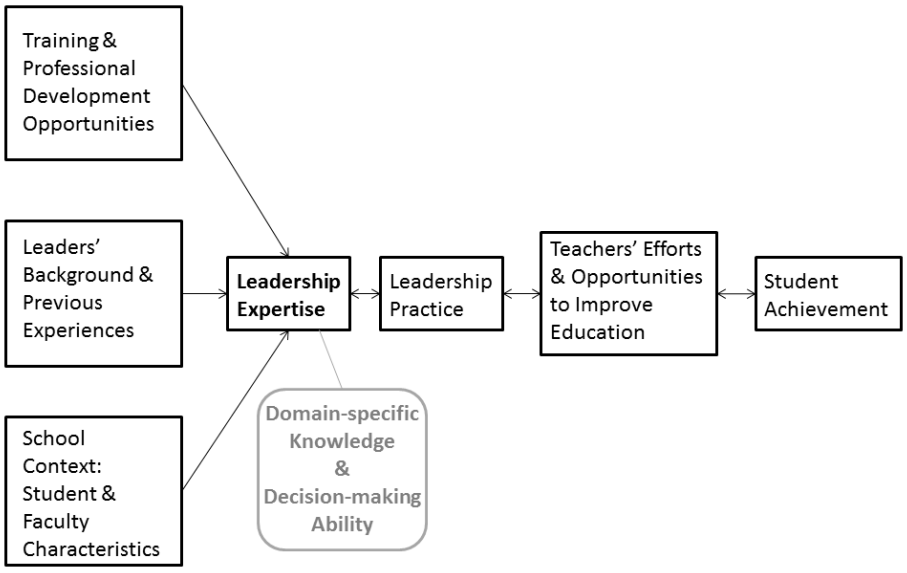
Cognitive approach to school leadership. Recognizing the importance of a contingent environment in making effective decisions, this study also draws on theory generated from empirical findings on principals' thought processes when making decisions. This study is rooted in a line of research that diverged from previous studies of educational leadership that tended to focus on leaders' behaviors, by instead focusing on leaders' thought processes. With an exclusive focus on observable behaviors, previous leadership theories could not answer why educational leaders performed those observable behaviors (Hallinger, Leithwood, Murphy, 1993). This shift in focus from behaviors to thought processes reflects the underlying assumption that what leaders think is generally related to what they do. This line of research worked on the assumption that leaders' decision-making ability is a central cognitive activity of the job and, consequently, a

characteristic that distinguishes expert from typical principals (Hallinger, Leithwood, Murphy, 1993; Leithwood & Stager, 1989).

The theory further hypothesizes that leaders with more expertise and knowledge should be able to work with and through other school staff to improve their schools (Goldring, Huff, Spillane, & Barnes, 2009). Expertise is a personal characteristic of the principal that interacts with organizational and contextual factors to work through teachers to impact students (Goldring et al., 2009). Goldring and colleagues (2009) provide the clearest logic model that situates the cognitive approach within the broader school ecosystem. Figure 2 shows that leadership expertise is one of several important factors that indirectly influence student achievement. The following sections discuss the theoretical issues involved in assessing expert decision-making and how those issues inform the proposed study.

Expertise. Relying on research into the nature of expertise, the Cognitive Approach to School Leadership posits that administrators' expertise is due to a combination of one's knowledge and one's decision-making ability. Knowledge is broken into two types: declarative and procedural (Ohde & Murphy, 1993). Declarative knowledge includes facts, concepts, principles, and their interrelationships; whereas, procedural knowledge reflects an understanding of how to apply the declarative knowledge (Ohde & Murphy, 1993). Knowledge is then organized into a schema (Anderson, 1982). Schemas connect previously unrelated facts into a coherent picture, increasing automaticity and eased access to the now-organized information, which reduces cognitive load and allows the expert to attend to other aspects of the problem.

Figure 2. The cognitive approach to school leadership situated in the broader context



Note. Figure taken from Goldring et al., 2009, p. 200. ADMin-SD assesses the constructs in the grey box which breaks down the components of leadership expertise.

Experts possess more knowledge and more refined knowledge schemas, and they make more connections among schemas and stimuli in the environment. They are more sensitive than novices to unique aspects of events that occur commonly, and they use those idiosyncrasies in their decision-making. Experts also identify and specify problems and solutions faster than novices and with fewer errors, partially due to better domain-specific pattern detection and domain-specific recall memory (Chi, Glaser, & Farr, 1988). For example, chess masters solved problems faster, with fewer errors than novices. But when they were given random positions, positions that were not common chess positions, that is, positions for which they did not have a schema, they performed as well as novices, demonstrating that expert performance requires domain-specific knowledge. Simply put, expertise is domain-specific (Chi, Glaser & Farr, 1988); e.g., expert plumbers are not necessarily expert photographers.

Ill-structured problems. Based on a key finding from the seminal study in this line of research, the cognitive approach relies on the use of ill-structured problems to discriminate expert from novice responses (Leithwood & Stager, 1989). This finding makes sense because, in real-life, most of the important problems people face are ill-structured in nature (Fredericksen, 1983). Problems are ill-structured when they require the decision-maker to bring structure to the problem by defining the problem and what should be done about it (Simon, 1973). That is to say, the decision-maker must shape and refine his/her sense of the problem and goal while determining solutions to achieve that goal (Klein & Weitzenfeld, 1978). For instance, in the situation where the Black student had a negative physical interaction with the school safety officer, one administrator may assess the problem to be one of bad training. A second administrator may think the student should have complied with the officer's requests immediately, and a third administrator may identify the problem as deep-rooted, institutional racism. Thus, an administrator's interpretation of an ill-structured problem provides insight into their thought process, including their values and their practical understanding of schools. Ill-structured (a.k.a., *ill-defined*, *swampy*, or *messy*) problems usually have more than one correct answer and being correct is more a matter of degree than the language of correct and incorrect suggests. It then follows that researchers must define the continuum of proficiency pertaining to these kinds of problems through the performances of identified-experts and novices. Thus, the field has adopted a research paradigm in which experts are compared with novices to determine the traits that constitute expertise.

Sampling. Identifying experts and novices is critical in this research paradigm. Defining experts has usually used two approaches: absolute and relative (Chi, 2006).

Absolute experts can be identified with objective, performance measures; this approach tends to take a trait-view of expertise in which people are endowed with the characteristic (Chi, 2006; McFall, 1982). For domains and situations without explicitly defined criteria – domains marked by ill-structured problems – the study of experts has taken a relative approach and defines experts on a continuum in comparison to novices. This approach assumes that novices can perform in such a way that they can become experts and tends to take a skill-view of expertise in which the ability is produced by skills that can be learned (McFall, 1982). Thus, a goal of the relative approach is to determine how to enable less skilled individuals to become experts (Chi, 2006). Because ill-structured problems do not have guaranteed correct answers, the relative approach must be used in developing and validating ADMin-SD. In the relative approach, experts have been identified by academic qualifications (e.g., graduate students vs. undergraduates), seniority or years performing the task, or consensus among peers (Chi, 2006).

The model. The line of research begins with Leithwood and Stager's (1989) qualitative, grounded theory analysis of 44 interview protocols during which principals were asked to solve vignettes such as the one presented below in the literature review. Based on these protocols from typical and expert principals, they determined six general categories that were reflected in principals' decision-making process: problem interpretation, goals, values, constraints, solution processes, and their mood (experts tended to be calmer while novices expressed more fear).

Briefly, problem interpretation statements reflect the principal's understanding of the specific nature of the problem, often identifying multiple potential problems. Goal statements included relatively short-term purposes the principals wanted to achieve.

Values were longer-term purposes, “operating principles, fundamental laws, doctrines, and assumptions that guided the principal’s thinking” (p. 133). Constraints were barriers or factors that narrowed the range of possible solutions, some of which could be overcome and some of which could not. Solution processes were the actions the principal described that were taken or to be taken to address the problem. They included mood as a sixth category in their analysis, which is surely important in decision-making. For the purposes of ADMin-SD, however, mood is not be assessed for practical reasons. The measure is completed online, in laboratory-type conditions, where one’s mood would be expected to be neutral. Moreover, assessing mood online, beyond self-report, does not seem feasible. It should be noted that these six components compose a model of decision-making, but the term *model* is used loosely in the sense that these are important aspects to consider when solving ill-structured problems. There is no particular order that they must be considered, and consideration is typically be iterative. As the only descriptive model of administrator decision-making, it is used as the basis of ADMin-SD.

In conclusion, the cognitive approach to school leadership posits that expert principals are strong decision-makers. In fact, empirical evidence from this pool of literature supports that theoretical position, and I review the literature that flows from the cognitive approach in the following section because it provides the substance that makes up ADMin-SD; whereas Contingency theory provides a macro view within which the cognitive approach operates.

Literature Review

I review the empirical studies of principal decision-making with a focus on methodological issues and the substantive findings. As a measure development study, I

provide a comprehensive review of the literature to ensure appropriate representation of the content domain. Following this review of the literature base, I review the predominant approaches to assessing decision-making and problem-solving.

How the search was conducted. To identify the empirical studies of administrator decision-making, I conducted a search of the PsychInfo, Academic Search Premier, and ERIC databases for peer-reviewed articles written in English in education journals with different combinations of the terms "principal," "problem-solving," "decision-making," "problems," "measure(ment)," "ill-structured," "student discipline," and "expertise." After reading the titles and abstracts, including only empirical studies, 23 articles and eight conference papers were retained for review.

After reviewing all 31 citations, I excluded two studies that focused on principals managing a team or group of people charged with solving student instructional or behavioral problems. That is, those studies focused on managing and facilitating group conversations rather than examining the decision-making thought processes involved in identifying problems, generating alternatives, and making decisions. I excluded six more articles and eight conference papers because they did not meet the following two criteria: (a) published in a peer-reviewed journal and (b) focused on the thought processes of principals (or asst. principals) in problem-solving and decision-making. The final literature pool for this review contains 15 journal articles.

Results of the review. According to the search conducted, there are not many empirical studies of principal decision-making in the English-written literature base. Although interest in school administrative decision-making goes back almost 60 years (e.g., Hemphill, 1958), rigorous empirical work did not begin until the 1980's when

Leithwood and colleagues began to examine the construct. Despite some conflicting results, most of the studies congeal into a cohesive literature base. The studies share relatively similar conceptions of decision-making and have used a set of similar methods and measures, but the same topic was rarely addressed twice. The literature presents some relatively uncommon methodological features that are important for understanding the substantive findings, which provide the foundation for the design of this study. See Appendix A for a table summarizing the studies and the relevant issues.

Review of methodological issues. Due to the difficulties of assessing decision-making performance on ill-structured problems, these studies use somewhat uncommon sampling techniques, item types, and scoring methods. As a result, in the following sections, I review how experts and novices were identified, the types of problems that were used across the studies, and how responses to written vignettes were scored to assess decision-making performance.

How experts and novices were identified. Of 12 studies, expert identification procedures were used in nine and can be categorized into the following four patterns. The first, and weakest, pattern was the use of one or two central office administrators to nominate highly effective or expert principals, an approach used in four studies with a few variations (Bullock, James, & Jamieson, 1995; Leithwood & Steinbach, 1993; Lazaridou, 2007a, 2007b, 2009; St. Germaine & Quinn, 2005). This approach, based on principals' reputation, confers the likely benefit of increased efficiency in expert identification, but it is subject to personal biases, especially if only one or two central office administrators are queried. Although this approach has been used across twenty years of studies, other studies found ways to improve it.

The second pattern supplements the reputational approach with a survey designed to assess leadership characteristics. This approach was used in two studies by Leithwood and colleagues. For example, in the seminal study, Leithwood and Stager (1989) relied on a two-step process that combined the reputational approach with scores derived from an interview protocol they had developed in prior research. They first asked two central office administrators in each of three districts to indicate independently which of all their principals they would recommend as highly effective. Only principals endorsed by both administrators made it to a second screening, which consisted of an interview (Leithwood, 1987) designed to assess principals' self-reports of how they lead their schools. Principals who obtained high scores (3s and 4s out of 4) were then designated *experts* (6 out of 22) and the remaining 16 principals were designated as *typical*.

The third pattern was used once, by Allison and Allison (1993). They did not use a reputational approach to avoid a halo effect in data collection and analysis. Instead, they used years of experience in role to demarcate novices from experts by recruiting elementary school teachers who had just earned the credentials to be principal but had not become one yet (deemed *rookies*), elementary principals who had 10-15 years of experience in the role (*seasoned*), and elementary principals with 20+ years of experience (*veterans*). They failed, admittedly, to capture any true novices when they discovered their rookies had an average of 9.4 years of experience as vice principals. As a result, they recruited 10 students from a graduate education program and deemed them *entrants*. Their two most experienced groups outperformed the two least experienced groups on average, but the *seasoned* group outperformed their *veterans*. They found that years of experience as principal did not predict performance linearly, but years in schools did.

The fourth pattern was employed in one larger study that produced two articles (Brenninkmeyer & Spillane, 2008; Spillane, White, & Stephan, 2009). Also avoiding a reputational approach, these researchers used teacher surveys about their principal's leadership characteristics to identify principals whose scores showed growth over three sequential time points (1997, 1999, & 2001) relative to other principals in the sample. Collected as a matter of district policy, the surveys had a response rate of about 75%. After excluding principals with inconsistent tenure over this time and schools with high student mobility rates, they labeled principals whose scores were rising as *experts* (n = 20) and those whose scores were flat or declining as *typical* (n = 16). Further, they used standardized test scores to confirm that the expert principals were leading schools whose test scores were rising relative to the sample. This analysis supported their teacher survey results, but the difference between groups was not large enough to achieve statistical significance. The second article from the study (Spillane et al., 2009) did not use the typical principals in their study. Instead, to compare with the group of 20 experts, they recruited 24 *aspiring* principals from a principal training licensure program.

These studies generally concluded that administrator experience does not equate to expertise in administration because the groups and individuals with more experience tended to demonstrate more expertise, but not always. Resolving that experience was not the sole source of expertise, the literature base invokes explanations like Kennedy's (1987) that experience contributes to expertise only if practitioners are able to learn from their experience. Across the studies that used an expert/novice paradigm, they all acknowledged the limitations in how they identified their experts and non-experts. The nomination procedure, while likely efficient, is inherently susceptible to various biases,

including who the central office administrator may happen to like, have a relationship with, or had the opportunity to observe professionally. As well, the line between novices and experts was blurred due to reasons of attrition, incorrect initial identification, and small sample sizes. Further, the novices that displayed expertise may not have been true novices in that they could have been individuals with several years of experience as teachers or asst. principals (e.g., Allison & Allison, 1993).

Examining the link between principal expertise and decision-making is the basic thrust of the Leithwood line of research; they failed, in my mind, to conclusively show that principals must be expert decision-makers to be expert principals. However, it does appear that making decisions is a central activity and skill in being an effective or expert principal. It may well be that decision-making skills distinguish highly effective principals from less effective principals, but more studies and better measures are needed to resolve that question, further spotlighting the need for the development of ADMin-SD.

Table 1. Expert/Novice Identification Procedures

Patterns	**# of studies**
1. Reputational	6
2. Reputational + Survey	2
3. Years of Experience	1
4. Teacher Surveys + Student Test Scores	2

Note. The rest of the studies did not use an expert/novice paradigm.

Types of problems. In this research paradigm the type of problem used to investigate principals' decision-making is critically important. The types of problems used in these studies take two general forms: either a problem from the principal's own practice (either current or past) or a researcher-created vignette presented to them. Use of vignettes allows for comparison across principals because the items are standardized across respondents. A total of six studies used principal-generated problems that were

either current problems they were working through or past problems they had solved well or poorly. These principal generated problems offer the benefit of knowing what the principal actually did to address a real problem.

In the earlier studies by Allison and Allison (1993), Leithwood and Stager (1989), and Leithwood and Steinbach (1992, 1993), they used "general administrative" vignettes that ranged in topics from problems with a school library to setting school objectives. Allison and Allison's (1993) single vignette was quite long (~750 words) and provided a lot of context. In comparison, Leithwood and Stager (1989) used vignettes that were short (<100 words). As an example, Leithwood and Stager (1989) used the following short vignette in their seminal study:

> Your new school is one in which staff have never been involved in the setting of school objectives and are not apparently very interested in doing so. You have come to believe that it is a *very* important thing for staff to set school objectives and to evaluate them at the end of the year. (p. 134, italics in original)

They found that expert and novice principals tended to respond similarly on well-structured problems, but ill-structured problems generated notable differences in responses.

Future research built on this finding by focusing on principals' performance on ill-structured problems. Spillane and colleagues (2008, 2009) examined performance on instructional problems in math and reading like this one:

> A large number of the elementary teachers in your school have admitted to you they are not comfortable teaching mathematics. Your mathematics test scores demonstrate a weakness in this area. However, the school district in which you work uses both mathematics and literacy test results to determine how well a school is doing academically. How will you address this situation?

Brenninkmeyer and Spillane (2008) found different patterns of responses based on the content of the problem. Specifically, they found that principals relied on follow-up

meetings when dealing with math problems, which the authors believed indicated principals' understanding that improving the math curriculum cannot be fixed with a single meeting or solution. For literacy problems, they found that experts were more likely to rely on successful anecdotes than typical principals, which the authors struggled to explain other than to say that experts preferred to use anecdotes to solve literacy problems. With such a weak explanation, this finding may also simply be an artifact of the dataset though it does align with extant results.

Table 2. Type of Problems

Authors	**Year**	**# of Vignettes**	**Content of Vignettes**
Leithwood & Stager	1989	6	"General Administrative"
Leithwood & Steinbach	1992	4	"General Administrative"
Allison & Allison	1993	1	Library Problem
Brenninkmeyer & Spillane Spillane, Stephan & White	2008 2009	6	"Instructional Problems"
Lazaridou	2007a, 2007b, 2009	5	"Human Relations" & "Strategic"
Goldring, Huff, Spillane, & Barnes	2009	3	"Instructional Problems"

Lazaridou (2007a, 2007b, 2009) presented five ill-structured problems that ranged from a conflict between a classroom teacher and an assistant to the impact of financial constraints and low enrollment on staff policies. One vignette involved disciplining a student after the student used abusive language toward a teacher. Lazaridou developed these vignettes based on principals' prior experiences, but she did not report content validation results. Goldring and colleagues (2009) also modeled their vignettes on those of Leithwood and Stager (1989) and Brenninkmeyer and Spillane (2008) and similarly did not report technical adequacy. After reviewing the literature, it is important to note

that no study in this line of research has yet examined principal decision-making systematically in the context of student discipline problems save for one vignette in the Lazaridou studies (2007a, 2007b, 2009).

Scoring decision-making. All the studies in the literature pool used think-aloud techniques (Ericsson & Simon, 1980), interviews, or written responses to elicit decision-making performance. Because ADMin-SD uses open-ended responses to written vignettes, this section focuses on how the literature scored these kinds of responses. A total of five studies used researcher-generated vignettes to elicit decision-making thought processes; one study scored responses to principal-generated problems. Only the studies that scored responses for quality of response are reviewed in this section because effectiveness is a focus of ADMin-SD. This decision excludes the articles by Lazaridou (2007a, 2007b, 2009), Spillane and colleagues (2008, 2009), and Goldring and colleagues (2009) as these studies did not score responses for quality. For example, Spillane and colleagues (2008, 2009) coded decision-making protocols for usage of the different processes associated with expert and typical principals. In their study, for example, *telling a successful anecdote* was an expert process for which they coded the protocols. The protocols were not scored for likely effectiveness or expertise or other outcome variables that are typically assessed in decision-making and problem-solving research.

Of the three remaining studies, two of them scored verbal reports that were transcribed; one of the studies scored written responses. First, Leithwood and Steinbach (1992) scored written responses to test the effects of an instruction program designed to improve decision-making abilities. Two judges rated responses to four vignettes (2 pre-test, 2 post-test). They rated the responses holistically, which is to say they gave a global

rating to the whole response and did not parse the response into segments. They gave each response two ratings of 0-3 (very poor to very good). The first rating was for the "thoroughness of the process or the quality of the thinking" (p. 335); the second rating was for the "quality of the solution or the product" (p. 335). They did not report reliability between the two judges.

Second, following this study, Leithwood and Steinbach (1993) scored transcribed verbal responses elicited during interviews about current school improvement problems. They did not find a strong relationship when they correlated these scores with teachers' survey responses on these principals' transformational leadership characteristics. They were hampered, however, by a small sample size of nine principals. Nonetheless, they scored principals' responses by coding for use of expert processes across the six components of their model (problem interpretation, goals, constraints, values, solutions, and mood). They coded for quality of use of these processes by assigning scores of 0-3. A score of 0 meant there was no use of the skill/process; 1 meant there was some indication of the skill being used. 2 meant the skill was demonstrated, and 3 indicated that the skill was used frequently or a "particularly fine example of the skill" (p. 320). This last rating introduces multi-dimensionality to their scale when it should be unidimensional. From 0 – 2, the scale reflects a frequency of skill use, but a score of 3 reflects frequency or quality. Scores were summed to provide a quantitative measure of the principal's *process* in response to their current school improvement problems. This quantitative measure was based on more scores, attending to a problem in the previous measure's construction. It is unclear why they did not separately rate the actual quality of the thinking.

Last, taking a different approach, Allison and Allison (1993) measured decision-making with responses to a longer vignette. They scored responses for level of abstraction, attention to detail, and effectiveness. To code for attention to detail, they preselected (un)important details in the vignette and then coded responses for discussion of these details. To code for level of abstraction, they coded principals' goals. Concrete goals were related to physical objects and personnel; abstract goals were programmatic and transformational in nature. Lastly, they scored responses holistically for expertise on a scale of 1-10. According to their coding scheme and judges, they found greater levels of abstraction and greater attention to relevant details were positively related to judged expertise. Using a long vignette, they provided a lot of context to each principal, thus structuring the problem for the principal more than the shorter vignettes do, which does not allow principals to impose their values on the situation as much as the shorter, ill-structured vignettes. As well, preselecting details that are important – to these researchers – may be details that are unimportant to their respondents in their local context. Further, they did not report results of technical adequacy studies with respect to these issues.

Table 3. Approaches to Scoring Effectiveness in Responses

Scoring Method	**Data Collection**	**# of studies**
1. Appraisal of Effectiveness, Coherence, Thoroughness	Verbal Report	1
2. Coding for demonstration of skill – Frequency and Quality scale	Written Response	1
3. Judged expertise, attention to detail, goal abstraction	Verbal Report	1

Overall, the literature has used a few ways to measure principals' decision-making, but each approach demonstrated substantial shortcomings including use of problematic scales and structured vignettes, as well as possibly questionable content

validity and reliability. Moreover, despite creativity's importance in responding to ill-structured problems, none of the studies attempted to measure it even when giving global appraisals of the responses. In sum, the literature presents four methods for sampling, several types of problems, and three ways to score responses for effectiveness. Having described these methodological issues, the following findings can be put into context.

Review of substantive issues. Despite rarely discussing the same topic twice, results from the literature can be categorized into four themes: differences between experts and novices, principals' personal characteristics, the skills used in decision-making, and the overall construct.

Differences between experts and novices. Across the components of their model (i.e., problem interpretation, constraints, values, goals, solutions, and mood), Leithwood and Stager (1989) found differences between their expert and typical principals. Table 4, above, presents a summary of these findings. Using the research on expertise (e.g., Chi, Glaser, & Farr, 1988) and Leithwood and colleagues' research as a guide, researchers have found differences between experts and novices (or another comparison group, typical principals for example). The next sections describe these differences within the context of Leithwood and Stager's (1989) model of decision-making.

Problem interpretation. In general, experts do a better job identifying and interpreting problems they face. They tend to spend more effort on problem interpretation and less on solutions because as problems become clearer, their solutions become more obvious. To illustrate the point, if one brings a car to a mechanic and says it's broken, the mechanic may not know where to start. In comparison, if one tells the mechanic that the transmission fell off when going over a speed bump, then the solution is obvious: replace

the transmission. The better the problem can be specified, the clearer the solution should become. Experts know that specifying the problem reaps more benefits, all things being equal, than spending time on solutions.

Table 4. Differences between experts and novices

	Experts	**Novices**
Problem Identification	• more if-then reasoning • focus on the consequences of the problem for the school, students, & programs • recount relevant & successful anecdotes • desire to collect information to understand the problem • more detailed, abstract, and comprehensive interpretations	• less if-then reasoning • focus on the consequences as related to themselves and their staff • recount irrelevant or unsuccessful anecdotes • made assumptions in lieu of information
Values	• Use of values in lieu of information • Use of values in lieu of org policy • Explicit use of values	• Did not use values explicitly
Goals	• Wanted to keep parents informed • Goals focused on school, students, & programs • More abstract goals	• Wanted to keep parents happy • Goals focused on staff • More concrete goals
Constraints	• Consider constraints when planning • Faces up to conflict	• Viewed some constraints as unsolvable • Prefers to avoid conflict
Solutions	• Spends more time planning, gathering data, building support for solutions • Relied on delegating appropriately • Wider repertoire of responses • Stressed the importance of following up on solutions	• Spends less time planning, collecting data building support • Uncomfortable delegating

Specifically, expert principals tend to provide more detailed, abstract, and comprehensive interpretations of problems by demonstrating more if-then reasoning,

while novices did not (Brenninkmeyer & Spillane, 2008; Leithwood & Stager, 1989; Spillane, Stephan, & White, 2009; St. Germaine & Quinn, 2005). While interpreting problems, expert principals tended to focus on the consequences of the problem for the school, students, and programs, whereas novices tended to focus on the consequences as related to themselves and their staff (Leithwood & Stager, 1989). It should not be surprising though that someone newer to a job is more concerned with keeping their job than someone who has more security in the position. Additionally, expert principals recount relevant and successful anecdotes from their practice to help interpret and understand current problems they face (Brenninkmeyer & Spillane, 2008; Lazaridou, 2007b; Leithwood & Stager, 1989; Spillane, Stephan, & White, 2009). As well, expert principals expressed a desire to collect information to understand the problem, while novices made assumptions in lieu of information (Leithwood & Stager, 1989). More recent research (Brenninkmeyer & Spillane, 2008), however, did not find that experts tried to collect data more than typical principals. These conflicting results may be due to the growth of a culture of data usage that occurred during the years between the studies.

Goals. Leithwood and Stager (1989) found that experts and typical principals differed in the goal statements they made. Experts tended to focus goals on students and programs, while keeping parents informed. Novices shared more staff-oriented goals and wanted to keep parents "happy." Allison and Allison (1993) found that their experts tended to conceive of goals in more abstract terms (i.e., programmatic and/or transformational vs. physical and/or related to personnel), which aligns with experts' more abstract interpretations of problems.

Values. As a personal characteristic, this topic is described in more detail in the next section. In terms of specific differences between experts and novices discussed in the literature, experts tended to consider more principles or values and use them as basis for determining longer-term goals. Principles or values may be reflected in statements such as: *Teachers deserve a safe work place*. Typical principals did not make these kinds of statements, demonstrating that they were not thinking at this level of abstraction. Further, expert principals rely on their values when they have incomplete information. Thus, if they have to make an assumption, they let their values guide them (Begley & Leithwood, 1990; Lazaridou, 2007b; Leithwood & Stager, 1989).

Constraints. Experts hardly indicated constraints as such; instead, they built the consideration into their solution process. For example, rather than viewing community opposition to a school consolidation problem as a constraint, an expert principal simply included mechanisms to give the community opportunities to voice their opinion and obtain information on the issue. Rather than viewing the opposition as a constraint explicitly, the experts folded it into their conception of the problem and solution, so experts did not state many constraints explicitly. When they did, however, they found ways to deal with those constraints. Typical principals, however, saw more constraints and tended to view them as potentially unsolvable. For example, one principal was discussing resources as a constraint: "it may be too much of a drain on the resources of the school, and we may not be able to handle all they think we can" (p. 148). When conflict was viewed as a constraint, expert principals were more adept at managing and facing up to it; whereas, novices understood the importance of communication but preferred to avoid conflict when possible (Bullock, James, & Jamieson, 1995).

Solutions. In terms of devising solutions, experts spent more time in planning a response, gathering information/data, and garnering support for the eventual solution; whereas, typical principals tended to seek extra information when they were not sure what to do (Leithwood & Stager, 1989). For example, two typical principals said they would seek counsel from the superintendent before doing anything else. In their solutions, experts are more likely to delegate: they identified clear strategies for delegating tasks, knew which kinds of tasks to delegate, and were comfortable with transferring their authority to others (Bullock, James, & Jamieson, 1995; Brenninkmeyer & Spillane, 2008). Novices, however, were uneasy about delegating, did not want to overburden their colleagues, and were less comfortable about transferring their authority (Bullock, James, & Jamieson, 1995). Experts stressed the importance of following up on solutions, while novices did not (Spillane, Stephan, and White, 2009), and experts had a wider repertoire of responses to unanticipated obstacles (St. Germaine & Quinn, 2005).

Personal characteristics. A total of four studies explored how different aspects of principals' personal characteristics influence their decision-making. Specifically, the studies explored principals' values, their personal and professional biographies, and dimensions of their personality. Although the literature has delineated several differences between experts and novices, some specific personality characteristics do not differ between groups of experts and typical principals. Brenninkmeyer and Spillane (2008) administered personality measures to their expert and typical principals intended to assess extraversion, conscientiousness, and emotional stability. There were no differences between groups along any of these personality measures.

Principals' values, however, help guide their decision-making, particularly when information is lacking or in the absence of relevant, organizational policy (Lazaridou, 2007b; Leithwood & Stager, 1989) and can change over time (Begley & Leithwood, 1990). Lazaridou (2007b) found principals used 7 distinct values in their decision-making, from least to most: confidentiality, nurturing, fairness, personal effectiveness, collaboration, consideration, and mission. Having a mission, and its related supports (e.g., aligning others behind goals, maintaining communication, being responsible to the public), was the most prominent value expressed by her sample of principals. For example, one principal expressed the importance of having everybody on board and moving in the same direction; the principal thought the student (and parent) experience would be better with that cohesion in place.

Beyond their values, principals' personal and professional biographies have been implicated in the decision-making process. Sleegers, Wassink, van Veen, and Imants (2009) conducted a small qualitative study with two new principals by interviewing them about problems they faced in their first two years and by observing them in practice. They selected these two principals because several of their demographics matched, but more importantly they faced similar problems. These authors noted that the differences were mainly in how they tried to solve the problem, rather than in how they interpreted it. Coming from a hierarchical world of business and politics, one principal relied on a top-down strategy. Rising through the education ranks, the other principal, however, used a bottom-up approach because he stated he valued autonomy and teachers taking professional responsibility for their development. With a sample of two, the findings are impossible to generalize, but they provide evidence and a description of how one's

values, personal, and professional background can influence principal's decisions. Table 5 presents a summary of the personal characteristics studied by this line of research.

Table 5. Personal characteristics reviewed

Characteristic	Findings
Personality traits of extraversion, conscientiousness, & emotional stability	No differences between expert and typical Ps
Values	• Ps rely on values in lieu of information or organizational policy • Ps rely on values when making assumptions
Personal & Professional Background	Impacts framing of solutions more than interpretations of problems.

Note. Ps = Principals

Related to the skill of decision-making. A total of four studies focused on factors related to one's decision-making skills. These studies have investigated the use of archetypal strategies, types of knowledge used, and attention to detail and goal abstraction. For example, Allison and Allison (1993) found their expert principals paid more attention to relevant details and espoused more abstract goals. In other words, they understood problems at a deeper level.

Representing problems and solutions abstractly requires different kinds of knowledge. This literature has demonstrated that principals use four kinds of knowledge in their decision-making, from most to least: knowledge of the organization, knowledge of the people involved, tacit knowledge, and knowledge of the task (Lazaridou, 2009). Knowledge of the organization involves its policies, procedures, mission, resources, etc.; knowledge of the people involved includes their strengths, weaknesses, preferences, responsibilities, etc.; and knowledge of the task includes an understanding of what the problem is and how to solve it. Tacit knowledge is squirrelly defined as "the kind of

knowledge that allows administrators to recognize previously encountered macro-patterns and to respond to them with 'rules of thumb'" (Lazaridou, 2009, p. 8). St. Germaine and Quinn (2005) defined tacit knowledge as: practical wisdom, intuition, and/or "knowledge that is bound up in the activity and effort that produced it" (Horvath, 1999, p. ix, as cited by St. Germaine & Quinn, 2005). Expert principals used tacit knowledge more than the novices (Lazaridou, 2009; St. Germaine & Quinn, 2005), which allowed the experts to exhibit better timing in their problem interpretation and solution generation processes. Novices either concluded their decision-making too soon or too late; additionally, they implemented actions without enough preparation or after waiting too long (St. Germaine & Quinn, 2005).

One small qualitative study (Leithwood & Steinbach, 1990) tried to make tacit knowledge explicit by conducting interviews with 11 identified expert principals. They interviewed them directly about how principals determined the priority of problems they faced, how they determined the difficulty of problems, and how they determined whether to involve others in the decision-making process. They found principals determined priority based on eight factors: number of staff capable of handling the problem, number of people involved, the content, the time frame, the fit with their conception of role, relationship to long term plans, perceived importance by others, and avoidance of problem escalation. They determined problem difficulty based on seven factors: availability of clear procedures, impact on staff morale, number of people required to solve, likelihood of value conflicts, likelihood of solution all can accept, type of people affected, and degree of control over solution. Lastly, they determined whether to include others based on six factors: the problem difficulty, the time available, the importance of

finding the best solution, the amount of relevant knowledge possessed by others, the problem's impact on others, and the need for ownership of the problem and solution.

These factors overlap and we see common threads throughout. The number of people involved, who they are, what their interests are, and other social considerations (e.g., staff morale) are evident in these lists. Time is another factor: if time is short, they may not be able to involve others to the degree they would if they had more time before a solution was needed. Although again based on a small number of interviews, these findings provide us some concrete factors that principals consider when determining priority, difficulty, and how to involve others.

Using different kinds of knowledge, principals use three archetypal strategies (decomposition, conversion, and reversion; Voss, Green, Penner, & Post, 1983) and possibly a fourth that is *solution-oriented* (Lazaridou, 2007a). In decomposition, experts break larger problems into smaller, easier sub-problems. In conversion, experts convert harder problems into a different kind of problem. For example, a problem of student discipline may be converted into a problem of inadequate professional development. Experts use reversion to identify and eliminate factors that contribute to the problem either before, during, or after the problem has occurred. Lazaridou (2007a) found evidence that the expert principals in her study used these strategies; however, they used a fourth strategy to a greater extent, labelled *solution-oriented*. When using this strategy, the principals in her study immediately started stating their solutions and plans. The emergence of this category in her analysis contradicts prior research in that experts spend more effort on problem interpretation than on planning solutions. Because the principals in her study acted more like novices than experts in this regard, I return to her sampling

procedures. She used a reputational approach, asking only one administrator to recommend highly effective principals, so I interpret this finding with caution.

The literature base has examined different skill-related aspects to making effective decisions, including paying attention to detail, forming abstract goals, and using different kinds of knowledge and archetypal strategies. These findings are still disjointed and do not present a comprehensive model of the thought processes, or cognitive skills, needed to make effective decisions, which further highlights the need for a standardized measure to help build empirical evidence on these and other topics. Table 6 presents a summary of these findings.

Table 6. Findings related to the skills used in making decisions

Topic	**Findings**
Use of archetypal strategies	• Decomposition • Conversion • Reversion • Solution-oriented
Use of different kinds of knowledge	• Knowledge of organization • Knowledge of the task • Knowledge of the people involved • Tacit knowledge
Attention to detail	Expert responses paid more attention to relevant details
Goal Setting	Expert responses incorporated more abstract goals

Related to the construct. The last theme of the literature review gives shape to the overall construct by addressing whether the skill can be taught, how it is related to leadership ratings, and how it should be measured. Evidence suggests that teaching the multi-component model of decision-making can result in gains from pre-test to post-test (Leithwood & Steinbach, 1992). According to their scoring, the experimental group showed bigger improvements than the control group but not always big enough to achieve statistical significance.

Two studies also suggest that highly rated leaders may be linked with strong decision-making performances (Brenninkmeyer & Spillane, 2008; Leithwood & Steinbach, 1993). First, Leithwood and Steinbach (1993) conducted a quantitative study linking leadership characteristics as reported in teacher surveys with decision-making performance on vignettes, but they did not find a clear relationship due to small sample size. Second, due to a methodological feature of their study, Brenninkmeyer and Spillane (2008) correlated principals' use of decision-making processes with scores on organizational and leadership measures that designated their principals as experts or typical. Expert processes included for example delegating tasks and responsibility. And indeed they found that principals who were classified as experts were positively related to the use of expert processes ($r = 0.36$, $p = .03$) and negatively related to the use of typical processes ($r = -0.37$, $p = .025$). This finding provides the most robust support for the notion that more effective principals use more effective decision-making processes.

Table 7. Findings related to the overall construct

Finding	**Authors**
Decision-making skills can be improved with instruction & practice.	Leithwood & Steinbach, 1992
Highly rated Ps (according to teacher surveys) tend to use expert decision-making processes	Brenninkmeyer & Spillane, 2008

Note. Ps = principals

Summary. The previous review has fleshed out cognitive approach to school leadership (CASL) by providing a comprehensive treatment of the methodological and substantive issues. This literature paints a detailed yet still disjointed picture. Results have offered methodological insights on sampling, writing items, and scoring responses, as well as more substantive findings on the differences between novices and experts, and insights into factors related to personal characteristics, to decision-making skills, and to

the construct. With a nascent literature base, the theory has yet to develop an assessment framework for measuring administrator decision-making as there have been few attempts to date. As a result, the following section reviews approaches to constructing assessments of decision-making from outside this theoretical framework and present an assessment framework based on this review that is designed to support evaluation of principal decision-making skills.

Review of Assessment Frameworks

I approach the development of ADMin-SD through the conceptual orientation of Evidence-centered Design (ECD; Mislevy, Almond, Lukas, 2003). ECD is a rigorous approach to constructing educational assessments to draw valid, reliable, and reasoned inferences based on performance. The approach views assessment as a special case of reasoning in which inferences are generated by observable evidence about unobservable traits or skills (Mislevy, et. al, 2003). In ECD, researchers identify the skill/attribute/construct they aim to assess, the observable behaviors that demonstrate that skill/attribute, the tasks that evoke those behaviors, and the scoring method that best captures the performance to make inferences about the target skill. Scores are aggregated and placed on a continuum of proficiency to make inferences about the respondent's proficiency. Through the ECD lens, the following review discusses the prevailing assessment frameworks for assessing the construct. Then, I present ADMin-SD's assessment framework. Appendix B presents a summary of the topics reviewed in the following sections.

How the search was conducted. I searched academic journals in education and the social sciences for literature on constructing assessments of decision-making and

problem-solving with different combinations of the key words: *assessment*, *measure(ment)*, *decision-making*, and *problem-solving*. Most of the education literature addressed problem-solving in math and science for K-12 students while the social science literature addressed decision-styles and social problem solving; I excluded literature on construction of IQ tests because of the de-contextualized nature of the items. The search yielded three applicable conceptual frameworks (D'Zurilla & Maydieu-Olivares, 1995; Sugrue, 1995; Zaccaro, Mumford, Connelly, Marks, & Gilbert, 2000). Only one (Zaccaro et al., 2000) explicitly addresses decision-making in ill-structured situations, while D'Zurilla and Maydieu-Olivares (1995) implicitly acknowledge that most of the problems people face are ill-structured. Sugrue's (1995) framework, while strong in its own ways, struggles to address decision-making for ill-structured situations because of the difficulty in defining the domain.

Upon guidance from my committee, I also searched the public websites of the large-scale assessment outfits that produce problem-solving assessments for K-12 students (e.g., Educational Testing Service [ETS], Center for Research on Evaluation Standards and Student Testing [CRESST], and Programme for International Student Assessment [PISA]). This literature offers the benefits of rigorous, large-scale assessment construction: strong arguments and precise content analysis along with reasoned and iterative development. This search yielded two frameworks for inclusion in the review (O'Neill & Schacter, 1997; OECD, 2010). I excluded manuscripts that used computer adaptive testing methods or that addressed technology-rich environments. Computer adaptive frameworks were excluded because ADMin-SD does not use those techniques. Frameworks that addressed technology-rich environments were excluded because all the

problems contained in the assessment require the use of at least one specific piece of information and communication software. The remaining large-scale assessment frameworks inform ADMin-SD's framework but not exclusively because they tend to address K-12 students solving problems in math and science. Usually these assessments score student responses as (in)correct, typically in the form of multiple-choice items. The following review suggests that these approaches can contribute to the assessment of administrator decision-making, but none of them can be applied exclusively.

Results of the review of assessment frameworks. With an ECD lens, I discuss how the frameworks defined the construct, the behaviors that exhibit the construct, the tasks that elicit those behaviors, and how those tasks are scored.

Defining the construct. The frameworks include four different components in their definition of the construct: knowledge, skills, meta-cognition, and motivation. Knowledge generally encompasses concepts, principles, and facts (Sugrue, 1995) and has been categorized into knowledge of the task, of the organization, and of the people involved (Zacarro et al., 2000). The skills included in the construct involve the skills associated with identifying problems, setting goals, determining constraints, identifying values, and generating solutions. In these frameworks, meta-cognition refers to planning, monitoring, and reflection. Motivation refers to self-efficacy or the degree to which one is interested in the task. Three of the frameworks include all four components to some degree; one includes knowledge and skills only (Zaccaro et al., 2000), and one includes only skills (D'Zurilla & Maydieu-Olivares, 1995). ***Defining the behaviors.*** Four of the five frameworks (D'Zurilla & Maydieu-Olivares, 1995; OECD, 2010; O'Neill & Schacter, 1997; Zaccaro et al., 2000) define the behaviors that demonstrate the cognitive

activity quite similarly. These behaviors include variations on problem identification and interpretation, generation of alternatives, and selection of the best one. Zaccaro and colleagues (2000) include solution implementation in their construct, while D'Zurilla and Maydieu-Olivares (1995) specifically do not because they believe solving a problem requires different skills than actually implementing the solution in reality. The fifth framework (Sugrue, 1995) does not define the behaviors at this level of specificity; rather, the author advocates for multiple measures to assess "behaviors indicative of" whatever skill or construct is being measured (e.g., knowledge, motivation, etc.).

Defining the tasks. The five frameworks include a wide range of tasks designed for different purposes. First, multiple-choice, sorting, and ranking items were widely recommended to assess domain-specific knowledge. O'Neill and Schacter (1997), instead, asked respondents to fill in a concept map to assess their knowledge. All five frameworks, however, advocated for performance tasks to assess respondents' ability as opposed to their knowledge. These performance tasks generally involve presenting the respondent with a problem to solve. The large-scale, computerized assessments built simulations for students to experiment with and run; performance would be measured by click-data that is recorded automatically. The clicks are then coded as processes that respondents are performing to solve the problem.

Second, the use of ill-structured and non-routine tasks was recommended for usage specifically by two frameworks (OECD, 2010; Zaccaro et al., 2000), though a third (D'Zurilla & Maydieu-Olivares, 1995) implicitly acknowledges that the difficult problems people face are ill-structured. Zaccaro and colleagues (2000) provide a detailed discussion of their use of ill-structured tasks to assess respondents' skills. These ill-

structured tasks generally take two forms: cued and uncued. Cued items mitigate a problem typical of constructed responses: the difficulty raters have in extracting the appropriate information to score. The use of simple cues (e.g., how do you identify the problem? or what goals would you set?) prompts respondents to demonstrate these sub-skills. One could argue, however, that the respondent would not have demonstrated the skill without the prompt, so Zaccaro and colleagues (2000) recommend supplementing these prompted items with unprompted items in which respondents are simply presented an ill-structured vignette and asked to respond, thereby eliciting how they would perform on their own. Unprompted items are placed before the prompted items to avoid cues influencing responses to unprompted items.

Variables used and scoring procedures. All frameworks either summed or averaged scores across dimensions and responses. Three of the five frameworks employed process and product variables in estimating respondents' scores. That is, process variables reflect the steps in the process and are scored for how well they are performed. Process variables can be scored with frequency counts of behaviors or attributed a qualitative rating of effectiveness or thoroughness. Product variables assess some dimension of the final product or solution and are usually scored as correct/incorrect, in (dis)agreement with *a priori* expert responses, or given qualitative ratings for effectiveness, feasibility, originality, and/or competence.

Two of the frameworks specifically mentioned assessing creativity, but only Zaccaro and colleagues (2000) included a measure of it. The PISA framework recognizes the importance of creativity in the problem-solving process, as evidenced by the inclusion of key concepts in their definition of problem-solving, which involves "the ability to

acquire and use new knowledge, or to use old knowledge in a new way, to solve novel problems (*i.e.* problems that are not routine)" (OECD, 2010, p. 13). Novel, ill-structured problems cannot, by definition, be solved by routinely applying available information (Baughman & Mumford, 1995). One's relevant knowledge and experience must be transformed to generate new interpretations of problems and potential solutions (Mumford, Zaccaro, Harding, Jacobs, & Fleishman, 2000). Generating new interpretations and solutions requires creativity (Mayer, 1992). Thus, solving novel problems requires creativity as using old knowledge in new ways is one of its definitions (Mayer, 1992). When reporting results, however, OECD (2010) do not report any score for creativity, even for constructed response items, while Zaccaro and colleagues (2000) do report a score for creativity.

Summary. In sum, each of the frameworks reviewed contributes to the ADMin-SD's proposed framework. ADMin-SD most closely resembles the framework espoused by Zaccaro and colleagues (2000) for several reasons. First, that framework specifically addressed decision-making for those in leadership positions and incorporates the use of ill-structured problems. As well, the variables used in that framework are most applicable to principals' reality (e.g., effectiveness, feasibility, and creativity). The other frameworks also make unique contributions. D'Zurilla and Maydieu-Olivares (1995) offer the clearest explication of how to use process vs. product variables. Sugrue's (1995) explanation of the relationship between domain-specific knowledge and decision-making skills resolves a difficulty in assessing responses in ill-structured domains by explaining how knowledge and skills are inseparable. OECD (2010) provides a rationale for

assessing creativity, even if they did not do it themselves, and CRESST offers 10 straightforward specifications to address when designing an assessment.

The previous reviews have discussed the relevant methodological features and substantive findings in the literature and the predominant assessment approaches. These reviews provide the guidance to develop and validate a measure of administrator decision-making.

Assessment Framework for ADMin-SD

Using the conceptual approach recommended by ECD, I define the construct, the behaviors that demonstrate the construct, and the tasks that elicit those behaviors. A chain of logic connects these three conceptual aspects: Decision-making is a cognitive ability that is demonstrated by the behaviors or sub-skills of interpreting problems, setting goals, foreseeing constraints, stating values, and generating solutions. Open-ended, ill-structured problems elicit the use of these skills and can discriminate between experts and novices and are therefore used as the tasks. Responses are scored with six variables, two of which are process-oriented and four of which are product-oriented. As a result, inferences drawn from performances about these sub-skills, and principals' corresponding ability levels, may be articulated with respect to the six variables that are described below.

Researchers from CRESST did not use an exact ECD approach, but their approach shares common features. They provide straightforward information on ten issues they recommend specifying when constructing a measure: (a) identifying the conceptual framework; (b) identifying what to measure; (c) selecting assessment approach (e.g., multiple-choice, performance, etc.); (d) criteria for judging the assessment

(e.g., price, validity, etc.); (e) type of technology; (f) purpose of assessment; (g) participants (e.g., team or individual); (h) level of stakes (high vs. low); (i) context; and (j) recommended testing time. Many of these issues are included in ECD but with different terminology. Following the conceptual explication as recommended by ECD, I detail the remaining practical assessment specifications as recommended by CRESST.

Defining the construct of interest. ADMin-SD defines its construct of interest as decision-making, which involves decision-making skills and domain-specific knowledge. This definition is directly aligned with Zaccaro and colleagues' (2000) view of the construct. As Sugrue (1995) points out, assessing knowledge in ill-structured domains is particularly difficult. Indeed, studies in this literature base have shown the knowledge principals used to solve ill-structured problems is tacit (e.g., St. Germain & Quinn, 2005). To explain, the content domain of *single-digit addition* is quite well-structured; the knowledge and skills needed are clear and articulable. In contrast, the knowledge and skills that principals need is not as easy to demarcate, hence its label as tacit.

Goldring and colleagues (2009) relied on the Interstate School Leaders Licensure Consortium (ISLLC) standards to demarcate what knowledge principals needed to solve ill-structured problems, but they were not able to specify what knowledge to measure further than the text in those standards. Leithwood and Steinbach (1992), when teaching the decision-making skills in professional development, only taught principals how to perform the steps in the decision-making model explicitly and expected the knowledge to pass along tacitly throughout the discussions they had in the class environment. Therefore, ADMin-SD's assessment framework must adopt Sugrue's (1995) view of ill-

structured domains in which it may not be possible to separate knowledge from skills as they grow and change together.

Thus, of the four components that were included in the K-12 frameworks (knowledge, skills, motivation, and meta-cognition), ADMin-SD includes only two (knowledge and skills) for two reasons. First, motivation was included in the K-12 frameworks to account for fourth graders' (lack of) motivation to respond to problems about trains passing in the night. If ADMin-SD is used as intended (e.g., incentivized for research studies, used in principal preparation and at the district level), then it is reasonable to expect principals to be highly motivated to perform their best on the measure, reducing the need to measure respondents' motivation. Second, the meta-cognitive functions referenced in the K-12 frameworks (e.g., reflection and monitoring) are not feasible to measure with this assessment for two reasons. Assessment takes place during a single administration, which does not give respondents adequate time to reflect, and it does not require any solution implementation, thus rendering any monitoring thereof to be meaningless.

Defining the behaviors. To make inferences about this latent cognitive ability, one needs observations of the construct on which to ground the inferences (Bennet, Jenkins, Persky, & Weiss, 2003; Mislevy, Almond, Lucas, 2003). These observations are exhibited as principals demonstrate their skills during their performance on the components of the decision-making process (i.e., interpreting problems, identifying goals, foreseeing constraints, specifying values, and generating solutions). *Interpreting problems* is defined as any statements related to identifying, defining, framing, or prioritizing how they view the problem(s) presented in the vignette. *Identifying goals* is

defined as any statements about goals or what they want to achieve to solve the problem. *Constraints* are defined as any statements that identify, define, or prioritize sub-problems, limitations, obstacles, barriers that may (not) be able to be overcome in the course of achieving the goals that were set. *Stating values* is defined as any statement reflecting a principle, doctrine, or belief. For instance, a principal may say that s/he believes, "All teachers deserve a safe workplace" or "Being inclusive is important." These statements tend to provide a rationale or support for respondents' thinking. *Generating solutions* is defined as any statement that is related to how they will achieve their goal(s), resolve negative circumstances, overcome obstacles, or otherwise solve the problem as they have identified it.

Defining the task. The task that evokes these observations is responding to ill-structured problems because these kinds of problems require principals to structure the problem in such a way that exhibits their individual abilities, knowledge, and values. This line of research shows that ill-structured problems elicit differences between expert and typical principals, while well-structured problems do not. These kinds of tasks are "inherently ambiguous and open-ended" (Wiggins, 1989, p. 85), thus warranting the use of constructed response items. On the one hand, constructed response items are generally viewed with higher face validity and as more authentic because they present respondents with tasks like what is encountered in real life (Braun, Bennett, Frye, & Soloway, 1990; Wiggins, 1989). As well, constructed response items can offer raters a view of how the respondent tried to solve the problem (Birenbaum & Tatsuoka, 1987; Quellmalz, 1989). Constructed response items also eliminate the chance for the respondent to work backward, so to speak, from the list of solutions provided with a multiple-choice or

ranking question until the respondent finds the right solution (Braun, Bennett, Frye, & Soloway, 1990).

On the other hand, constructed response items also present significant challenges in that they are resource-intensive to score and are usually associated with lower reliability in large-scale assessment efforts (Bennett, 1991; Wiggins, 1989; cf. Birenbaum & Tatsuoka, 1987). The low reliability results partially from the items needing to be scored by human raters and partially from the variable nature of constructed responses. Although their reliability is generally lower than that of multiple-choice items, reliability can be increased by using more than one item and/or judge (Bennet, 1993; Bennet, Rock, & Wang, 1991). Typically, constructed response items such as the essay portion of the Scholastic Aptitude Test record student performance on one item, the essay itself. ADMin-SD includes multiple, shorter items to enhance reliability.

The tasks come in two forms: prompted and unprompted vignettes as Zaccaro and colleagues (2000) recommend. Unprompted items are placed before prompted items so the prompts do not influence responses to unprompted items. There are five simple prompts, one for each component of the decision-making model. The five prompts are as follows: How do you define the **problem(s)** presented? What **goal(s)** do you set to solve the problem(s) presented? What **value(s) or principle(s)** guide your thinking? What **constraints** do you foresee? What **solution(s)** do you propose? After field testing, ADMin-SD includes two unprompted and two prompted items. The vignettes depict the student discipline categories that principals must address as part of their job. Definitions of student discipline problems came from the School-wide Information System (SWIS; see www.pbisapps.org) *Referral Form Definitions* document (Todd, Horner, Tobin,

Eliason, & Conley, 2013). SWIS is an online database for school staff to enter student discipline referrals. The system allows for customization but is loaded with pre-defined student discipline categories, such as: defiance, fighting, vandalism, and others. About 8,000 schools use this data system across the country (www.pbisapps.org, n.d.). Although the system can be customized, the definitions of student discipline problems provide the most widely accepted and comprehensive list of student discipline problems available.

Defining the variables and scoring methods. Based on the substantive review and the review of frameworks, ADMin-SD includes six variables: two process and four product. A simple measure of accuracy cannot be used on this assessment because there are no single, guaranteed correct answers to ill-structured problems. The two process variables evaluate performance of the components (or steps) in the decision-making process according to its thoroughness and coherence. The four product variables evaluate the response holistically according to its overall quality, creativity, feasibility, and effectiveness. Thoroughness scores are averaged across components (i.e., the steps of the decision-making process) and across judges for item level scores. Coherence, like the product variable scores, are averaged across judges for an item level score. All item level scores are then averaged for a total score. All variables are scored on a slider scale from 0-100 with four labels that mark zero and the quartiles. The following sections describe and provide a rationale for the proposed variables.

Thoroughness. ADMin-SD assumes that cognitive sub-skills (e.g., identifying problems, determining constraints, etc.) can be observed through one's consideration of the components of the decision-making process. The nature of written, constructed responses requires level of detail to use as a proxy for how much the person thinks about

a topic or idea, thus requiring thoroughness as a measure of the respondent's process. A thorough process would include detailed consideration of each component of the model, where 0 represents *No Discussion* of the component. If respondents do not state any goals, for example, then they are not demonstrating the skill. The next label on the slider, *General Discussion,* indicates the respondent has made a general mention of the component. For example, saying that one would set a goal is more general than actually stating the goal. The next label on the slider scale, *Specific Discussion,* indicates the component has been discussed somewhat specifically. For example, specifying that one will set a goal to train teachers in positive behavior supports is more specific than writing that one will set a goal related to teacher professional development. The final label, *Detailed Discussion,* indicates that component has been discussed specifically and in detail. For example, after setting the specific goal to train teachers in positive behavior supports, the respondent may also discuss their rationale, logistics, or other details.

Coherence. Experts' explanations of their decision-making process tend to be more coherent than novices' explanations. Logically, coherence is needed to evaluate how the components of the decision-making process are performed together, not in isolation from each other. A coherent discussion demonstrates alignment, interrelatedness, and/or consistency within the response (Leithwood & Stager, 1989). A strong response demonstrates coherence among the components and ideas within the components. For instance, if the problem is defined as a lack of staff training in PBIS, then the goals should aim to increase training, and the solutions should involve actually providing the staff with the appropriate PBIS training. These components would then be aligned or consistent with one another.

Judges score coherence based on the ideas put forth. If all the ideas are aligned, consistent, or interrelated, the response should be scored as *Complete Alignment*. If most of the ideas are aligned, it should be scored as *Strong Alignment*. If few of the ideas are aligned, score the response as *Weak Alignment*. If none of the ideas are aligned, score the response as No Alignment. If a respondent only includes one component – solutions for example – judges determine if the ideas within that component are consistent with each other. In the exceptionally rare case that only one basic idea is offered, the response is scored as *No Alignment*.

Quality. This variable simply asks for the judge's overall impression of the quality of the response. Each vignette is constructed so that respondents should address the direct problem(s) presented in the vignette; however, they can also use the problem(s) as an indicator that the issue(s) may need to be addressed at the school-level as well, which judges should consider when evaluating the overall quality of the response. Additionally, if judges thought a response was biased, inequitable, or illegal, they used this variable to reduce its score.

Creativity. Creativity is critical in solving problems (Mayer, 2013; OECD, 2010). Creativity is defined as something original, unique, novel, or a combination of two existing ideas in a new way (Amabile, 1982). Solutions themselves have been defined as a creative idea or a combination of existing ideas (Davis, 1966). The act of solving problems calls for "creative interpretation of situations and production of meanings and possibilities" (Meacham & Emont, 1989, p. 10). In this study, creativity of responses is assessed because novel solutions are often required for ill-structured or novel problems (Leithwood, Cousins, & Smith, 1990; Leithwood & Steinbach, 1991). Despite the

obvious subjectivity involved in rating creativity, 40 years of research has shown that groups of experts can achieve satisfactory reliability when evaluating the creativity of a product or performance (Amabile, 1982; Hennessey, Amabile, & Mueller, 2011). This variable assesses the quality of the ideas put forth and is therefore a product variable. Judges use the labels *Not Creative*, *A Little Creative*, *Somewhat Creative*, and *Extremely Creative* to locate their score on the slider scale.

Feasibility. Feasibility is assessed because solutions must be feasible in order to be effective. "In organizations, it is often far more important to have a workable solution at the right time than one truly best solution" (Mumford et al., 2000, p. 15). As a result, leaders must consider feasibility when evaluating potential solutions (McCall & Kaplan, 1985). Feasibility is defined in four ways: (a) as the possibility of doing something easily, conveniently, or practically; (b) as the ability to accomplish one's goals, usually in light of one's capacity and resources; (c) as attending to constraints and possible negative consequences; (d) as considering whether alternatives conflict with broader organizational efforts, goals, or policies (Mumford et al., 2000). Logically, one could invent a creative solution to a problem, but if it is not feasible to implement, the solution may as well not exist. Ideally, a solution would be high on creativity, feasibility, and likely effectiveness. Judges use the labels *Not Feasible*, *A Little Feasible*, *Somewhat Feasible*, and *Extremely Feasible* to locate their score on the slider scale.

Effectiveness. Likely Effectiveness is defined as the degree to which a solution is likely to be successful in producing a desired result. Indeed, Voss and Post (1988) have defined solutions to ill-structured problems as "good if other solvers find little wrong with it and think it will work" (p. 281). In contrast, they defined a solution as poor if

others can point out why it will not work (Voss & Post, 1988). There is no way to judge likely effectiveness of these kinds of responses other than subjectively, especially because there could be more than one right answer to ill-structured problems. However, if the judges agree with each other, then we can be more confident that the rating is reliable. Judges use the labels *Not Effective*, *A Little Effective*, *Somewhat Effective*, and *Extremely Effectiveness* to locate their score on the slider scale.

Assessment specifications. To supplement the conceptual ECD approach, I follow the recommendations of O'Neill and Schacter (1997) by presenting the remaining six practical and logistical specifications of the assessment that have not yet been discussed.

Purpose of the assessment. The purpose of ADMin-SD is to assess administrator decision-making performance to make inferences about principals' (and assistant principals') ability to make decisions regarding student discipline situations.

Recommended testing time. The assessment should last about 30 minutes.

Type of technology. The assessment is conducted over the internet. Respondents need a computer or laptop with reliable internet access.

Participants. There are two types of participants: the respondents and the judges. Principals, assistant principals, and aspiring principals are the intended respondents. Local qualified judges need to be recruited to score responses; this point is crucial to the validity of the instrument. According to contingency theory, the best answers to ill-structured problems are the solutions that best fit the context; therefore, judges who know the local context must be used. A local qualified judge includes authority figures and experts who routinely evaluate principals' decision-making as well as principals

themselves who have previously been judged as particularly effective decision-makers (D'Zurilla & Maydieu-Olivares, 1995). For example, a superintendent (or another central office district administrator), who supervises principals, would be a qualified judge. Professors and program director(s) would be qualified judges if ADMin-SD is used in principal preparation or licensure programs. When ADMin-SD is used for research purposes, these qualified raters should recruited from the locale in which the study is taking place. For example, if the study was taking place in Portland, Oregon, qualified judges should be recruited from the city itself. Judges from rural eastern Oregon may not be familiar enough with the local customs, rules, and procedures particular to Portland, which makes it harder for judges from outside the locale to judge whether an idea is feasible, creative, or will likely be effective. In other words, aligned with Contingency Theory, the best solution in an urban setting may not be the best solution in a rural or suburban setting. This notion suggests that a judge from one setting – who holds norms and values particular to that setting – should not impose and apply their standards and judgment when evaluating a solution intended for another setting. Therefore, local judges need to be recruited to maintain the ecological validity of the instrument.

Context. The assessment can be administered in any office or location in which the respondent is comfortable and can maintain a reliable internet connection.

Level of stakes. Until empirical evidence demonstrates that ADMin-SD can generate valid and reliable inferences about principals' decision-making ability, it should be used in low-stakes situations, including but not limited to: principal preparation programs, initial and continuing licensure programs, professional development programs.

Criteria for judging assessment. The assessment should be judged according to three basic criteria. First, the measure must demonstrate adequate reliability. Second, it must demonstrate technical adequacy in discriminating between high and low performances, without demonstrating bias based on gender and/or race/ethnicity. Third, after those two conditions have been satisfied, the assessment's practical utility and efficiency must be the ultimate criteria for judging the assessment's value. ADMin-SD has relatively high practical value in that it is be the only measure of its kind for such an important area of principals' practice. Its efficiency is a balancing test among the gains in practical value versus the resources required to administer and score the assessment. It takes about three to four minutes to score each vignette for each judge with a total of four vignettes, totaling about 12-15 minutes per respondent per judge.

The preceding chapter has reviewed the literature on principal decision-making, the assessment literature related to measuring that construct, and presented an assessment framework designed to generate inferences about principals' decision-making skills in addressing student discipline situations.

CHAPTER II

METHODS

The purpose of the current study is to develop and validate an instrument to assess administrator decision-making in student discipline situations. To that end, I pose three research questions that reflect the instrument's iterative development, regarding its content validity, its pilot test, and its field test. Then, I describe the methods used to answer those questions.

Research Questions

RQ1: What is the content validity of the instrument?

RQ2: What are the initial psychometric characteristics of the instrument after pilot-testing?

RQ3: What are the psychometric characteristics of the instrument after field-testing?

Methods

I used a modified form of the Behavior Analytic Model (BAM; Goldried & D'Zurilla, 1969) to develop and validate ADMin-SD. BAM has been used to construct and assess individuals' competence in social problem-solving skills across various kinds of situations (e.g., Bullis, Bull, Johnson, & Johnson, 1994). The model was developed in the 1960's by social learning theorists who believed that people's behavior should be interpreted as an interaction with their situation and their environment. The model is used to identify, train, and evaluate performances' level of competence. BAM views decision-making skills molecularly in that the overall ability can be broken into specific behaviors; it does not view it as a static trait on which the individual is high or low (McFall, 1982).

The BAM requires five phases for development and validation of the measure: situational analysis, response enumeration, response evaluation, development of the measure format, and evaluation of the measure. Situational analysis involves analyzing the domain to cover the depth and breadth of content. Response enumeration and evaluation involve identifying the possible responses and their relative strength. The fourth stage, development of the measure format, must be performed to compile the items and scoring system into their final form. The last step involves evaluating the measure's reliability and construct validity. To adapt the BAM to fit the construction of a measure of administrator decision-making with respect to ill-structured problems, I modified the response enumeration and evaluation stages because ADMin-SD uses a constructed response format. In short, I developed and validated the items and solicited content validity information from judges. I piloted the vignettes and scoring method, refining both based on their performance statistically and on qualitative feedback from respondents and judges. Lastly, I conducted a field test to make final revisions and appraise its reliability and discriminant validity.

Initial instrument development. When evaluating an instrument's construct validity, the content included in the measure must be considered (Goldfried & D'Zurilla, 1969; Messick, 1995). According to Messick (1995), three criteria should be used to evaluate whether the content included is valid: content relevance, representativeness, and technical quality. Content relevance refers to the idea that the content included in each item reflects at least a portion of the content domain being assessed. Representativeness refers to the idea that content is sampled proportionally to its occurrence in the domain. Technical quality refers to the technical construction of the items, which in this case

means that the vignettes are sufficiently ill-structured and written clearly by removing ambiguities. Traditionally, content relevance and representativeness are assessed by expert professional judgment, and the documentation generated by that assessment "serves to address the content aspect of construct validity" (Messick, 1995, p. 6). In this study, the content domain is reflected in the various types of problems administrators face regarding student discipline and is derived from the School-wide Information System (SWIS) database, which is an online database for school staff to enter student discipline referrals. The definitions of student discipline problems from the SWIS system provide the most widely accepted, comprehensive list of student discipline problems available.

I wrote vignettes for selected categories of discipline problem in the SWIS database. I removed categories that elicited pre-determined, legal responses such as a bomb threat. In these instances, principals are taught to respond with a pre-determined procedure such as calling the police and cooperating with law enforcement. By using a pre-determined procedure that affords them no discretion, principals are not able to demonstrate their decision-making ability, thus rendering these categories useless in estimating their skill. After eliminating these categories (n = 2) and the category of *other*, 23 categories remained. It would not be feasible to represent all 23 categories in a constructed response measure, so the categories had to be reduced further. Thus, I selected the eight categories that are cited as problematic and the most disproportionate categories according to Skiba and colleagues (2002, 2011). These categories reflect moderate infractions that are subjective in nature to some degree and include: abusive language, defiance, disrespect, disruption, fighting, physical aggression, harassment, threat. The eight categories were further collapsed into six by combining defiance,

disrespect, disruption into one category, which is a common practice in quantitative studies in this literature. I wrote 13 vignettes to reflect the final six categories.

Based on real events in every case, I wrote vignettes to be ill-structured using a few strategies. For example, for some vignettes, I wrote the vignette so that both parties involved were going to be seen as somewhat right and somewhat wrong in their actions. Further, the situation had to be muddied somehow, usually by withholding information that would help resolve it. For other vignettes, I included some issues that relate to family values, which can bring up a conflict for educators who may have opinions that they cannot impose on a student or family in certain instances, such as views of different cultures. Initially, I referred to characters as the first student or the second student; I tried to avoid using pronouns to avoid gender bias and to seek a neutral condition. However, this practice confused some respondents and judges, so names were later added to the characters. See Appendix C for the vignettes sent out for content validity, Appendix D for the vignettes used in the pilot test, and Appendix E for the vignettes used in the field test. Beyond the content of the vignettes, these appendices show the revisions I made and the pictures I used in the different forms.

The pictures. After the content validity study, pictures of the characters in the vignettes were added to concretize and standardize the vignettes to a degree. Without a picture, each respondent would conjure his or her own image of the characters in the vignettes; thus, the pictures help standardize how each respondent interprets the vignettes. To secure the pictures, I searched stock photo websites for pictures of adults (teachers/paraprofessionals) and middle school male and female students without a smile or a frown who were White, Hispanic and/or Latinx, Black, and Asian. These criteria

were not easy to satisfy, particularly for pictures of Black, Hispanic, and Asian people. Three to five pictures were selected for each character where possible; I selected and later used the only suitable photo I could find for the shooter vignette.

To estimate the pictures' reliability, I assembled those pictures into a survey and sent it to the judges for rating. If a picture is intended to depict a middle school, Hispanic or Latina young woman, then respondents should reliably interpret the picture in that way. Judges were asked to rate the age, gender, race/ethnicity, and attractiveness of the people in the pictures. Typically, attractiveness has been defined with six dimensions: attractive, classy, handsome/beautiful, elegant, sexy, and likable (Ohanian, 1990). As it is inappropriate to ask administrators to rate the beauty or sexiness of students, I queried the judges on attractiveness and likability. I did not query on elegance and class as they are difficult to assess from a two-dimensional portrait of a middle school student. In fact, one judge was uncomfortable with rating the attractiveness and likability of the students in the pictures and rated all of them 100, which I believe demonstrated a discomfort with evaluating and/or differentiating between students on those characteristics. In retrospect, I should have only included likability. See Appendix F for a summary table of the results of the survey.

I aimed to give pictures to about a third of the pilot respondents across the three groups to see if the pictures had an influence on responses. Employing this methodological choice made the pilot test an underpowered two-way ANOVA design (discussed in more detail below). For the pilot test, I used pictures of white males to the extent possible to keep that constant to control for any bias that may be introduced by using pictures of students of different genders and race-ethnicities. Including only white

males, however, does not reflect reality; principals address student discipline situations with students of all backgrounds. Additionally, one pilot respondent noted all the vignettes involved White males and commented that there should be more representation. If one respondent felt strongly enough to comment, more were thinking and feeling it. From a measurement standpoint, I believe the comment speaks more generally to respondents' motivation to complete the measure and perform their best, so I took heed to change the pictures for the field test.

Finding pictures of students who could be reliably rated as middle school age was more challenging than I expected. I think that is partially because middle school students can look awkward, and stock photographers have a harder time selling pictures of awkward looking people. In context, however, the borderline students should be able to pass as middle schoolers. After pooling the reliably rated pictures where agreement > .50 on gender, race-ethnicity, and age, I randomly selected the pictures to fill the roles in the vignettes for the field test. For example, three pictures were rated reliably that were suitable for the cyber-bullying vignette: a White young man, a White young woman, and a Black young woman, all with computers. From those three pictures, I randomly selected the picture that would be used in the field test. I used this procedure to fill all roles. See Appendix G for the pool of reliably rated pictures.

Development of the scoring method. As detailed in the Assessment Framework, ADMin-SD embodies six variables: quality, creativity, feasibility, effectiveness, coherence, and thoroughness. Judges are shown the vignette, the response, and then asked for their ratings. For the pilot test, I asked judges to score coherence two ways: overall and by decision-making component. I had them rate, in order, the response's: quality,

creativity, feasibility, effectiveness, coherence both ways, then thoroughness for each component. Judges made 15 ratings for the pilot test and 10 for the field test because coherence by component was removed. Judges used sliders scales for the ratings, placing the slider where they thought it belonged on the scale's continuum. See Appendix H for screenshots of the slider scales.

Content validation – answering RQ #1. Following the development of the vignettes, their content must be validated as one aspect of the measure's construct validity (Messick, 1995). The judges recruited for the study were surveyed as to each vignette's frequency of occurrence, importance, and realism. Vignettes were ranked according to their mean scores across those characteristics, and the ranks were summed. The survey also included an open-ended field after each vignette to prompt respondents for qualitative feedback about the technical quality of each item (e.g., grammar, syntax, clarity of writing) as well as suggestions to improve the vignettes. Qualitative suggestions were logged and either incorporated or rejected. The data generated by this survey serves as the evidence for the measure's content validity (Messick, 1995).

Recruiting judges. As discussed in the Assessment Framework, qualified judges include people who supervise, mentor, train, and/or evaluate principals and assistant principals. Judges were recruited through a nomination process in which a state of Oregon board of education member, a former state of Oregon chief education officer, a director of a research unit focused on positive behavior supports, and a director of an administrator licensure program all nominated judges who met three criteria. The nominees had to be either a current or former principal, a current or former central office administrator, and an expert in handling student discipline according to the nominator's

opinion. Because no absolute measure exists to identify who is an expert judge, I had to use this relative method for identifying them (Chi, 2006).

The nominators produced a list of 40 candidates, with one judge producing about half of all those nominated. After checking they met the three criteria, I had to remove three nominees. I did not want 37 judges, nor could I compensate that many. I wanted the fewest number of judges possible that would make the scoring workload feasible. I was aiming for about 15 judges, so I had to determine the order in which I would contact them because I wanted to allow for equal representation across the nominators. Using the following procedures, I emailed judges, asking them to participate in rounds. Three nominees were nominated by two separate judges; I solicited their participation first with an email that outlined the expected workload and an offer to meet or talk on the phone to answer any questions they might have; one agreed to participate. Then, I solicited the first two to four nominees across all four nominators to participate. I followed this process until I had contacted 24 nominees in total, 13 of whom agreed to participate. I kept the remaining 13 available to contact later in case they were needed.

The group of 13 judges who agreed to participate was reduced to eight over the course of the study due to attrition. Thirteen judges finished the content validity survey; 10 finished scoring the pilot responses and responding to the post-pilot feedback; eight judges finished scoring the field test responses and responding to the final post-field test feedback survey. I collected data on their demographic and professional background, including their years in schools, previous positions held, and self-rated expertise in addressing student discipline.

First revision of the instrument. Following the content validity survey, I dropped or retained and edited vignettes based on their importance, difficulty, and realism. Clarity was used primarily as an indicator of which vignettes needing revising. I had the ability to pilot test eight vignettes due to budgetary considerations, as more vignettes to score costs more in payments to judges. I used the content validity rankings along with theory to select which eight. Beginning with the eight best ranked, I removed one for defiance because I did not want to overrepresent defiance at the expense of another category. I also wanted to include the vignette about a teacher feeling threatened, ranked 9th, because it is an important issue in theory. Of the eight vignettes, odd numbered ranks became unprompted vignettes and the evens became the prompted vignettes. They were then revised for clarity based on their rankings and qualitative feedback from the judges. For example, the vignette about homophobic harassment and fighting needed revising as respondents mixed up who was harassing and being harassed in their responses, and even a judge emailed me to clarify who was doing what to whom. As a result, I considered adding names to that vignette to increase clarity, but I decided to add names to the characters in all vignettes so that feature would remain constant.

Pilot test – answering RQ #2. After editing and compiling the items that scored best on the content validity survey into a pilot form, I tested the instrument with a small group of master's students, aspiring principals, and established administrators to answer the second research question. The vignettes should be understood by respondents and should generate reasonable constructed responses. Judges should demonstrate adequate reliability and find the scoring method usable.

Recruiting respondents. Based on the literature reviewed, I recruited three groups of respondents: graduate students earning their Master's in Teaching, aspiring administrators, and established administrators. First, I recruited established administrators earning their Professional Administrative Licensure (ProAL) at the University of Oregon to compose my group of *experts* (14.37% response rate). Students enrolled in this program are typically, but not always, current administrators who are maintaining their licensure. Second, following the example of Brenninkmeyer and Spillane (2008) and Spillane, Stephan, and White (2009), I recruited aspiring principals from the Preliminary Administrative Licensure (PreAL) program in EMPL at the University of Oregon to compose my group of *novices* (19.70% response rate). Students enrolled in this program are usually teachers who want to become administrators. Third, following the example of Allison and Allison (1993), I recruited a sample of graduate students earning their Master's in Teaching from the University of Oregon to compose a group of *true novices*. Students were recruited from both UO Teach (6.06% response rate) and the K-12 Special Education program (22.22% response rate). Students enrolled in these programs are typically pre-service teachers. Occasionally, for example, someone who taught in another state that did not require a master's degree to teach has now moved to Oregon where it is required. To recruit these respondents, secretaries from the respective departments sent an email from me that solicited their consent, provided directions for their participation, and informed them they would be provided a research incentive of $30 for their time; the email linked them to the Qualtrics survey. I collected data on respondents' demographic and professional background. The pilot form of ADMin-SD was administered to 69

respondents, 49 of whom finished. Data from respondents who did not finish were used to train the judges.

Training the judges. I trained the judges via PowerPoint Slides that included voice over explanations of the material. Online training for educators can be as effective as face-to-face training (e.g., Fishman et al., 2013), yet more efficient (Dede, Ketelhut, Whitehouse, Breit, & McCloskey, 2009). A copy of the training can be downloaded from Google Drive at this link; it must be opened in PowerPoint for the audio to work. Taking about an hour to complete, the PowerPoint slides included definitions and examples of the types of vignettes, the decision-making components, and how to score the variables.

I conducted the training this way for three reasons. First, I want the training to be as automated as possible so districts, schools, and licensure programs can use it in the future, without me present. Second, from a practical standpoint, I did not think I could schedule all my judges together at the same time because they were mostly full-time administrators, working 40-60 hours per week. Third, I wanted (and budgeted for) the training to take an hour. If we trained in person, it would have taken longer with just including driving, parking, etc..

At the end of the training, judges were directed to click a link that brought them to four practice exercises in Qualtrics, two unprompted and two prompted. These practice exercises, like the examples in the training, were taken from respondents who did not complete all the vignettes. To check the effectiveness of the training, I calculated an intra-class correlation (ICC) using procedures recommended by Hallgren (2012). The training prior to the pilot test generated an ICC(c,1) of .66 and an ICC(c,k) of .94 including all 60 ratings and 11 raters, and the training prior to the field test generated an

ICC(c,1) of .59 and an ICC(c,k) of .94 with all 40 ratings and 8 raters. The ICC(c,1) coefficients reflect the agreement of the individual ratings, indicating moderate reliability. The ICC(c,k) coefficients suggest the average of the judges' ratings, which is reported by ADMin-SD and used in hypothesis testing, achieved excellent reliability.

Second revision of the instrument. Following the pilot test, vignettes were dropped or retained and edited based primarily on their mean ICC, then their ability to discriminate, and then their content validity ranking. I wanted to select the best of the unprompted and prompted vignettes to ensure both were on the field test. The specific pilot test results are presented in Chapter III; however, regarding my methods, "Threat" was dropped because of its low ICC and its inability to discriminate between novices and experts. "Defiance" was dropped also because of its inability to discriminate and its lower content validity ranking (5^{th}). The "Shooter" vignette was retained because of its mean ICC, ability to discriminate, and its high content validity ranking (1^{st}). The "Cyber-bullying" vignette was retained because it had a similar ICC, and showed some ability to discriminate with larger effect sizes than "Defiance" and "Threat," but not large enough to achieve statistical significance.

Of the prompted vignettes, "Elbow" had an acceptable ICC, but it showed no ability to discriminate while the other three did; it was also ranked low in content validity (8^{th}), so I dropped it. The remaining three were all strong in different ways, so I retained them all, hoping evidence from the field test would shed light on which were the strongest. Although "Homophobia" had the lowest ICC, that was largely due to one pair of judges disagreeing with each other. Nonetheless, the vignette still demonstrated some ability to discriminate and a high content validity ranking (2^{nd}), so I kept it. The "Para"

vignette demonstrated the best reliability and an ability to discriminate, so I kept it. “Go Back” demonstrated the same reliability, potential to discriminate, and was ranked highly in content validity (3rd), so I wanted to see how it would perform in the field test. The content of the three prompted vignettes relates to homophobic harassment and fighting, harassment of a Hispanic student and near fighting, and a negative physical interaction between a student and a paraprofessional. The “Homophobia” and “Go Back” vignettes are similar in nature, so I would have to choose only one for the final form. They are both about students being harassed and then fighting back or making a threat to fight back.

Refining the scoring method. Judges also responded to a separate post-pilot feedback survey, which I administered through Qualtrics. I queried them about the directions and operational definitions in the judges’ training, about the clarity of the slides and voiceover, about their experience scoring responses with the selected variables, and about the graphical representation of what they saw on the computer screen in Qualtrics. I used feedback from this survey to refine the variables, the training, and what the judges saw in Qualtrics. For example, I asked judges if they developed a preference for scoring coherence, either by component, overall, or no preference. About half responded no preference, while the other half said they preferred scoring it overall. I used this feedback, along with statistical analysis, to remove coherence by component. Definitions for the other variables were specified further by providing more precise examples. The log of revisions is available upon request.

Field test – answering RQ #3. After revising the instrument, I conducted a field test to estimate its technical adequacy, answering the final research question. The design of the field test is identical to the pilot test.

Recruiting respondents. Although the design of the field test is identical, I had to widen the sample to licensure programs in the state of Oregon. I recruited respondents the following school year from the same programs at the University of Oregon. Response rates were: 23.08% for ProALs, 19.40 for PreALs, 8.91% for the UOTeach program, and 5.00% for the K-12 Special Education program. Then, I recruited respondents from the comparable programs at Portland State University (response rate of 10% for ProALs and 10% for PreALs), George Fox University (6.45% ProALs, 0% PreALs), Concordia University (4% ProALs, 12.94% PreALs, 19.03% MATs). Dr. Keith Hollenbeck connected me with educational administration professors at these schools by sending the same recruitment email for me them. A professor from Portland St. posted my solicitation to their "Desire to Learn" management system for 30 PreAL students (10.00%), while another professor forwarded it to her 10 ProAL students (10.00%). A professor at George Fox University had my recruitment email forwarded on to 42 PreAL students (0.00%) and 31 ProAL students (6.45%). Concordia required me to go through their IRB and then forwarded my email to 247 Master's students (19.03%), PreAL students (12.94%), and ProAL students (4.00%). The overall response rate was 18.15%, 143/788, of which 118 finished. Data from respondents who did not finish were again used to train the judges. See Appendix I for a graphical representation of the lack of differences between the institutions on their total scores.

Training the judges. The judges were trained in the same fashion. The training was edited to reflect the changes to the instrument as well as based on feedback from the post-pilot feedback survey. More and better examples were added to help define the components of the decision-making model and the variables.

Analysis plan. I conducted several analyses to answer the research questions and evaluate ADMin-SD's technical adequacy. Scores were analyzed such that they demonstrated use of the full range of scales, that judges could produce adequate reliability for each vignette, variable and overall, and that the instrument could discriminate according to respondents' program enrollment, their current role in schools, their self-rated level of expertise in addressing student discipline, and their years spent professionally in schools. Ideally, all variables and vignettes should demonstrate differences in scores according to an analysis of variance on all four variables that I refer to as the *proxy variables* because they are acting as a proxy for the continuum between experts and novices. I presumed that more experienced educators will perform better on all variables, except less experienced respondents may be more creative, offering fresh ideas.

Content Validity. To answer RQ #1, judges were asked to rate each vignette's difficulty, importance of occurrence, and the realism depicted in the vignette. I conducted descriptive analyses, including examination of means and standard deviations for importance, difficulty, and realism. These means were ranked and aggregated to decide which vignettes would be dropped, revised, or retained. The content validity survey also included an opportunity for judges to offer open-ended feedback on how to improve the vignettes. This feedback was logged and either used or rejected; the log (excel spreadsheet) is available upon request.

Pilot test. To answer the second research question, I conducted several analyses to appraise the technical adequacy of the instrument after pilot testing. After cleaning the data, I ran descriptives (i.e., mean, SD) for the variables and vignettes. I ran checks for

normality and outliers. Then, I estimated the reliability of the variables and vignettes to revise the text of the vignettes and the scoring method, and I estimated the vignettes' and variables' ability to discriminate between novices and experts based on the study's four proxy variables.

Checks for normality, outliers, and local independence. I relaxed my assumptions of normality because of the small sample size. The univariate distributions of each variable approached uni-modal normal distributions. Additionally, I did not remove the rare outliers that I identified with box plots because I knew all values were reasonable; none were typographical errors. And from a measurement perspective, I want to see the full range of the scales are used. The assumption of independence was, however, violated as responses are clustered in vignettes and respondents and nested in judges. Because of the small sample size, mixed models were inappropriate, so I used a robust estimator with a simple linear regression to account for the local independence and outliers (Yaffee, 2002) when assessing the relationship between current role, years in schools, and self-rated expertise.

Reliability. To estimate the reliability of the instrument, I conducted several complementary analyses including using Krippendorf's Alpha (KAlpha), ICCs, and the inter-item correlation matrix. I estimated KAlpha with the kripp.boot package in R (Proutskova & Gruszczynski, 2017). KAlpha is becoming more widespread in its use because of its flexibility; it works with two or more judges using nominal, ordinal, interval, and scale variables and can handle missing data (Hayes & Krippendorf, 2007). Krippendorf (2011) suggests variables above .80 demonstrate adequate reliability, while variables above .67 demonstrate some reliability, but he notes that these are just his

general impressions and lack empirical support. Like Cohen's Kappa, Krippendorf's Alpha is a chance-corrected reliability coefficient. Therefore, it may make sense to use similar cut points to those recommended for Kappa by Landis and Koch (1977). They view the agreement of Kappa coefficients < 0 as *poor*, 0 to .20 as *slight*, .21 to .40 as *fair*, .41 to .60 as *moderate*, .61 - .80 as *substantial*, and .81 to 1 as *near perfect*.

There are multiple versions of the ICC, and three determinations must be made to select the appropriate form to use (Shrout & Fleiss, 1973). First, a two-way model is appropriate because I sampled my judges from a larger population by randomly assigning judges to respondents. Second, a measure of consistency, *c*, rather than absolute agreement is appropriate because of the scale being used and because it is more important that raters agree in rank than in their absolute number (Hallgren, 2012). Third, the average unit, *k*, is appropriate because this choice relies on how the measurement protocol is implemented in the field (Koo & Li, 2016) and on the unit of analysis used to make inferences (Shrout & Fleiss, 1973). Thus, the average unit ICC is appropriate for two reasons: a) the averages of the judges' ratings would be reported to actual respondents, and b) the averages are used in this study's hypothesis testing. However, I have also included the ICC for single units as Appendix J for the pilot test and Appendix K for the field test. ICCs should be interpreted on a scale where 1.0 is perfect agreement and 0.0 is random agreement. ICCs can be negative and suggest systematic disagreement. There are two similar sets of standards by which to interpret the ICC. Commonly cited, Cichetti (1994) recommends cut points where reliability reflected by ICC values less than .40 is poor, by values between .40 and .59 is fair, by values between .60 and .74 is good, and by values between .75 and 1.0 is excellent. More recently, Koo and Li (2016)

recommend slightly more stringent cut points, where the reliability indicated by values less than .50 is poor, .50-.74 is moderate, .75-.90 is good, and above .90 is excellent.

With 12 judges beginning the pilot test, I divided the 12 into four distinct, randomly assigned trios for providing ratings on the responses. Two did not finish, but they occupied separate trios, turning those trios into duos. I conducted reliability analyses for each vignette and variable, as well as the total (mean) scores across vignettes. Because two or three judges scored each response with interval scales, I report the mean two-way ICC(c, k) across groups of judges (Shrout & Fleiss, 1973); I include the mean's range. To calculate the mean ICC, I subset the data into the distinct groups of judges, calculated their ICC with the irr package in R (Gamer, Lemon, & Singh, 2012), then averaged the ICC coefficients across groups of judges. Following an initial estimation of the ICCs, I re-ran the analyses by systematically removing variables and vignettes. For example, I had to choose which method of coherence would remain, so I used these comparative analyses to make that decision.

Beyond the ICC, I also constructed an inter-item correlation matrix. I examined the matrix looking for correlations that were either very high, indicating the variables may be redundant, or very low, indicating they are unrelated to the construct. These correlations then help suggest whether variables should be dropped or refined.

Discriminant validity. Ideally, each variable, vignette total, and the measure overall should be able to discriminate between novices and experts, specifically the four proxy variables in this study. Thus, analyses of variance are appropriate to determine if there are statistically significant relationships between the proxy variables and ADMin-SD's variables, vignette totals, and overall. If the relationship is significant, then it could

be said that the variable, vignette, or total overall discriminates according to the proxy variable. I calculated and report the achieved power for each analysis.

To assess program enrollment's relationship with the outcomes of interest, because of the small sample sizes, I ran separate ANOVAs for each variable and vignette total with two independent factors: the pictures and one of the proxy variables. Ratings were averaged across judges for use in these analyses. I tested the null hypothesis that there were no differences between the groups for the categorical proxy variables and no significant relationship between the proxy variables and ADMin-SD's outcomes of interest, using an alpha of .05 and noting if alpha was < .10 as well. Alphas were adjusted post-hoc when conducting Tukey tests. I report the unstandardized mean differences between the groups as well as eta^2 and Cohen's f^2 as effect sizes for the overall effect of program enrollment. I used these effect sizes more than the alpha to estimate the vignette's or variable's discrimination.

To assess whether one's current role is a predictor of respondents' performance, I coded their current role as a dummy variable where school-based administrators (i.e., current principals and asst. principals) were coded as "1" and everyone else was coded as "0." I entered this variable, along with if they received pictures in their form, into a linear regression with a robust estimator, regressing the outcomes of interest onto the proxy variables. I followed the same procedures with years in schools and self-rated expertise, which are continuous variables.

Post-pilot feedback survey. judges were surveyed about the training and scoring method after they scored the pilot test responses. Questions concerned how clear the slides were, how many questions they had during the training, and how appropriate the

variables were for appraising the responses, etc.. I calculated frequencies on the responses (e.g., judges were asked if they preferred scoring coherence overall or by component). I logged qualitative feedback and then either incorporated or rejected it. The log of revisions is available upon request. Based on the content validity, reliability, discriminant validity, and the post-pilot feedback survey, I made several revisions to the instrument, including: deciding to assess coherence overall, changing how it was operationalized, and reducing the vignettes to the five that demonstrated the best reliability and validity properties.

Field test. To answer the third research question, I followed the same general strategy. There is only one main difference. With respect to the reliability analysis, and because 10 judges began the field test (two more did not finish), I used a crossed design to randomly assign judges to respondents. The two who did not finish occupied two of the same trios, so those trios were not included in the reliability analyses. I estimated ICC(*c*,*k*) with confidence intervals for the remaining trios and calculated their arithmetic mean, as in the pilot test. Krippendorf's alpha (KAlpha) is appropriate for crossed designs as well, so I report it with confidence intervals for the variables only. In fact, because KAlpha can handle missing data, it was much easier to run that analysis than the ICCs for various groups.

With a sample size that is small for mixed models, leading to worse power, I again ran the linear regressions with a robust estimator. For the variables' total (mean) scores and the grand total (mean) score, there was no nesting, so simple ANOVAs were run for program enrollment, and the linear regressions were run with the robust estimator again to maintain consistency across the analyses, though the standard OLS estimator

may have been equally appropriate. Alphas were adjusted post-hoc when conducting Tukey tests. Each outcome of interest was regressed onto each proxy variable. In each case, I tested the null hypothesis that there were no differences between the groups and no significant relationship between the proxy variables and ADMin-SD's outcomes of interest, using an alpha of .05. I report the power of each analysis, unstandardized mean differences for program enrollment, unstandardized betas and standard errors for current role, standardized and unstandardized betas and standard errors for years in schools and self-rated expertise.

CHAPTER III

RESULTS

This chapter presents the results of the various analyses conducted to develop and validate ADMin-SD. Descriptive statistics are provided for the participants (including judges and respondents), the vignettes, and the variables. The research questions are answered in order by presenting the results related to: the instrument's content validity, technical adequacy and revisions after the pilot test, and the instrument's technical adequacy after the field test.

Participants

The following sections present the participants' demographics and background, which includes the judges, pilot test respondents, and field test respondents.

Judges demographics and background. Of 13 total judges, five identified as female, seven identified as male, and one preferred not to answer. One judge identified as Hispanic, while 12 did not. All 13 identified as White. One of the judges was a current superintendent; three were asst. superintendents; one was an other central office administrator, two were current principals; three were retired, two current educational consultants but former administrators, and a current principal supervisor and instructor who was a former principal. Their mean age was 54.85 (*SD* 10.35), and they have an average of 29.38 years in schools (*SD* 9.86), with a range of 17 to 43. Last, they rated themselves on their expertise in handling student discipline (*M* 83.23, *SD* 7.62, mdn 85).

Pilot respondent demographics and background. Please see table 8 for a detailed breakdown of the demographics and background of the pilot test respondents.

With a sample of 49 respondents, the sample was 61.2% female, 87.8% White, with an average age of 38.9 and 12.39 years in schools.

Table 8. Pilot test respondent demographics

	Admins (n = 24)	Aspiring (n = 13)	Grad Students (n = 12)	Overall (n = 49)
Pictures	18	2	9	29
No Pictures	6	11	3	20
Gender				
Male	9	6	4	19
Female	15	7	8	30
Hispanic or Latinx				
Yes	-	-	1	1
No	23	13	11	47
Prefer not to answer	1	-	-	1
Race-Ethnicity				
White	23	12	8	43
Hispanic or Latinx	-	-	1	1
Black or Afr. Am.	-	-	-	-
Asian	-	-	2	2
Am. Ind. or Al. Nat.	-	-	1	1
Prefer not to answer	1	1	-	2
Current Role				
Gen Ed Tch Candidate	-	-	5	5
SpEd Tch Candidate	-	-	3	3
Gen Ed Tch	4	7	-	11
SpEd Tch	-	-	3	3
Asst. Principal	6	1	-	7
Principal	10	-	-	10
Other Cent Off	3	-	-	3
Admin	1	5	1	7
Other				
Age Mean (SD)	45.33 (8.52)	38.00 (9.50)	31.08 (6.39)	38.90 (10.04)
Years in Schools Mean (SD)	16.62 (5.78)	12.38 (7.17)	3.92 (4.68)	12.39 (7.80)
Years Experience w/Discipline Mean (SD)	10.45 (7.40)	10.46 (6.77)	2.08 (1.62)	8.41 (7.18)
Self-Rated Expertise Mean (SD)	69.83 (16.55)	74.92 (12.84)	34.33 (24.60)	62.49 (24.04)

Field respondent demographics and background. Please see table 9 for a detailed breakdown of the demographics and background of the field test respondents.

Table 9. Field test respondent demographics

	Admins (n = 33)	Aspiring (n = 28)	Grad Students (n = 57)	Overall (n = 118)
Gender				
Male	13	7	13	33
Female	20	20	44	84
Prefer not to answer	-	1	-	1
Hispanic or Latinx				
Yes	2	4	4	10
No	24	22	52	98
Prefer not to answer	7	2	1	10
Race-Ethnicity				
White	30	22	43	95
Hispanic or Latinx	1	1	4	6
Black or Afr. Am.	-	-	4	4
Asian	-	1	4	5
Am. Ind. or Al. Nat.	-	2	1	3
Prefer not to answer	2	2	1	5
Current Role				
Gen Ed Tch Candidate	-	-	29	29
SpEd Tch Candidate	-	-	1	1
Gen Ed Tch	4	13	7	24
SpEd Tch	-	2	-	2
Asst. Principal	8	2	-	10
Principal	15	2	-	17
Other Cent Off	3	-	1	4
Admin	2	9	11	22
Other	-	-	8	8
Educ. Assistant	1	-	-	1
Retired				
Age				
Mean (SD)	43.30 (10.25)	37.42 (6.91)	31.32 (7.95)	36.12 (9.81)
Years in Schools				
Mean (SD)	17.45 (7.68)	12.43 (5.51)	3.56 (3.89)	9.55 (8.21)
Years experience w/Discipline				
Mean (SD)	12.30 (8.27)	9.46 (5.88)	2.65 (2.86)	6.97 (7.02)
Self-Rated Expertise				
Mean (SD)	74.61 (13.39)	66.35 (19.26)	35.28 (23.89)	53.65 (27.10)

With a sample of 118 respondents, the sample was 71.2% female, 80.5% White, with an average age of 36.12 and 9.55 years in schools.

Content Validity – Answering RQ #1

Thirteen judges rated 13 vignettes on their importance, difficulty, realism, frequency, and clarity. Table 10 presents the results of the content validity study.

Table 10. Content validity results

	Category	Clarity	Freq.	Importance		Difficulty		Realism		Total	Final Rank
		Mean	Mean	Mean	Rank	Mean	Rank	Mean	Rank		
Shooter	Threat	75.54	15.85	97.00	1	84.62	2	72.62	3	6	1
Homophobia	Harass	55.54	43.92	93.92	2	86.38	1	68.69	6	9	2
Cyber	Harass	76.69	65.31	76.92	8	77.31	4	85.38	1	13	3
GoBack	Harass	47.92	48.54	79.23	5	58.54	6	73.00	2	13	3
NoHist	Def	81.23	24.77	82.00	3	53.23	8	69.62	5	16	5
OneTeacher	Def	71.00	68.15	81.85	4	52.54	9	72.23	4	17	6
Para	PhysAgg	77.08	16.69	78.85	6	69.92	5	53.85	9	20	7
Elbow	Fight	58.77	18.46	78.46	7	83.31	3	36.15	12	22	8
Teacher	Threat	86.85	22.23	66.54	9	52.31	10	44.77	11	30	9
TeachNoAct	Def	73.00	39.31	61.15	13	54.46	7	45.54	10	30	9
NoRead	Def	20.15	52.08	63.62	11	14.23	13	62.54	7	31	11
TeacherPara	AbLang	64.46	73.85	63.46	12	30.92	12	60.92	8	32	12
Student	AbLang	66.15	28.15	66.46	10	39.54	11	27.38	13	34	13

Note. AbLang = Abusive language; Def = Defiance; PhysAgg = Physical aggression; Harass = Harassment.

In short, I ranked and summed judges' mean scores on importance, difficulty, and realism and then selected the top eight for budgetary reasons. Despite querying the judges on frequency and clarity, I did not use those metrics to select the pilot vignettes because some vignettes were meant to be frequent and some infrequent, which does not help when ranking them overall. As a result, after their selection, I checked to ensure there was a range of both frequent and infrequent but important vignettes included. Further, I

did not use clarity in the total ranking because that metric conveys how much the vignette needs to be revised. In selecting the top eight, I skipped the vignette highlighted in grey in the table because I only had room for one defiance vignette, and I wanted to include the teacher threat vignette because that issue is discussed by Skiba and his colleagues.

The Pictures

A total of 12 judges finished the survey about different characteristics of pictures of 24 students and eight staff; 13 began the survey. One judge did not finish (data kept) and one judge was removed because the judge rated all pictures the same on these categories: "Not Sure" for gender, "Other" for race-ethnicity, and 100 for both physical attractiveness and likability. Because those responses were constant and added more potential uncertainty (i.e., "Not Sure," "Other") that was more systematic than reflective of the actual pictures, it seemed appropriate to remove the judge from this analysis.
I had to adopt a low threshold of 50% to define acceptable agreement on gender, race-ethnicity, and age. I assumed gender could achieve 100% agreement, but it did not. I had to adopt this low threshold to have a large enough pool from which I could randomly select pictures to use in the vignettes, and I surmised that borderline students could pass for the role I assigned them once respondents were cued as to how to perceive them. Table 11 summarizes the roles, pictures needed for those roles, and the number of pictures that were acceptable.

A total of 23 pictures of students were rated reliably for gender, 21 for race-ethnicity, and 11 for age, while a total of 11 pictures of students were rated acceptably across all three. A total of eight pictures of staff were rated reliably for gender, five for

race-ethnicity, while a total of five pictures were rated acceptably across both. Age was not queried for the staff members.

Table 11. Number of pictures needed and available

Vignette	Role	# Needed	# Acceptable
Cyber	Student on computer	1	3
Shooter	Shooter	1	1
GoBack	Male "Foreign" Student	1	2
	Male White Student	1	6
Homophobia	2 Male Students	2	5
Para	Student	1	4
	Paraprofessional	1	2

Pilot Test – Answering RQ #2

The results of the field test were focused on estimating the technical adequacy of the different parts of the instrument to make revisions to it. The results are interpreted in that light. The pilot test afforded the opportunity to test two ways of assessing coherence; I had to choose one. The pilot test also afforded the opportunity to look at how the pictures may affect respondents. In short, the results were encouraging but left room for improvement.

The pictures. The pictures were entered as a second independent factor in the ANOVAs and as a covariate in the robust regressions. As a predictor, it explained significant amounts of variance in nearly every analysis. Those who received the pictures across all three groups scored lower than those who did not; the pictures, as covariate, was negatively related to the outcomes of interest. See Appendix L for a representative

sample of figures that demonstrate this effect. One difficulty in interpreting the effect of the pictures is because the sample sizes of those with and without pictures were uneven. Not only were sample sizes unbalanced, but the proportions were not comparable either; see Table 8 for reminder of sample sizes. For both established administrators and graduate students, more people within those groups received pictures by about 2:1. For the aspiring administrators, the reverse was true, and by a larger margin (~1:5), so it is hard to separate the effects of the pictures from that of the group, especially because only two aspiring administrators received pictures while 11 did not.

I wanted to see if the pictures had any effect at all, hoping they did not. From the figures, it is clear they did, though that effect need not be bias and/or measurement error. It is possible that those with pictures were less secure in their responses, more doubtful, more uncertain in how to approach the situation because it felt more real. I would have to conduct a qualitative analysis of the responses to determine other possible reasons for the discrepancy. I kept the pictures for the field test for two reasons: the effect was fairly consistent across groups (see Appendix L), and I felt it imperative to make sure every respondent conjures the same character in their mind when responding. I reasoned that, despite an effect, everyone would have the pictures, so it would be more of a constant effect, and keeping the pictures would bolster the phenomenological validity of the instrument. As well, future studies could control for racial bias by including a racial attitudes scale or an implicit attitudes test (IAT) at the end of the measure.

Reliability analyses. The reliability analyses estimate the agreement for each variable's individual ratings according to Krippendorf's alpha (KAlpha) and their mean ICCs across groups of judges, each variable's average scores across the unprompted,

prompted, and all the vignettes, as well as the comparable results for each vignette and each vignette's total score.

Variables. Table 12 presents the reliability coefficients for the variables.

Table 12. Reliability coefficients for pilot test variables

	Individual Ratings				Average Ratings		
	All 8 Pilot				Unprompted	Prompted	All
	Kalpha	95% conf. int.	Mean ICC (c,k)	Range	Mean ICC (c,k)	Mean ICC (c,k)	Mean ICC (c,k)
Qual	0.36	(0.34 - 0.38)	0.62	.45 - .73	0.61	0.83	0.81
Crea	0.31	(0.29 - 0.34)	0.59	.52 - .65	0.61	0.74	0.80
Feas	0.19	(0.14 - 0.22)	0.42	.10 - .64	0.22	0.49	0.36
Effct	0.35	(0.33 - 0.38)	0.60	.42 - .71	0.45	0.76	0.79
cohOvr	0.24	(0.20 - 0.28)	0.43	.08 - .62	0.53	0.63	0.69
cohProb	0.37	(0.34 - 0.40)	0.60	.41 - .71	0.54	0.76	0.77
cohGoal	0.21	(0.17 - 0.23)	0.50	.23 - .64	0.62	0.64	0.81
cohVal	0.33	(0.3 - 0.36)	0.59	.36 - .71	0.62	0.75	0.73
cohCon	0.45	(0.41 - 0.47)	0.67	.51 - .79	0.70	0.83	0.81
cohSol	0.24	(0.2 - 0.28)	0.40	.25 - .64	0.34	0.50	0.61
thorProb	0.41	(0.37 - 0.44)	0.65	.47 - .75	0.50	0.63	0.77
thorGoal	0.28	(0.25 - 0.31)	0.55	.41 - .70	0.64	0.79	0.81
thorVal	0.45	(0.41 - 0.48)	0.68	.57 - .78	0.70	0.81	0.80
thorCon	0.44	(0.41 - 0.47)	0.67	.57 - .75	0.66	0.89	0.86
thorSol	0.24	(0.2 - 0.26)	0.53	.38 - .70	0.55	0.63	0.75
cohTot	0.4	(0.37 - 0.42)	0.67	.46 - .81	0.73	0.74	0.86
thorTot	0.45	(0.43 - 0.47)	0.71	.55 - .80	0.75	0.83	0.88
vignTot	0.47	(0.42 - 0.49)	0.71	.52 - .81	0.73	0.82	0.88

On the left side, the table presents KAlpha coefficients and its confidence interval along with the mean ICC across groups of judges and its range. The right side of the table, under average ratings, presents the mean ICCs of scores that have been averaged across the prompted, unprompted, and all eight pilot vignettes (i.e. the reliability for the total scores, across those conditions).

Despite KAlpha being known for generating lower coefficients than other indices, the KAlpha coefficients are still quite low and indicate little agreement among the judges. However, the mean ICC(c,k) suggests that the averages of the judges' ratings demonstrate generally moderate reliability. Feasibility and coherence overall did not achieve even moderate reliability. Their ranges indicate that groups of judges varied widely in their ability to agree but that some groups achieved moderate reliability. When judges' ratings were averaged across the prompted, unprompted, and all vignettes, the mean ICCs generally increased. The prompted vignettes generated better agreement, which is in line with previous research. Further, this result makes sense because the responses are broken into their component parts for the prompted responses. For the unprompted responses, judges must sort through the text to pull out the components of the decision-making process, which requires more judgment on their part, leading to worse agreement. Although it would of course be better for the instrument to demonstrate better reliability, I was encouraged that it appeared to be functioning as predicted.

The variables' reliability for the total scores across vignettes was generally acceptable. Feasibility was poor, which was surprising. Other than feasibility, the other variables demonstrated moderate to good, even excellent reliabilities (Cichetti, 1994).

Inter-item correlation matrix. The inter-item correlation matrix has been included as Appendix M. The matrix shows the variables are related with correlations ranging from the .30s to the .80s. Overall quality, effectiveness, and the vignette total are highly correlated and suggest either quality or effectiveness may be redundant.

Vignettes. Table 13 presents the reliability coefficients for the vignettes and their total scores. The two columns on the left presents the mean ICC and its range for all individual ratings across the variables for each vignette. The next two columns provide the mean ICC and its range for the reliability of each vignette's total score.

Table 13. ICCs for the pilot vignettes

	Individual Ratings		Total Scores		Individual Ratings w/o Coherence Components		Spearman-Brown	
	Mean ICC(c,k)	Range	Mean ICC(c,k)	Range	Mean ICC(c,k)	Range	ICC if doubled	Factor for .80
Defiance	0.53	.45 - .70	0.67	.61 - .73	0.55	.45 - .71	0.71	3.27
Cyber	0.48	.19 - .66	0.60	.08 - .87	0.52	.29 - .58	0.68	3.69
Threat	0.37	.13 - .62	0.41	.12 - .90	0.4	.20 - .65	0.57	6.00
Shooter	0.57	.17 - .74	0.69	.25 - .86	0.55	.10 - .71	0.71	3.27
Elbow	0.67	.57 - .72	0.79	.72 - .86	0.66	.49 - .72	0.80	2.06
GoBack	0.67	.57 - .76	0.80	.68 - .91	0.67	.67 - .73	0.80	1.97
Homophobia	0.47	-.01 - .74	0.56	-.05 - .85	0.46	.12 - .53	0.63	4.70
Para	0.67	.56 - .75	0.79	.68 - .87	0.69	.68 - .71	0.82	1.80
Unprompted	0.49	.33 - .64	0.59	.40 - .81	0.51	.32 - .62	0.68	3.84
Prompted	0.62	.46 - .70	0.73	.52 - .83	0.63	.56 - .68	0.77	2.35
All 8	0.59	.42 - .69	0.71	.52 - .81	0.59	.45 - .70	0.74	2.78

Note. Spearman-Brown figures based on mean ICC when coherence components are removed.

The set of columns on the right provide the Spearman-Brown prophecy ICC if the

number of items were doubled and how many more items would be needed to achieve an ICC of .80. I included how the reliabilities changed after removing the individual coherence components. Generally, the mean ICCs either remained the same or improved slightly, which indicates that removing the coherence components does not degrade the instrument's overall reliability. In fact, it appears to help slightly. Maintaining the instrument's reliability was positive enough for me to decide to remove the components, if only for parsimony.

Validity analyses. Interpreting the results of the validity analyses was murkier than desired for at least two reasons: established administrators performed the worst of the three recruited groups and because of the pictures. Tables 14 - 17 present the validity analyses for the variables and vignettes. Appendix N presents the classic ANOVA and regression summary tables for the pilot test validity analyses.

Variables. The outcome variables were regressed onto the four proxy variables (program enrollment, school-based administrator status, self-rated expertise in addressing student discipline, and years in schools). Unstandardized and standardized coefficients are presented where appropriate; I did not provide standardized betas for the categorical IVs. The achieved power for each analysis is presented as well.

Program enrollment. Table 14 shows the unstandardized mean differences between the groups on each variable. On every variable save the thoroughness of defining the problem, the aspiring administrators performed best. On every variable, save four of them, administrators and aspiring administrators performed differently, indicating the instrument can discriminate between these two groups of respondents on those variables with significant relationships. However, the administrators scored very similarly to the

graduate students, which seems plausible but unlikely. The effect of the pictures is the likely explanation.

School-based administrator status. Table 14 also presents the unstandardized betas for being a school-based administrator. Only six out of 18 variables demonstrated the ability to discriminate. Contrary to expectations, school-based administrators again scored lower, which makes sense because of how many administrators received pictures.

Table 14. Pilot variables' discrimination on program enrollment and current role

	Program Enrollment						School-based Admin		
	Unstandardized Mean Differences			Standardized Effect Sizes		Power	Unstandardized		Power
Variables	Ad – As	Ad – Gr	As – Gr	Eta^2	Cohen's f^2	df(3,379)	beta	(SE)	df(3,380)
Qual	**-9.02****	-4.21	4.82	.03	.19	.91	-2.97	(2.45)	.99
Crea	-5.65	-3.83	1.82	.01	.12	.51	-2.14	(2.57)	.96
Feas	**-6.64***	-0.26	**6.37**$^{\sim}$	.02	.15	.74	1.87	(2.50)	.95
Effct	**-6.29**$^{\sim}$	-3.16	3.13	.01	.12	.55	-2.13	(2.76)	.92
cohOvr	**-6.77***	-0.29	**6.49***	.03	.17	.87	-0.97	(2.04)	.99
cohProb	**-8.81***	**-7.66**$^{\sim}$	1.14	.02	.16	.78	**-6.97***	(3.24)	.99
cohGoal	**-7.06***	-4.43	2.63	.02	.14	.70	-3.73	(2.49)	.98
cohVal	**-11.12****	-5.18	5.94	.03	.18	.90	**-6.49***	(3.13)	.99
cohCon	**-8.58***	**-8.09**$^{\sim}$	0.50	.02	.15	.76	-6.17	(3.39)	.98
cohSol	-2.71	-0.70	2.01	.00	.06	.17	-3.35	(2.13)	.93
thorProb	**-6.60**$^{\sim}$	**-7.36***	-0.76	.02	.14	.71	**-4.64**$^{\sim}$	(2.85)	.97
thorGoal	**-9.07***	-5.32	3.74	.03	.18	.89	**-5.78***	(2.61)	.99
thorVal	**-9.97***	-4.76	5.21	.03	.17	.86	-4.28	(2.88)	.99
thorCon	-8.52	-6.10	2.43	.02	.15	.77	-4.53	(2.98)	.97
thorSol	-3.56	-1.55	2.01	.01	.07	.23	-3.15	(2.42)	.99
cohTot	**-7.66***	**-5.21**$^{\sim}$	2.44	.03	.17	.84	**-5.01***	(2.41)	.99
thorTot	**-7.54***	**-5.02**$^{\sim}$	2.53	.03	.17	.96	**-4.38***	(2.25)	.99
vignTot	**-6.91***	-4.17	2.73	.03	.18	.88	**-3.52**$^{\sim}$	(2.01)	.99

Note. ~ = p < .10; * = p<.05; ** = p <.001.

Self-rated expertise. Table 15 presents the unstandardized and standardized betas and standard errors for the effect of respondents' self-rated expertise in addressing student discipline situations. See Table 9 on page 71 for a reminder of how the groups rated themselves on average. Administrators rated themselves highest, then aspiring

administrators, and then the graduate students. The coefficients show the variable has a small to moderate relationship with all of the variables, except five. The relationship is negative, again possibly due to the pictures reducing the administrators' scores. Respondents self-rated their expertise on a scale from 0-100, so the unstandardized coefficients should be interpreted as a one unit change in their self-rated expertise (e.g., from 49 to 50) is linked with a reduction of about a tenth of a point on that variable. In other words, someone who rated themselves a '0' should score about 10 points lower on the variable than someone who rated themselves a '100'; the standardized coefficients suggest the relationship is weak.

Table 15. Pilot variables' discrimination on self-rated expertise and years in schools

	Self-rated Expertise					Yrs In Schools				
	Unstandardized		Standardized		Power	Unstandardized		Standardized		Power
	Beta	SE	beta	SE	df(3,380)	beta	SE	beta	SE	df(3,380)
Qual	-0.07	(0.04)	-0.07	(0.05)	.99	-0.21	(0.15)	-0.08	(0.06)	.99
Crea	**-0.14****	(0.05)	**-0.15**	(0.05)	.99	-0.21	(0.16)	-0.08	(0.06)	.97
Feas	-0.05	(0.04)	-0.06	(0.05)	.96	-0.05	(0.13)	-0.02	(0.05)	.95
Effct	**-0.10***	(0.05)	**-0.11**	(0.05)	.97	-0.12	(0.15)	-0.04	(0.05)	.92
cohOvr	-0.05	(0.04)	-0.06	(0.05)	.99	-0.02	(0.12)	-0.01	(0.05)	.99
cohProb	**-0.17****	(0.05)	**-0.15**	(0.05)	.99	**-0.48***	(0.19)	**-0.14**	(0.05)	.99
cohGoal	**-0.08***	(0.05)	**-0.09**	(0.05)	.98	-0.14	(0.15)	-0.05	(0.05)	.97
cohVal	-0.09	(0.06)	-0.08	(0.05)	.99	-0.19	(0.19)	-0.06	(0.05)	.99
cohCon	**-0.14***	(0.06)	**-0.12**	(0.05)	.98	-0.29	(0.19)	-0.08	(0.05)	.97
cohSol	-0.03	(0.04)	-0.04	(0.05)	.89	0.00	(0.12)	0.00	(0.05)	.88
thorProb	**-0.15****	(0.05)	**-0.14**	(0.05)	.99	**-0.40***	(0.18)	**-0.12**	(0.06)	.99
thorGoal	**-0.12***	(0.05)	**-0.13**	(0.05)	.99	-0.14	(0.15)	-0.05	(0.05)	.99
thorVal	**-0.11***	(0.05)	**-0.11**	(0.05)	.99	-0.18	(0.19)	-0.05	(0.06)	.99
thorCon	**-0.14***	(0.05)	**-0.13**	(0.05)	.99	-0.17	(0.18)	-0.05	(0.05)	.96
thorSol	**-0.10***	(0.04)	**-0.12**	(0.05)	.99	-0.07	(0.14)	-0.02	(0.05)	.99
cohTot	**-0.10***	(0.04)	**-0.12**	(0.05)	.99	-0.19	(0.14)	-0.07	(0.05)	.99
thorTot	**-0.13***	(0.04)	**-0.15**	(0.05)	.99	-0.22	(0.15)	-0.08	(0.05)	.99
vignTot	**-0.10****	(0.04)	**-0.14**	(0.05)	.99	-0.18	(0.12)	-0.08	(0.05)	.99

Note. ~ = p < .10; * = p < .05; ** = p < .001

Years in schools. Years spent professionally in schools only demonstrated a significant relationship with two variables: coherence of problem definition and thoroughness of problem definition. The inter-item correlation matrix shows these two items are highly correlated so it is not surprising they have similar results. Despite these two relationships, years in schools does not appear to have a strong relationship with performance on the variables.

Vignettes. As with the variables, the vignette total scores were regressed onto the four proxy variables. I report the same standardized and unstandardized statistics.

Program enrollment. Table 16 presents the results of the ANOVAs with program enrollment and pictures as the independent factors and each vignette's total score as the dependent outcome.

Table 16. Pilot vignettes' discrimination on program enrollment and school-based administrator status

	Program Enrollment						School-based Admin		
	Unstandardized Mean Differences			Standardized Effect Sizes		Power	Unstandardized		Power
	Ad – As	Ad – Gr	As – Gr	Eta^2	Cohen's f^2	df (3,44)	beta	(SE)	df(3,44)
Defiance	-3.34	-5.24	-1.90	.02	.14	.12	-0.29	(5.78)	0.26
Cyber	-6.92	-2.96	3.96	.04	.22	.24	-1.90	(3.82)	0.89
Threat	-3.37	0.56	3.93	.01	.12	.10	0.65	(3.75)	0.14
Shooter	**-11.32**˜	-6.23	5.10	.09	.34	.50	-1.80	(5.23)	0.84
Elbow	-0.78	-0.87	-0.09	.00	.03	.05	-1.17	(4.89)	0.24
GoBack	-4.93	-2.92	2.01	.02	.13	.11	-3.56	(4.00)	0.23
Homo-phobia	**-10.19**˜	-4.66	5.53	.09	.32	.45	-5.30	(5.86)	0.58
Para	**-13.47**˜	-10.80	2.67	.11	.36	.54	-5.97	(4.80)	0.73

Note. Ad = Administrator; As = aspiring administrator; Gr = graduate student.

Again, administrators scored lowest; aspiring administrators scored the highest except on the "Defiance" and "Elbow" vignettes. The three groups hardly scored differently at all on the "Elbow" and the "threat" vignettes, making them quick likely candidates to be

dropped. Four of the vignettes, "Cyber," "shooter," "Homophobia," and "Para" demonstrated moderate to large effect sizes. A fifth vignette, "Go Back," also demonstrated a possible small effect. The analyses were underpowered, which places in doubt whether the null effects are indeed null.

School-based administrator status. Table 16 also presents the unstandardized betas and standard errors for being a school-based administrator. None of the vignettes demonstrated an ability to discriminate according to whether the respondent was a school-based administrator or not, though the same four did demonstrate larger effects.

Self-rated expertise. Table 17 presents the standardized and unstandardized coefficients for the relationship between self-rated expertise and the vignettes' total score. Only the fourth vignette, "Shooter," demonstrates a significant relationship with the predictor. For every unit increase respondents' self-rated expertise, they lose about a fifth of a point on their total score. Again, the relationship is negative, which was counter to expectations, and possibly an artifact of the pictures condition.

Table 17. Pilot vignettes' discrimination based on self-rated expertise and years professionally in schools

	Self-rated Expertise					Year In Schools				
	Unstandardized		Standardized		Power	Unstandardized		Standardized		Power
	Beta	(SE)	Beta	(SE)	df(3,44)	Beta	(SE)	Beta	(SE)	df(3,44)
Defiance	-0.06	(.13)	-0.08	(.18)	.29	-0.15	(.37)	-0.07	(.16)	.28
Cyber	-0.08	(.08)	-0.11	(.11)	.88	-0.2	(.25)	-0.09	(.11)	.87
Threat	-0.05	(.08)	-0.07	(.11)	.21	0.21	(.24)	0.09	(.11)	.24
Shooter	**-0.21***	(.09)	**-0.28**	(.12)	.97	**-0.37***	(.27)	**-0.16**	(.12)	.92
Elbow	-0.06	(.10)	-0.08	(.14)	.25	0.09	(.30)	0.04	(.13)	.22
GoBack	-0.08	(.11)	-0.10	(.15)	.23	-0.07	(.32)	-0.03	(.14)	.19
Homo-phobia	-0.08	(.09)	-0.10	(.12)	.53	-0.16	(.27)	-0.07	(.12)	.49
Para	-0.15	(.12)	-0.21	(.16)	.52	**-0.71~**	(.32)	**-0.31**	(.14)	.73

Note. ~ = p < .10; * = p < .05; ** = p < .001

Years in schools and the variables. Table 17 also presents the comparable results for years in schools as a predictor. Again, the "Shooter" and "Para" vignettes demonstrate some statistical significance while the "Cyber" and "Homophobia" vignettes also intimate possible small effects.

In summary, the results of the pilot test were encouraging, but the pictures and the disproportionate sample sizes made interpreting the results more difficult than desired. Feasibility was the least reliable variable and only discriminated according to program enrollment. The other variables showed moderate reliability and an ability to discriminate according to self-rated expertise. The vignettes did not generate as strong reliability as the variables, but the reliability for the vignettes' total scores was generally in the moderate to good range. The "Shooter" vignette and the vignette about a paraprofessional getting injured were the best vignettes at discriminating based on the four proxy variables. Two other vignettes, "Cyber" and "Homophobia," demonstrated potential to discriminate.

Post-Pilot Feedback Survey and Revisions

The results of the pilot test were encouraging and yet still indicated some obvious revisions to make. At the end of the pilot test analyses, I kept five vignettes, removed coherence by component, and made some smaller adjustments. See Appendix O for the results of the post-pilot feedback survey. In short, in terms of the training, the judges reported that the slides were clear, that the examples provided were somewhat or very helpful, and they felt the training to score the responses was sufficient or thorough. They felt the variables were appropriate and explained relatively clearly. Half the judges said they had developed a preference for scoring coherence overall as opposed to by each component; the other half said they had no preference. I used this item directly in

deciding to remove coherence by component, along with the mean ICC values of the vignettes when they were removed as mentioned earlier. Besides this item, the most helpful feedback came from the optional constructed responses about suggestions for improving the training or scoring method.

Besides removing coherence by the components, I made several changes, many of them practical in their nature, for both the respondents and judges. For example, several respondents commented (in an optional qualitative text field after the final question) that they were not sure what I meant by "constraints" or "values" when I prompted them to think through those steps of their decision-making process in the prompted vignettes. Second, I noticed that some non-completers would stop after reaching the first prompted vignette; I surmised they did not like the format and discontinued their participation. A few respondents said they did not like the format outright, so they typed their whole answer into the first prompted text box and left the rest blank, effectively turning the vignette from prompted into unprompted for scoring purposes. In response, I added a page in between the unprompted and prompted vignettes and told respondents the format of the questions was going to change; they would be asked a set of questions, and I provided the definitions of the key terms. I told them they did not have to memorize the definitions because they would be provided with each subsequent vignette. Lastly, I asked them to respond in the different text boxes provided so "their responses could be scored appropriately." I believe these changes were effective in the field test as nobody crammed their whole response into the first text box as they did in the pilot test.

I added a similar set of key definitions for judges to their scoring template, in case they wanted a quick reference, which a few judges said would have been helpful. A few

judges also commented that they wanted a text box for writing comments, so I added that. Some judges used them more than others. Future research should use these comments to conduct a qualitative analysis on characteristics of strong responses. I further used the judges' feedback to address areas of the training that needed further specifying. Please see the log for more details.

Field Test – Answer RQ #3

The results of the field test are focused on estimating the technical adequacy of the final form. The results were not interpreted with the intention of making any large revisions, at most perhaps removing another vignette and/or a variable. In the end, the results pointed toward removing one vignette, and I opted to keep all the variables although an argument could be made to remove either quality or effectiveness. These issues are discussed further in chapter four. Table 18 presents the mean and standard deviation for the variables across vignettes and overall.

Table 18. Variable means across vignettes

	Cyber		Shooter		Go Back		Homophobia		Para		Overall	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Qual	37.70	24.15	41.62	24.09	46.83	26.56	48.83	25.03	48.00	24.67	44.60	25.24
Crea	29.90	25.86	32.98	25.42	38.27	30.33	35.94	28.02	35.51	27.92	34.52	27.67
Feas	58.45	27.11	58.58	24.34	63.67	24.50	65.49	23.39	59.82	23.80	61.20	24.79
Effct	37.57	25.92	50.18	25.45	52.23	26.73	51.18	26.01	51.61	26.04	48.56	26.58
Coh	43.82	21.86	44.62	22.78	55.36	22.35	56.88	22.19	55.91	21.92	51.32	22.94
thorProb	42.65	27.28	33.87	28.31	56.63	27.38	59.11	24.09	53.28	26.63	49.11	28.37
thorGoal	31.56	28.31	29.51	25.87	63.81	24.88	63.94	23.08	61.72	22.88	50.11	29.73
thorVal	30.36	27.08	37.92	28.31	62.43	25.51	58.64	26.81	60.73	22.54	50.02	29.25
thorCon	28.92	27.77	21.11	24.64	60.05	25.38	59.61	22.95	50.98	25.75	44.13	30.02
thorSol	50.95	27.37	63.79	24.31	64.84	25.13	66.53	26.46	63.83	26.39	61.99	26.52
thorTot	36.89	19.96	37.24	18.07	61.55	21.18	61.57	19.88	58.11	20.10	51.07	22.93
vignTot	39.19	18.32	41.42	16.94	56.41	20.07	56.62	19.03	54.14	19.57	49.55	20.28

Reliability analyses. Identical to the pilot test, I report the reliability for each variable and vignette. I examine the individual ratings and total scores for both as well.

Variables. Table 19 presents the reliability coefficients for each variable. Again, feasibility demonstrated the worst reliability across all the analyses. Other than feasibility, KAlpha increased somewhat, still lower than Krippendorf (2011) recommends, though his recommendation is not based on empirical studies.

Table 19. Field test reliability coefficients for the variables

	Individual Ratings Across				Average Scores Across		
	All				Unprompted	Prompted	All
	KAlpha		Mean ICC		Mean ICC	Mean ICC	Mean ICC
	Est.	95% conf. int.	(c,k)	Range	(c,k)	(c,k)	(c,k)
Qual	0.40	(0.36 - 0.41)	0.65	.45 - .74	0.76	0.72	0.79
Crea	0.23	(0.21 - 0.25)	0.56	.38 - .68	0.63	0.74	0.75
Feas	0.08	(0.04 - 0.11)	0.32	.00 - .46	0.36	0.45	0.52
Effct	0.36	(0.33 - 0.38)	0.60	.45 - .69	0.72	0.69	0.74
Coh	0.31	(0.27 - 0.33)	0.55	.22 - .71	0.67	0.52	0.62
thorProb	0.53	(0.51 - 0.55)	0.75	.57 - .89	0.8	0.78	0.78
thorGoal	0.49	(0.47 - 0.51)	0.71	.38 - .89	0.43	0.73	0.73
thorVal	0.52	(0.50 - 0.54)	0.71	.45 - .87	0.51	0.71	0.72
thorCon	0.52	(0.49 - 0.54)	0.73	.49 - .91	0.57	0.74	0.79
thorSol	0.37	(0.34 - 0.38)	0.68	.50 - .79	0.76	0.77	0.84
thorTot	0.59	(0.57 - 0.61)	0.80	.62 - .92	0.81	0.81	0.85
vignTot	0.55	(0.53 - 0.57)	0.77	.66 - .89	0.84	0.79	0.84

Like Kappa, KAlpha is a chance-corrected reliability coefficient. If one uses similar cut points that are recommended for Kappa, then the values fall mostly in the fair to moderate range according to Landis and Koch (1977). Except for feasibility, the mean

ICCs on the individual ratings improved, settling in the moderate to good range based on both Koo and Li's (2016) and Cichetti's (1994) recommendations. Interestingly, the difference between the mean ICCs for the average scores across the unprompted and prompted vignettes narrowed, which probably indicates that the judges became more comfortable picking out the pertinent information from the unprompted responses. Last, the reliability for total scores across all five field vignettes (the last column) is acceptable, especially considering these are constructed responses regarding complex student discipline problems.

Additionally, the thoroughness variables all show quite similar and adequate reliability. These variables are less subjective and were easier to define and operationalize in training. In general, judges were trained to look for how much detail with which each component was discussed. It is encouraging to see the variables generate higher reliability and improve after refining the training.

Inter-item correlation matrix. Last, the inter-item correlation matrix for the field test has been included as Appendix P. The matrix shows the variables are related with correlations ranging from the .30s to the .80s. Just as in the pilot test, overall quality, effectiveness, and the vignette total are highly correlated and suggest either quality or effectiveness may be redundant.

Vignettes. Table 20 presents the reliability figures for each of the five field tested vignettes for the individual ratings and total scores. I include the Spearman-Brown prophecy formula predictions for how the ICC would change if the number of items (vignettes in this case) were doubled and what number of items (vignettes) would be

required to achieve an ICC of .80. The Spearman-Brown formula was based on the mean ICC for the individual ratings.

The analysis of the variables' reliability indicated the judges achieved similar agreement on the prompted and unprompted vignettes, and this analysis of the vignettes confirms that finding. This time, the unprompted actually achieved better reliability, which is probably because the prompted scores included the vignette "Go Back" which had the lowest reliability of the five.

Table 20. ICCs for the field test vignettes

	Individual Ratings		Total Scores		Spearman-Brown	
	Mean ICC(c,k)	Range	Mean ICC(c,k)	Range	ICC if doubled	Factor for .80
Online	0.65	.58 - .80	0.81	.68 - .91	0.79	2.15
Shooter	0.66	.53 - .82	0.74	.64 - .80	0.80	2.06
GoBack	0.57	.40 - .74	0.64	.46 - .86	0.73	3.02
Homophobia	0.67	.49 - .85	0.77	.56 - .94	0.80	1.97
Para	0.64	.20 - .79	0.66	-.02 - .86	0.78	2.25
Unprompted	0.65	.55 - .81	0.84	.76 - .91	0.79	2.15
Prompted	0.62	.43 - .75	0.79	.53 - .96	0.77	2.45
All 5	0.68	.53 - .80	0.84	.63 - .94	0.81	1.88

Note. Spearman-Brown formulae based on mean ICC of individual ratings.

Validity analyses. Interpreting the validity results for the field test was more straightforward as all respondents had the pictures; administrators scored highest on most metrics, suggesting the pictures complicated the validity results in the pilot test. The identical analyses were run, except the pictures were not entered into the ANOVA for program enrollment, nor as a covariate in the analyses for school-based administrator status, self-rated expertise, and years professionally in schools. Appendix Q presents the classic ANOVA and regression summary tables for the field test validity analyses.

Checking for normality and equal variances. Appendix R presents the univariate plots for each variable. Creativity is skewed to the right, while Feasibility is skewed to left. The others all approach normality. Using Levene's test where each variable was the outcome and each categorical proxy variable was the independent factor, variances were homogenous every time, except for feasibility and program enrollment and creativity and school-based administrator status; however, using the robust estimator should help correct for this violation of homogeneity of variances. Appendix S presents these results for every variable and categorical proxy variable. Appendix T presents the qq plots for the vignettes and variables that were regressed onto school-based administrator status, self-rated expertise, and years spent professionally in schools.

Variables. The individual ratings of the variables, not the totals, were regressed onto the four proxy variables using the same analyses as in the pilot test. A separate round of analyses was run with the total scores for the final form only, see Appendix U. I did not want to run the total scores until there was a final form because including unreliable or non-discriminating vignettes would lead to worse estimates of the final form's technical adequacy.

Program enrollment. Table 21 presents the unstandardized mean differences between the groups of respondents on each variable. Unlike the pilot test, aspiring and established administrators scored quite similarly to each other, no significant differences on any variable. However, it appears that both groups were significantly different from the graduate students on quality, feasibility, and the thoroughness of discussing their solutions, and nearly both groups were different on effectiveness. On the whole, the variables do not do a good job discriminating according to respondents' program

enrollment. Appendix V presents the discriminant validity relationships with the categorical proxy variables.

Table 21. Field variables' discrimination based on program enrollment and school-based administrator status

	Program Enrollment						School-based Admin		
	Unstandardized Mean Differences			Standardized Effect Sizes		Power	Unstandardized		Power
Variables	Ad – As	Ad – Gr	As – Gr	Eta^2	Cohen's f^2	df(2,587)	beta	(SE)	df(2,588)
Qual	0.25	**4.56**~	**4.31**~	.01	.11	.66	**6.15****	(2.03)	.76
Crea	-1.41	0.08	1.49	.00	.03	.09	1.37	(2.10)	.08
Feas	-0.68	**4.50***	**5.18***	.01	.13	.84	2.70	(1.69)	.28
Effct	-0.90	4.19	**5.09**~	.03	.11	.65	**6.21****	(2.14)	.72
Coh	-2.16	1.36	3.52	.01	.08	.38	0.74	(1.83)	.06
thorProb	-3.49	-1.25	2.24	.00	.05	.17	-3.35	(2.77)	.19
thorGoal	4.04	3.03	-1.02	.00	.06	.24	**5.76***	(2.82)	.47
thorVal	0.35	-0.60	-0.95	.00	.02	.06	3.20	(2.82)	.17
thorCon	-0.24	1.39	1.64	.00	.03	.09	2.70	(2.91)	.14
thorSol	-0.51	**4.75**~	**5.26***	.01	.12	.72	**6.15****	(2.03)	.74
thorTot	0.03	1.46	1.44	.00	.04	.11	3.13	(2.12)	.26
vignTot	-0.48	2.20	2.68	.00	.07	.32	**3.41**~	(1.78)	.39

Note. ~ = p < .10; * = p<.05; ** = p <.001. Ad = Administrator; As = aspiring administrator; Gr = graduate student.

School-based administrator status. Table 21 also presents the results of the analysis on whether the variables can discriminate between those who are school-based administrators and those who are not. This proxy variable does a better job predicting the variables than program enrollment. Again, quality, effectiveness, and the thoroughness of solutions demonstrated significance and the largest effects. The thoroughness of their goals and the total score overall indicated a significant relationship when alpha is set to .10. These betas indicate that school-based administrators scored about five to six points higher than the others on quality, effectiveness, thoroughness of their goals and solutions, and about three and a half points overall. Interestingly, school-based administrators scored about 3.35 points worse on how thorough they defined the problem. Perhaps that

occurred because the administrators were less engaged or the problems were more obvious, meriting less discussion in their minds. Or possibly, it is that non-school-based administrators were so inexperienced that they felt the need to talk through the problem more so they could define it for themselves. Other than that variable, school based administrators scored higher than everyone else, although the differences did not always achieve statistical significance.

Self-rated expertise. Table 22 presents the standardized and unstandardized betas for each variable's regression onto respondents' self-rated expertise. See Table 9 on page 71 for a reminder of the average self-ratings for the three different groups of respondents. On average, as would be expected, administrators rated themselves highest, then aspiring administrators, then the graduate students. The same variables show an ability to discriminate according to self-rated expertise: quality, feasibility, effectiveness, thoroughness of solutions and the overall total. The effects seem small to moderate, where one unit increase of self-rated expertise on scale of 0-100 results in about a tenth of a point increase in the scores of the variables with significant relationships. The effect seems about the same size as the difference between being a school-based administrator and not, about 5-6 points if the respondents self-rated their expertise about 50-60 units lower. Appendix W presents the discriminant validity relationships with the continuous proxy variables graphically.

Years professionally in schools. Table 22 also presents the standardized and unstandardized coefficients for the variables' relationship with respondents' years in schools. None of the variables could discriminate according to this proxy for being a novice or expert.

Table 22. Field variables' discrimination based on self-rated expertise and years professionally in schools

	Self-rated Expertise					Years in Schools				
	Unstandardized		Standardized		Power	Unstandardized		Standardized		Power
	beta	SE	beta	SE	df(2,588)	beta	SE	beta	SE	df(2,588)
Qual	**.10****	(.03)	**.13**	(.04)	.76	.01	(.10)	.01	(.04)	.05
Crea	.03	(.03)	.04	(.04)	.11	-.11	(.11)	-.04	(.04)	.13
Feas	**.09****	(.03)	**.14**	(.04)	.88	.03	(.09)	.01	(.04)	.06
Effct	**.11****	(.04)	**.14**	(.04)	.83	.08	(.11)	.03	(.04)	.09
cohOvr	.04	(.03)	.07	(.05)	.26	0.0	(.10)	0.0	(.05)	.05
thorProb	.00	(.04)	.00	(.05)	.05	-.17	(.14)	-.06	(.05)	.19
thorGoal	.02	(.04)	.02	(.04)	.07	0.0	(.15)	0.0	(.05)	.05
thorVal	.02	(.04)	.03	(.04)	.08	-.12	(.14)	-.04	(.04)	.11
thorCon	-.02	(.04)	-.02	(.04)	.08	-.13	(.14)	-.04	(.04)	.12
thorSol	**.08***	(.04)	**.10**	(.05)	.53	.07	(.12)	.03	(.04)	.08
thorTot	.02	(.03)	.03	(.04)	.09	-.07	(.10)	-.03	(.04)	.09
vignTot	**.05**$^{\sim}$	(.03)	**.07**	(.04)	.32	-.04	(.08)	-.02	(.04)	.06

Note. ~ = p < .10; * = p < .05; ** = p < .001

Vignettes. As in the pilot test, each vignette's total score was regressed onto each of the four proxy variables. Overall, the vignettes did not discriminate as well as the variables did.

Program enrollment. Table 23 presents the unstandardized means between the three groups on each vignette total score. None of the vignettes could discriminate between the groups to the degree needed to achieve statistical significance, although the vignette named "Para" came close, $p = .15$. These analyses are underpowered to reject the null for small effects.

School-based administrator status. Being a school-based administrator was slightly better, achieving significance with the "Homophobia" vignette and the "Shooter" and "Para" vignettes demonstrated similarly practical effect sizes. These same three vignettes continually perform the best, and from a practical standpoint, their effect sizes are similar.

Table 23. Field vignettes' discrimination based on program enrollment and school-based administrator status

	Program Enrollment						School-based Admin		
	Unstandardized Mean Differences			Standardized Effect Sizes		Power	Unstandardized		Power
	Ad – As	Ad – Gr	As – Gr	Eta^2	Cohen's f^2	df (2,115)	beta	(SE)	df(2,116)
Cyber	-0.32	-0.36	-0.04	.00	.01	.05	-1.74	(3.41)	.07
Shooter	2.80	3.96	1.16	.01	.12	.20	4.64	(3.40)	.24
GoBack	-3.29	0.60	3.89	.01	.10	.15	1.35	(3.37)	.06
Homo-phobia	1.62	2.95	1.33	.01	.08	.11	**6.57**~	(3.45)	.38
Para	-3.20	3.85	7.05	.03	.18	.39	5.07	(3.32)	.22

Note. Ad = Administrator; As = aspiring administrator; Gr = graduate student.

Self-rated expertise. Table 24 presents the standardized and unstandardized coefficients for self-rated expertise' ability to predict each vignette's total score. Again, the "Shooter" and "Para" vignettes discriminated, and it appears that self-rated expertise acts as the best proxy variable for identifying novices and experts.

Years professionally in schools. Again, as with the variables, years spent professionally in schools was not related to any of the vignettes' total scores.

Table 24. Field test vignettes' discrimination based on self-rated expertise and years professionally in schools

	Self-rated Expertise					Year In Schools				
	Unstandardized		Standardized		Power	Unstandardized		Standardized		Power
	beta	(SE)	beta	(SE)	df(2,116)	beta	(SE)	beta	(SE)	df(2,116)
Cyber	0.0	(.05)	0.0	0.08	.05	-0.14	(.13)	-0.07	(.06)	.10
Shooter	**0.08**~	(.05)	**0.12**	0.07	.30	0.18	(.16)	0.09	(.07)	.16
GoBack	-0.02	(.05)	-0.03	0.08	.06	-0.12	(.18)	-0.05	(.08)	.08
Harass	0.08	(.05)	0.13	0.08	.27	-0.03	(.18)	-0.02	(.08)	.05
Para	**0.11***	(.05)	**0.18**	0.09	.43	0.04	(.18)	0.02	(.09)	.05

Note. ~ = p < .10; * = p < .05; ** = p < .001

Final form. After all the analyses, I decided to remove the third vignette, "Go Back," because it failed to discriminate along any of the proxy variables, and it had the lowest reliability of all the vignettes. I then ran the reliability analyses again without that vignette and the reliability coefficients improved slightly. I also ran the analyses with removing the cyber-bullying vignette, which also did a poor job discriminating, but the reliability coefficients worsened after removing it, so I kept it, leaving the instrument with two unprompted and two prompted vignettes. Appendix U contains the final round of reliability and discriminant validity tables with the coefficients for the final form. In short, the results essentially the same; point estimates shifted a bit, but, substantively, the inferences do not change. The instrument overall demonstrates good reliability, but its ability to discriminate is moderate. It best discriminates between people who are school based-administrators and those who are not as well as along respondents' self-rated expertise in addressing student discipline. The instrument discriminates poorly between people enrolled in master's in teaching, PreAL, and ProAL students, and it does not discriminate at all according to respondents' years spent professionally in schools.

Other Findings of Interest

There are a few other findings of interest in the data. Being male was not related to vignette total score, kendall's $\tau = -.02$, $p = .63$, which is an encouraging finding. Being white, however, was related to vignette total score, kendall's $\tau = .14$, $p < .001$. which is concerning. This variable was coded White (1) and not-White (0). The sample was largely white (~80%), and I did not have the cell size to conduct proper comparisons, but this relationship is worth exploring in future research. Age was negatively related to vignette total score, $r = -.10$, $p = .03$, which means that for every year increase in age,

respondents lose about a tenth of a point on their total score. Although administrators usually had the highest scores, we can also see that younger respondents had better scores as well, suggesting that younger administrators provide better responses. Appendix X presents the relationship between age and total score graphically.

CHAPTER IV

DISCUSSION

The purpose of this study was to develop and validate a measure of administrator decision-making in student discipline. Specifically, this study evaluated the content validity, technical adequacy after a pilot test, and technical adequacy of the instrument after field testing. Based on the field test results, ADMin-SD is comprised of two unprompted and two prompted vignettes. Using a two-way mixed, consistency, average measures ICC (McGraw & Wong, 1996), its total score demonstrates reliability of .84, which is in the good (Koo & Li, 2016) to excellent (Cichetti, 1994) range. The instrument's total score discriminates between those who are school-based administrators and those who are not (unstandardized beta = 3.84, p = .06), and its total score also can discriminate high and low proficiency performers according to respondents' self-rated expertise in addressing student discipline situations (b = .10, p = .04). Given the absence of any performance measure that can assess principal decision-making in student discipline situations, ADMin-SD fills a critical gap in the literature. From a measurement perspective, however, limitations and other issues with ADMin-SD's development and validation should be discussed for future researchers to consider when using it. In this chapter, I begin with the limitations to put the rest of the discussion in context, then I discuss the major findings related to the instrument's content validity, pilot test, and field test. Finally, I discuss the study's implications for researchers and practitioners.

Limitations

There are several limitations in the study that also, in turn, point to directions for future research. First, the study was underpowered. The second potential limitation lies in

the sampling procedure. Third, the vignette writing process may have been a large weakness of the study. Fourth, the sampling related to the pictures presented some limitations in the pilot test.

The required sample sizes for both the pilot and final validation studies indicated that I needed to recruit in total about half of the 200 current principals in the state of Oregon. Falling short of this number, the discriminant validity analyses were substantially underpowered to reject the null for small effects.

Using program enrollment as the basis of recruiting respondents presented both advantages and disadvantages. The main disadvantage is that I used these groups as a proxy for true novices, novices, and experts when the groups are more of a proxy for experience. Allison and Allison (1993) found that experience was linked to expertise but did not guarantee it, much like Kennedy (1987) explained. In other words, the respondents who have been identified as experts and novices may not in fact have been so. Recruitment of graduate students as *true novices* helped catch true novice responses, but the measure would be most useful if it was sensitive enough to discriminate between aspiring and established principals. This recruitment strategy, thus, allowed the opportunity to see if the instrument was sensitive enough to discriminate between aspiring administrators and established administrators, as well as allowing comparison of all administrators against everyone else.

Moreover, there is scant evidence (Brenninkmeyer & Spillane, 2008) that expert administrators are also expert decision-makers. However, this study provides some evidence that respondents who perceive themselves as adept in addressing student discipline problems were also rated higher by the judges in this study, suggesting that

more experienced administrators and aspiring administrators are indeed better decision-makers in this domain of their practice.

Taking extra time upfront to interview expert administrators may have resulted in vignettes that would have discriminated better. I could have asked them to identify problems to which they have learned to respond differently from when they were novice administrators. Conducting those initial interviews may have helped identify specific instances (like the injury to the paraprofessional) that elicit more differences between respondents.

In some ways, the effect the pictures had on the pilot test results is one of the more interesting findings of the whole study. Ideally, from a measurement perspective, the pictures would have had no effect, so they would not introduce measurement error. Unfortunately, that was not the case. Indeed, it may have been unreasonable to hope for no difference. The pictures add another layer of realism to the vignettes; we should expect an effect. As shown in Appendix L, the effect appears to be constant across the groups, which is consolation that, if there is an effect, at least it is constant. Future research should conduct studies with larger and more proportionate samples to examine the effect that the pictures have on responses. Despite these limitations, this study is the first to develop and validate a theoretically-grounded quantitative measure of administrator decision-making in student discipline.

Content Validity – RQ #1

Despite not taking more time upfront to write the vignettes, the judges' assessed them to be very realistic and representative of problems that principals face. For example, beyond the quantitative scores they gave, one judge said about the cyber-bullying

vignette, “This is about as real as it gets.” About the same vignette, another judge said, “This is a good vignette and has a higher level of difficulty as administrators will need some background knowledge in school law to understand the community activity timestamp and the bleed in to the school environment.” Three of the four vignettes that made the final form were ranked first, second, and third, while the fourth vignette ranked seventh. When developing future forms, within student discipline or in other sub-domains of principals’ practice, it would be wise to spend time up front with experts to identify those instances in which they have learned to respond differently than when they were a novice; those kinds of events should discriminate well.

Pilot Test – RQ #2

The pilot test lead to some interesting findings, not only regarding the pictures. I tested scoring Coherence two ways. Scoring Coherence by component was simply too much to ask of the judges; the cognitive load required was too much, especially combined with making other ratings. Judges were asked to consider whether each component of the decision-making process (i.e., problem definition, identifying goals, values, constraints, solutions) was aligned to the whole response; if the component was absent, then that component should get a zero because it cannot be aligned if it is not there. While it may make some sense to operationalize the variable this way, it was simply too complicated to ask of the judges. I surmised that the overall reliability would increase with the field test if only because what they were asked to do was less complicated and required less judgment on their part.

Besides getting rough estimate of the instruments’ reliability and discriminant validity, the qualitative feedback I collected from judges and respondents was the most

helpful in shaping the structure of the overall instrument. For example, using feedback from respondents that did not like the prompted format to add the page of directions and definitions could have never been uncovered if I had only looked at the statistical analyses. I believe this change contributed to a higher rate of completion by respondents and ensured that respondents understood the terms being used in the assessment, which again speaks to the instrument's phenomenological validity.

Field Test – RQ #3

The field test offered the opportunity to estimate the instrument's technical adequacy with respect to its reliability and discriminant validity with a larger sample and with all respondents receiving the pictures. Important issues related to measurement of this construct in general and specific issues related to the development and validation of this instrument emerged with respect to the variables, vignettes, and what I have been calling the proxy variables.

The variables. There are several specific issues related to the variables that merit discussion including: the variables' ability to discriminate, Feasiblity's reliability, and the possible redundancy of Quality and Effectiveness. Ideally, in terms of the instruments' construct validity, every variable should discriminate. That was not the case here; however, that may not have been a failing of the individual variable or instrument overall. For example, Creativity did not discriminate once, but it is plausible that administrators and non-administrators are not actually different in their levels of creativity; there is no reason to assume that being an administrator automatically confers one with greater creativity. I had reasoned that more experienced administrators would have had more opportunities to see creative ways of addressing student discipline that they could adopt,

but that was not the case. Given that it was moderately reliable, Creativity is worth keeping to determine if the trait can be enhanced by professional development. Indeed, the results on creativity's reliability are directly in line with reliability achieved by prior research for this variable (Amabile, 1971; Amabile, 2011). Moreover, creativity is an important variable in theory as proposing solutions is a creative act in and of itself (Davis, 1966); thus, removing creativity as a sub-scale could lead to an unmeasured source of variance.

The thoroughness of four of the five components did not discriminate either, only the thoroughness of solutions discriminated between school-based administrators and non-administrators, which makes sense because that is where most administrators could show their expertise by offering in depth discussions of what they would do, why they were doing it, and how they would accomplish it. But, as with creativity, the lack of discrimination along the other four components of the decision-making process may not be a failing of the instrument, it may be that the groups are not different, or that differences will emerge after an intervention is delivered for instance. It makes sense that the thoroughness of defining the problem, stating goals, clarifying values, and foreseeing constraints did not discriminate because, as found in prior literature (e.g., Leithwood & Steinbach, 1989), most decision-makers do not consider those aspects; they skip ahead to solutions. Expert decision-makers, however, take their time defining the problem because the more specific a problem is defined, the easier it is to solve (Nutt, 1993). I kept these four components of the decision-making process because they demonstrated adequate reliability and they may be able to detect differences after intervention.

Feasibility demonstrated the lowest reliability; given the pilot test, this finding should not be surprising. This variable should have been one of the easier ones upon which to agree because judges were asked to rate how practical the response was. They had to assess issues such as: What solutions were proposed? How many steps were involved? How difficult were those steps to execute? Were the stated goals practical, achievable? Measurement error may have been large, but there also may be substantive variability in how judges viewed the practicality of the steps to address the problem(s) and solution(s). Despite the low reliability, I decided to keep Feasiblity for a few reasons. First, again, this variable is important in theory. Second, it discriminated well (but it was not reliable so that discrimination should be discounted). Third, the variable should be able to achieve better reliability as in prior research (Zaccaro & Mumford, 2000), so efforts will be made to improve the training, and by extension its reliability.

The last issue to discuss regarding the variables is my decision to keep both quality and effectiveness when they may be redundant. Effectiveness is important in theory; effectiveness, creativity, and feasibility work together – a response may be creative and/or effective, but not feasible which makes it virtually worthless. Quality, however, was more reliable and provides an overall score. While effectiveness is a more specific dimension to assess, it is closely related to the overall quality of the response. Because the issue cannot be settled obviously, I decided to keep both to see if a factor analysis generates new evidence to decide if one should be dropped.

The vignettes. The decision to drop vignettes, the lower reliability than the variables, and the ability to discriminate were crucial matters to work through to develop and validate the instrument. The decision to drop "Go Back" was straightforward as it

had the lowest reliability and did not discriminate along any of the proxy variables. After removing it, reliability for the overall instrument was enhanced on the variables by about .02 - .04 on the mean ICC (see Appendix U for final form reliability coefficients). As well, the core of the problem(s) presented in the vignette were nearly identical to those presented in the homophobia vignette, on which both respondents and judges commented. The homophobia vignette had an extra wrinkle in it about providing support to the student knowing that the student may not want to disclose his thoughts and feelings to his parents.

Although the reliability for the vignettes was in the moderate (Koo & Li, 2016) to good (Cichetti, 1994) range, there were a few groups of judges who did not agree well, as evidenced by the ranges of ICCs that comprise the means. For instance, there was one group of judges that demonstrated about as much as agreement as would be expected by chance on Feasibility. Future studies should systematically examine the extent to which the judges themselves may have entirely different approaches to how the situations should be handled.

Three of the four vignettes showed evidence they could discriminate. Two discriminated according to self-rated expertise and the third according to school-based administrator status, but all three had similar effect sizes from a practical standpoint. The vignettes that discriminated involved a school shooting incident, an injury to the paraprofessional, and an incident of homophobia that lead to a fight; the fourth vignette involves a case of cyber-bullying. It is curious that the Cyber-bullying vignette did not discriminate, even though it seemed like there were two basic kinds of answers. Either the respondent said it was not a school issue since it was happening off school grounds,

outside of school hours, and without school property. Other respondents said it was a school issue because it was affecting the student's desire to come to school. It is plausible that there is another variable that could act as a proxy that I did not check.

The vignette about the para injury, however, was strong in discriminating; it allows for the demonstration of more domain-specific knowledge. Without conducting a systematic qualitative analysis of the responses, it seems the best answers to the school shooter vignette referenced ALICE protocols, and the best responses to the para injury vignette involved references to Human Resources issues with which experienced administrators would have experience while most educators, or pre-service teachers for example, would not. The best answers to the homophobia vignette may include a sensitivity to the students' desire not to disclose to his parents, with which only experienced administrators would have experience.

The proxy variables. With reason to believe that all four proxy variables (program enrollment, school-based administrator status, self-rated expertise, and years of professional experience in schools) would be related to respondents' performance, it was interesting to see how they performed. Self-rated expertise appears to have been a decent, if small, predictor of respondents' scores. Contrary to the literature that novices overestimate their expertise (e.g., Kruger & Dunning, 1999), the sample in the field test appears to have estimated its own ability well, at least according to the judges in this study. These data should be collected in future studies to use as a predictor or covariate depending on the research question. However, perhaps this variable is more a measure of respondents' motivation or sense of self-efficacy in addressing student discipline rather than their actual expertise. Future research should disentangle these possible confounds.

Being a school-based administrator (i.e., current principal or asst. principal) was also a good predictor of respondents' scores on the variables, which is positive evidence of the measure's construct validity. Unfortunately, it would be ideal if the measure could discriminate between aspiring and established administrators, which is a more selective comparison. Currently, the instrument does not appear sensitive enough to detect those differences; a larger, more representative sample should help answer this question. Future research should monitor this variable's relationship with scores on the variables.

Program enrollment was not a good predictor, but that is likely an artifact of the variability within those groups as shown by Tables 3 and 4. The graduate student group for example was comprised of both 22-year-old graduate students who had never been in a classroom as well as general education teachers adding a master's degree. Moreover, general education teachers could be found in all three programs, so the groups were not precise enough in providing distinct populations, thereby hurting the variable's ability to act as a predictor.

Years of professional experience in schools was generally not related to any of the dependent variables. I interpret this result more as a statement about the weakness of years in schools as a proxy for who is stronger and who is weaker rather than as a failing of the instrument. Indeed, Kennedy (1987) notes that experience is not equated expertise; one must learn from experience to convert it into expertise.

Implications for Researchers

With a constructed response, quantitative measure of administrator decision-making for student discipline situations, researchers are able to conduct various kinds of studies including qualitative, quantitative, and mixed methods studies. Due to this

flexibility, ADMin-SD is a useful tool to answer a range of research questions. Eventually, policy-makers should be able to use these substantive results to inform the issues they address, the research they fund, or policies they consider.

Future research. Specifically, there are at least two issues that emerged from this study that should be addressed by future research. The first is to conduct a factor analysis, which can be done with the current data set, to see how the variables hang together, and to see if Quality and Effectiveness are indeed redundant. The second issue must be addressed by a separate study, which should recruit a larger and more diverse, representative sample of the population to evaluate the instrument, specifically examining how the measure functions with respect to race-ethnicity. In this study, being White was related to vignette total score, kendall's $\tau = .139$, $p < .001$. This undesirable relationship should be explored to understand why it exists so it can be attenuated.

Beyond these two follow-up studies, ADMin-SD can be used to test interventions designed to improve school-based administrators' decision-making skills in student discipline; however, a second form should be made if the intervention testing requires a pre/post design. Leithwood and Steinbach (1992) conducted a study in which they taught the decision-making model to school principals who showed increases in their decision-making skills, but they were not always statistically significant differences. The first intervention to test would be whether or not simply using the decision-making model that ADMin-SD is based on would result in better decisions. In this study, responses to the prompted vignettes received higher scores, particularly in how thorough respondents were in their decision-making process. This finding suggests that being prompted to think

through the different components of the model results in a more thorough decision-making process.

In addition to testing decision-making interventions, ADMin-SD can also be used to evaluate how school-based administrators' responses to the vignettes are different when faced with different characters in the vignettes. To answer this kind of question, ADMin-SD should be administered to equivalent groups of principals, where the different groups are administered different forms where the pictures, names, race-ethnicities, and genders of the characters involved in the vignettes have been experimentally altered. For instance, the two groups would receive the exact same text; however, in one form, the character's name could be Thomas and in the other form, the character's name could be Tyrone. This experimental paradigm is becoming common in sociology and human resources. For example, Bertrand and Mullainathan (2004) examined whether people were likely to be hired based on their names (i.e., Emily vs. Lakisha). Although I would expect differences to emerge, as in other professions and based on the disproportionate student discipline outcomes, ADMin-SD may be able to uncover some specific differences in school-based administrators' thought processes related to this phenomenon.

Lastly, the instrument can be used to compare groups' ability to address student discipline compared to individuals. The group would respond together, and the group would get a score, just as an individual did in this study. Studies that examine these differences can evaluate whether decisions are more creative and better quality if made by a group or individual. Based on prior research (e.g., Paulus, 2000), groups should be more creative; however, they may take longer to come to their decision. Researchers can

appraise the extent to which the extra time is beneficial. Perhaps individuals make nearly as good decisions but in much less time, which would indicate greater efficiency.

Implications related to training and evaluation. This study had offers insight related to administrators' training and evaluation. As discussed above, it appears that respondents benefited from using the decision-making model as their scores were enhanced when prompted to think through the different steps of the model, which suggests that school-based administrators should be trained to use the model. Beyond that, it also appears people espoused answers that were more aligned with experts' judgments of the appropriate responses to the vignettes, given that age was negatively related to total score. The judges in this study averaged an age of 54, and three were retired, so there is no generational bias working between the judges and respondents. In other words, the judges did not only view the responses of people from their generation as strong; to the contrary, they took that view of respondents from younger generations. In general, these younger people have received training more recently, which perhaps suggests that preparatory programs are doing a good job. This finding may also suggest that experienced administrators are stuck in their ways, not adopting newer practices, and should receive current training.

Implications for Practitioners

With a quantitative measure of administrator decision-making for student discipline situations, school and district leaders can identify who are strong decision-makers in addressing student discipline, who needs training to improve, if the training was effective, possibly as a screener in the hiring process, and as a tool to facilitate difficult conversations between administrators. Licensure preparatory programs can use

the instrument in various ways as well. Judges and respondents have commented on this throughout the study. For example, in a qualitative text box seeking feedback at the end of the survey, an aspiring principal wrote:

> These are great scenarios and I wish they had been brought up in our institutes with possible solutions because they are tough. I felt that they are current, relevant and addressed ideas that we will most likely encounter during our time as administrators.

There were several comments like this, including from the pre-service teachers who appreciated the opportunity to think from a principal's perspective. From the other perspective, ADMin-SD offers judges the opportunity to assess respondents' thought process and espoused values in dealing with these difficult student discipline issues. In fact, one judge who is also an asst. superintendent remarked, "Scoring the scenarios is frightfully interesting and informative." Clearly, this judge read some concerning responses that could be helpful to district leaders when developing and training their principals and asst. principals.

Judges also offered different uses for the instrument that I had not yet considered. One judge remarked that "it would be a nice addition to the screening process when hiring. Several judges thought the instrument could "be useful as a continued licensure training component" with one judge "wish[ing] there was a class using this instrument [that was] required for administrative licensure." One judge thought of an interesting way to use the instrument to facilitate difficult conversations to help bring a group to consensus in applying intervention:

> It would be interesting to have administrators scoring other administrators' responses. It would be good conversation. Having scenarios to work through that are real life, thought provoking and multifaceted can aid in a more uniform application of interventions with fidelity. It would be great to have this as a tool at ODE or when presenting at COSA.

As the tool allows for multiple strong responses with the constructed response format, it offers the opportunity for rich discussion of multiple ways to approach the same problem. As this judge has implied, peers can use it to challenge and learn from each other.

Conclusion

ADMin-SD is grounded in theory from both educational leadership and decision-making; it has demonstrated adequate construct validity in its reliability, content validity, and discriminant validity. ADMin-SD can be used for several applications, including for both researchers and practitioners. Developing and validating ADMin-SD has laid the foundation for a program of research to follow that will ultimately help improve administrators' ability to make effective, efficient, and equitable decisions.

APPENDIX A

SUMMARY TABLE OF LITERATURE REVIEW TOPICS

Authors	Year	Study Type	Sample Size	Sample Composition	Expert Identification	Type of Problem	Collected Data	Findings	Theme
Leithwood & Stager	1989	Qual	22	6 experts, 16 typical	Reputation + Leadership Survey	6 Vignettes	Interview	Generated grounded, descriptive model of DM process. Found differences between groups.	Experts & Novices
Begley & Leithwood	1990	Qual	15	2/3rds of elem Ps from 1 urban district	N/A	From Practice	Interview	Examined influence of values on process. Ps rely on values of consequence and consensus, i.e. outcomes and social support.	Personal
Leithwood & Steinbach	1990	Qual	11	all expert	Reputation + Leadership Survey	From Practice	Interview	Interviews focused on how Ps gauge difficulty of problems, award priority, and determine how/when to involve others.	Skill
Leithwood & Steinbach	1992	Mixed	38*	22 Treatment; 16 Control	N/A	4 Vignettes	Written responses	Taught DM skills. Experimental group showed more growth but not always statistically sig.	Construct
Leithwood & Steinbach	1993	Quant	9	all expert	Reputation	From Practice	Interview	Correlated performance with leadership characteristics based on teacher surveys. Possible positive relationship but unclear evidence.	Construct
Allison & Allison	1993	Quant	32	6 veteran, 7 seasoned, 8 rookie, 8 aspiring, 10 true novice	Years of Experience	1 Vignette	Interview	Examined link between experience and expertise; found link is not linear. Level of abstraction and attention to detail positively related to judged expertise.	Skill
Bullock, James, & Jamieson	1995	Qual	28	13 novice, 13 typical, 2 new-to-post	Credentials + Reputation	From Practice	Interview	Experts and novices differed along 3 dimensions: delegation, facing conflict, approach to DM.	Experts & Novices
St. Germaine & Quinn	2005	Qual	6	3 expert, 3 novice	5 yrs experience + Reputation	From Practice	Interview & Obs.	Examined role of tacit knowledge in process. Use of tacit knowledge distinguishes experts from novices.	Experts & Novices

Authors	Year	Study Type	Sample Size	Sample Composition	Expert Identification	Type of Problem	Collected Data	Findings	Theme
Lazaridou	2007a	Qual	10	all expert	Reputation	5 Vignettes	Think-aloud	Coded responses for archetypal strategies and domain-specific processes. Found Ps used 4 strategies and used similar processes in extant literature.	Skill
Lazaridou	2007b	"	"	"	"	"	"	Examined influence of values on process. Found 7 distinct values.	Personal
Lazaridou	2009	"	"	"	"	"	"	Examined types of knowledge used: knowledge of org., of task, of people, & tacit knowledge.	Skill
Brenninkmeyer & Spillane	2008	Quant	36	20 experts, 16 typical	Teacher surveys & standardized test scores	6 Vignettes	Interview	No differences between groups on personality measures. Found small differences based on content of vignette. Demonstrated some differences with novices, some corresponding to previous findings but not all.	Experts & Novices
Spillane, White, & Stephan	2009	Quant	44	20 experts, 24 aspiring	"	6 Vignettes	Interview	Tested differences between experts' and aspiring Ps use of processes. Only 5 of 22 processes showed differences.	Experts & Novices
Goldring, Huff, Spillane, & Barnes	2009	Quant	48	all eligible Ps from 1 urban district	N/A	3 Vignettes	Written	Ill-structured vignettes can generate variance. Self-report expertise was related to experience, but not judged expertise. Broader, generic responses were judged worse than specific responses. Coded application of ISLLC concepts not application of Leithwood's model.	Construct
Sleegers, Wassink, van Veen, & Imants	2009	Qual	2	early career Ps	N/A	From Practice	Interview	Personal and professional biographies shape how Ps frame/interpret problems.	Personal

Note. Ps = Principals. DM = Decision-making. Obs. = Observation. * = 24 Ps and 14 Asst. Ps. ISLLC = Interstate School Leaders Licensure Consortium. Reputation identification methods involved using 1 or 2 central office admins to nominate expert or effective Ps.

APPENDIX B

SUMMARY OF ASSESSMENT FRAMEWORKS

Framework	Discipline	Construct Definition	Cognitive Processes	Operationalized Behaviors	Task Definition	Item Format	Variable Types	Variables Used
Sugrue, 1995	K-12 Education	Knowledge; Skills; Meta-cognition; Motivation	Planning & Monitoring	Did not provide this level of specificity	3 types of tasks: Selection, Generation, and/or Explanation of responses	Recommended Multiple Formats	Product	Correct/Incorrect; Motivation
O'Neill & Schacter, 1997 (CRESST)	K-12 Education	Knowledge; Skills; Meta-cognition; Motivation	Representing; Planning; Executing; Monitoring	Concept Mapping; Internet Searching	Concept Map; Simulated Webspace;	Constructed response; Survey Items	Product	Agreement w/Expert Maps; Links Searched
OECD, 2010 (PISA)	K-12 Education	Knowledge; Skills; Meta-cognition; Motivation	Recognize & specify problems; plan & execute solutions; monitor & evaluate progress	(Non-)numerical answers; Exploration strategies; Extended explanations	Ill-structured; Non-Routine; Interactive vs. Static; Computerized problems	Mult-choice; Constructed response; Drag & drop	Product	Correct/Incorrect; Partial Credit
D'Zurilla & Maydieu-Olivares, 1995	Social Problem-Solving	Skills	Problem identification; Generation of solutions; Perspective-taking; Mental simulation	Written, ranking, & MC responses that exhibit the cognitive processes	Vignettes; Interviews	Recommended Multiple Formats	Process; Product	Correct/Incorrect; Effectiveness; Recommended Multiple Variables
Zaccaro et al., 2000	Military Leadership	Knowledge; Skills; Meta-cognition	Problem identification; Information encoding; Category search, specification, re-organization; Evaluate, implement, monitor solutions	Written responses that exhibit the cognitive processes	Ill-structured Vignettes	Performance; Cued & Uncued Items	Process; Product	Effectiveness; Feasibility; Creativity; Ability to Coordinate;

Note. Rows in grey shadow indicate the framework is generic and not reflective of a framework for a specific, published assessment. OECD = The Organisation for Economic Co-operation and Development. MC = multiple choice.

APPENDIX C

CONTENT VALIDITY VIGNETTES

Defiance

1. You walk by a classroom and hear a teacher ask a student read a passage aloud in class. The student says, “No thanks Teach,” and looks down at the passage. The teacher, feeling disrespected, pressed the student to comply, who continually refused. The teacher, exasperated, sends the student to the office for being defiant. How do you handle the situation?
2. A student is continually being written up for defiance across several teachers. The student does not have a history of office discipline referrals and has average grades. Once teachers realized that the student is being defiant in nearly all of their classrooms, they decide to come to you seeking counsel, help, anything because the student has stopped following directions and talks back whenever asked to comply. How do you handle the situation?
3. A student is continually being written up for defiance with one specific teacher. The student has no history of office discipline referrals, and other teachers have not shared any negative anecdotes. How do you handle the situation?
4. You are walking by a classroom when you see and overhear a student being exceptionally disrespectful to a veteran teacher, but the teacher does not address it at all. In fact, you think the teacher looks scared and unsure of what to do. How do you handle the situation?

Harassment

5. Concerned parents complain to you that their child is being harassed over the internet by students in their child's grade. The parents are not combative; they are pleading for your help. Their child is being sent intimidating messages that cause their child to want to stay home from school. Based on the time stamps of the messages, the bullying is not taking place on school grounds, nor with school equipment. Moreover, it's unclear who sent the messages because the screen names used online cannot be linked easily with student names on class rosters. How do you handle the situation?
6. Three students are sent to the office for fighting. Two of them say the third student started the fight and attacked them. The third student was found fighting against the other two and had to be pulled off of one of them to break up the fight. The third student explains to you, in confidence, that the other two were "taunting me with gay slurs." The student admitted to snapping and attacking them because the student is gay but hasn't disclosed that to anybody at the school yet. How do you handle the situation?
7. Two students are sent to the office for harassment. The first student, a White student, was heard repeatedly yelling "Go back where you came from!" to a group of Hispanic students. The Hispanic student, sent to the office with the first student, was stopped before he physically attacked the first student but was heard threatening his physical well-being. How do you handle the situation?

Threat

8. A teacher felt threatened by a student and promptly issued an office discipline referral that sent the student to the office. You consult with both of them separately and they both described similar, but slightly different events. The teacher reprimanded the student for talking during the lesson and the student smiled at the teacher. From the teacher's point of view, the student glared and smiled menacingly. The student recalls staring back at the teacher, smiling awkwardly because of not being sure what to say or do. How do you handle the situation?
9. You receive a phone call from the local police department. The police officer tells you they have been notified that a person has been seen on your school campus visibly carrying a shotgun. The person appeared to be the age of a student, but it is not clear, and the person was most recently seen on your school's running track, about half of a mile from the main school building. The police are on their way but will not be there for at least 15 minutes. The day is nearing its end and students are scheduled to leave the building within 30 minutes. How do you handle the situation?

Abusive/Inappropriate Language

10. From an inclusive classroom setting, a student with an IEP was sent to the office for reportedly standing up in the middle of class and calling one of his classmates "a full retard" after dropping his pencil. In the office with you, the student explains that he saw this same classmate say that exact thing to another student

who receives special education services earlier in the week and he was just trying to fit in. How do you handle the situation?

11. You are walking through the cafeteria and overhear a student delivering abusive language to another student. You notice that the teacher does not respond, to the dismay of the paraprofessional who then promptly sends the student to office for using inappropriate language. In your office, the student is not sure what was inappropriate since that language is used at home. How do you handle the situation?

Fighting/Physical Aggression

12. Two students are fighting and a teacher steps in to break up the fight. As the teacher grabs one student to separate them, the teacher accidentally elbows the other student in the face, causing a bloody nose and possibly breaking it. The fight ends after everyone sees the blood and the student screams in pain. The next day, parents come in to complain that the teacher broke their child's nose; they are considering legal action. How do you handle the situation?
13. A larger student suddenly pushes back his desk and stands up in frustration. An older paraprofessional happens to be walking behind the student at the time. The student knocks over the paraprofessional, causing an injury to the paraprofessional's wrist upon falling. Compounding the incident is the fact that these two had a negative interaction recently. The paraprofessional is convinced the student did it intentionally, but student says he had no idea the paraprofessional was behind him because he was arguing with the student in front

of him who confirms the student's story after you checked. How do you handle the situation?

APPENDIX D

PILOT TEST VIGNETTES

1. A student is being repeatedly written up for a quiet, non-compliant form of defiance with one specific teacher. The student has no history of office discipline referrals, has had mixed grades in this subject matter previously, and other teachers have not shared any recent negative anecdotes. How do you handle the situation?

2. Concerned parents complain to you that their child is being harassed over the internet by students from their child's school. They are considering filing a police report because their child is being sent intimidating messages that cause their child to want to stay home from school. The parents show you screenshots of a messaging app that is commonly used by middle school students. Based on the time stamps of the messages, the bullying is not taking place on school grounds, nor with school equipment. Moreover, it is not definitive who sent the messages because the screen names used online cannot be linked easily with student names on class rosters. The student has been verbally bullied at school by a few boys and girls and assumes it is these same students. How do you handle the situation?

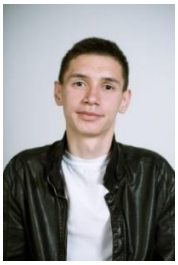
3. A teacher felt threatened by a student and issued an office discipline referral. You consult with both of them separately and they both described similar, but slightly different events. The teacher reprimanded the student for talking during the lesson and the student smiled at the teacher, muttering something under his breath. From the teacher's point of view, the student glared back at the teacher and smiled menacingly, muttering "you'll be sorry." The student recalls staring back at the teacher and smiling awkwardly because of not being sure what to say or do and says he did not mutter anything. How do you handle the situation?

4. You receive a phone call from the local police department. The police officer tells you they have been notified that a person has been seen visibly carrying a shotgun on your school campus's running track, about a quarter of a mile from the main school building. The person appeared to be the age of a student, but it is not clear. The police are on their way but will not be there for at least 5 minutes. The police officer tells you to go into lockdown, but the students were just released for the day 5 minutes before you took this phone call. How do you handle the situation?

5. Two students are fighting and a teacher steps in to break up the fight. As the teacher puts his hands on one student to separate them, the teacher accidentally elbows the other student in the face, causing a bloody nose. The fight ends after everyone sees the blood and the student screams in pain. The bloodied student receives medical attention; both students' parents are called to pick them up; the teacher completes an accident report. The next day, the parents come back to complain that the teacher gave their child a bloody nose; they are considering legal action. How do you handle the situation?



6. Two students were issued office discipline referrals for harassment. They were escorted separately to the office and have not been allowed to go back to class yet. The first student, a White student, was heard yelling repeatedly "Go back where you came from!" to a Latino student. The Latino student was stopped by a teacher before he physically attacked the White student, but the teacher heard the Latino student say that he was "going to get" the White student. How do you handle the situation?

7. Two students have been separated and escorted to your office for fighting. The first student was found attacking the other student. He explains to you, in confidence, that the other kid was “calling me gay and harassing me.” The first student admitted to attacking the other student first because he thinks he might be gay but has not disclosed that to anybody at the school yet. How do you handle the situation?

8. A larger student suddenly pushes back his desk chair and stands up in frustration. A paraprofessional happens to be walking behind the student at the time. The student knocks over the paraprofessional, causing an injury to the paraprofessional’s wrist upon falling. Compounding the incident is the fact that these two had a negative interaction recently. The paraprofessional is convinced the student did it intentionally, but student says he had no idea the paraprofessional was behind him because he was arguing with the student in front of him. You follow up with the student in front of him, who confirms the student’s story. How do you handle the situation?

APPENDIX E

FIELD TEST VIGNETTES

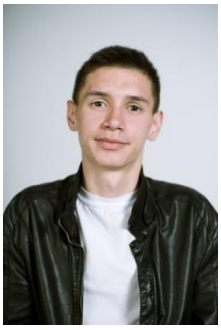
1. Concerned parents complain to you that their child, Kim, is being harassed over the internet by students from her school. They are considering filing a police report because their child is being sent intimidating messages that cause Kim to want to stay home from school. The parents show you screenshots of a messaging app that is commonly used by middle school students. Based on the time stamps of the messages, the bullying is not taking place on school grounds, nor with school equipment. Moreover, it is not definitive who sent the messages because the screen names used online cannot be linked easily with student names on class rosters. Kim has been verbally bullied at school by a few classmates, and the family assumes it is these same students.

Kim

2. You receive a phone call from the local police department. The police officer tells you that they have been notified that a person has been seen visibly carrying a shotgun on your school campus's running track, about a quarter of a mile from the main school building. The person appeared to be the age of a student, but it is not clear. The police are on their way but will not be there for about 5 minutes. The police officer tells you to go into lockdown, but the students were just released for the day about 5 minutes before you took this phone call.

3. Two students were issued office discipline referrals for harassment. They were escorted separately to the office and have not been allowed to go back to class yet. The first student, a White student, John, was heard yelling repeatedly “Go back where you came from!” to a Latino student, named Jose. Before physically attacking John, Jose was stopped by a teacher because the teacher heard Jose say that he was “going to get” John.

John



Jose

4. Two students have been separated and escorted to your office for fighting. The first student, Bruce, was found attacking another student, Tom. Bruce tells you that Tom was “calling me gay and harassing me.” Bruce then admitted to throwing the first punch because, he explains to you in confidence, he thinks he might be gay but isn’t sure and has not disclosed that to anybody at the school yet.

Bruce

Tom

5. A larger student, Will, suddenly pushes back his desk chair and stands up in frustration. A paraprofessional, Mrs. Bell, is walking behind Will at the time. Will knocks over Mrs. Bell, causing an injury to her wrist upon falling. Compounding the incident is the fact that these two had a negative interaction recently. Mrs. Bell is convinced Will did it intentionally, but Will says he had no idea Mrs. Bell was behind him because he was arguing with the student in front of him. You follow up with the student in front of him, who confirms Will's story.

Will

Mrs. Bell

APPENDIX F

RESULTS OF PICTURES SURVEY

Picture	Target AGE	Match Target Age?	Target Race-Eth	Match Target Race-Eth?	Target Gender	Match Target Gender?	Like mean	Phys mean	Attract Index
Stud01	MS student	2/12	white	12/12	male	12/12	60.25	52.33	112.58
Stud02	MS student	11/12	white	6/12	male	8/12	58.92	57.75	116.67
Stud03	MS student	6/12	hisp	8/12	male	12/12	61.58	53.50	115.08
Stud04	MS student	6/12	white	7/12	male	11/12	60.58	62.17	122.75
Stud05	MS student	3/12	black	12/12	female	12/12	62.17	65.58	127.75
Stud06	MS student	2/12	white	8/12	male	12/12	50.42	53.83	104.25
Stud07	MS student	3/12	white	12/12	male	12/12	61.17	59.58	120.75
Stud08	MS student	6/12	asian	11/12	male	7/12	62.83	61.00	123.83
Stud09	MS student	12/12	black	9/12	female	12/12	65.58	65.25	130.83
Stud10	MS student	4/12	black	12/12	male	12/12	68.75	70.92	139.67
Stud11	MS student	9/12	white	11/12	male	12/12	55.83	59.75	115.58
Stud12	MS student	4/12	hisp	1/12	female	12/12	68.92	70.25	139.17
Stud13	MS student	10/12	black	11/12	male	8/12	57.17	56.58	113.75
Stud14	MS student	5/12	white	8/12	female	12/12	55.83	67.08	122.92
Stud15	MS student	11/12	black	11/12	male	9/12	58.50	63.08	121.58
Stud16	MS student	0/11	white	6/11	male	3/11	26.64	28.91	55.55
Stud17	MS student	8/11	white	11/11	female	11/11	62.09	60.09	122.18
Stud18	MS student	5/11	hisp	8/11	male	10/11	52.18	52.64	104.82
Stud19	MS student	5/11	hisp	1/11	female	10/11	56.09	55.73	111.82
Stud20	MS student	10/11	white	10/11	male	7/11	52.18	46.45	98.64
Stud21	MS student	2/11	white	11/11	male	10/11	54.45	62.00	116.45
Stud22	MS student	0/11	black	11/11	male	11/11	57.73	60.64	118.36
Stud23	MS student	5/11	hisp	0/11	male	7/11	48.45	49.36	97.82
Stud24	MS student	9/11	white	10/11	female	11/11	55.82	61.82	117.64
Staff1	n/a	n/a	black	11/11	female	11/11	68.09	68.55	136.64
Staff2	n/a	n/a	white	5/11	male	11/11	64.09	63.64	127.73
Staff3	n/a	n/a	white	10/11	female	11/11	53.36	49.18	102.55
Staff4	n/a	n/a	hisp	5/11	male	11/11	69.09	62.36	131.45
Staff5	n/a	n/a	white	11/11	female	11/11	58.91	58.64	117.55
Staff6	n/a	n/a	black	10/11	male	11/11	61.00	60.64	121.64
Staff7	n/a	n/a	Black	11/11	female	11/11	64.73	55.73	120.45
Staff8	n/a	n/a	white	3/11	male	11/11	47.91	49.55	97.45

Note. Like = likability; phys = physical attractiveness; attract = attractiveness; Stud = student; MS = Middle School; hisp = Hispanic or Latinx.

APPENDIX G

RELIABLY RATED PICTURES

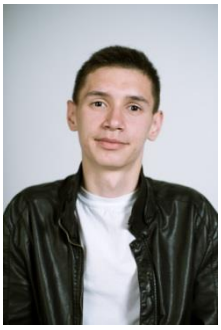
White Male Student 1

White Male Student 2

White Male Student 3

White Female Student

Latino Male Student



Asian Male Student

Black Male Student 1

Black Male Student 2

Large Male Student 1

Large Male Student 2

Shooter

Computer Student 1

Computer Student 2

Computer Student 3

(Staff Pictures on Next Page)

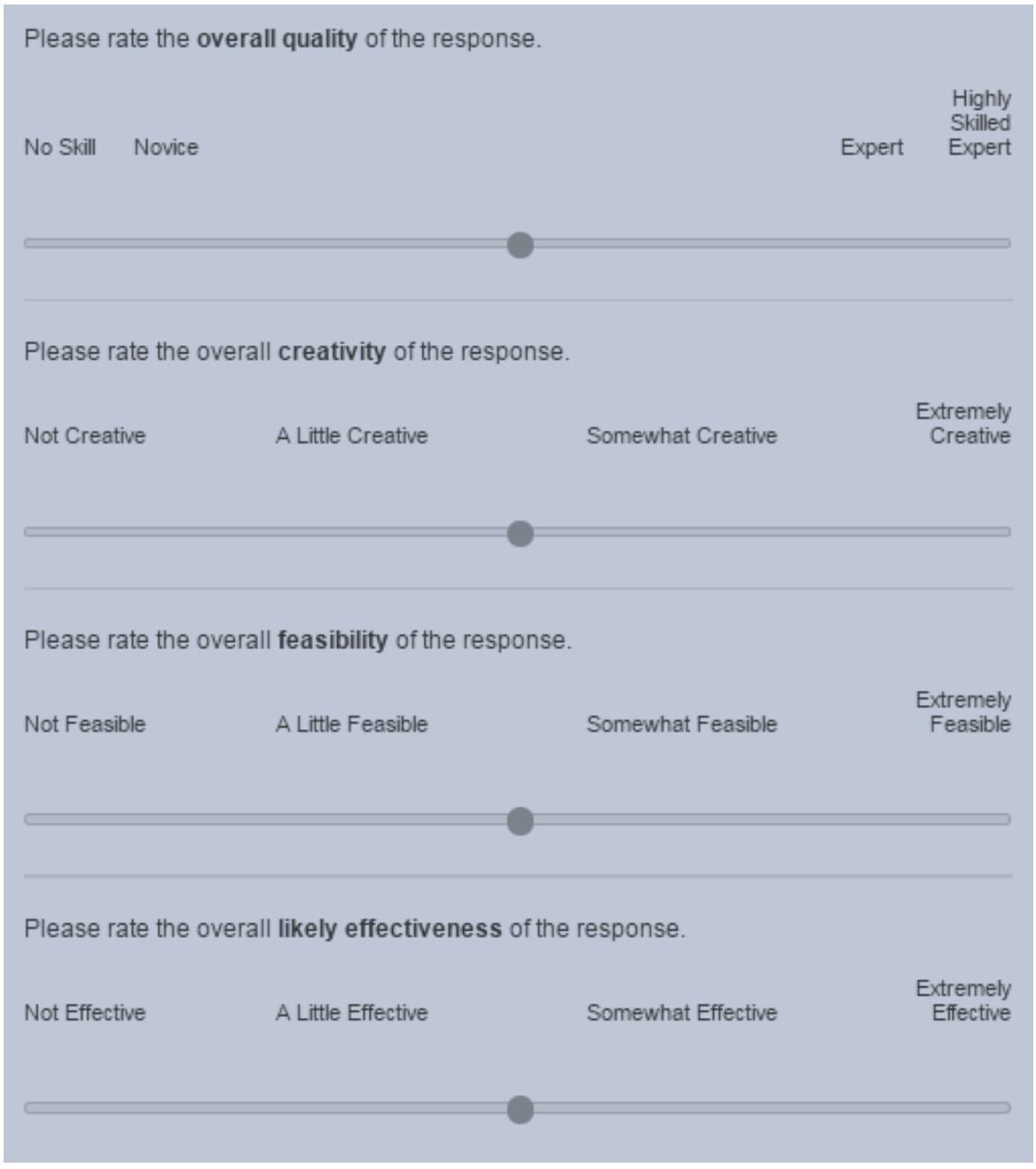
Teacher 1	Teacher 2	Teacher 3
		

Paraprofessional 1	Paraprofessional 2
	

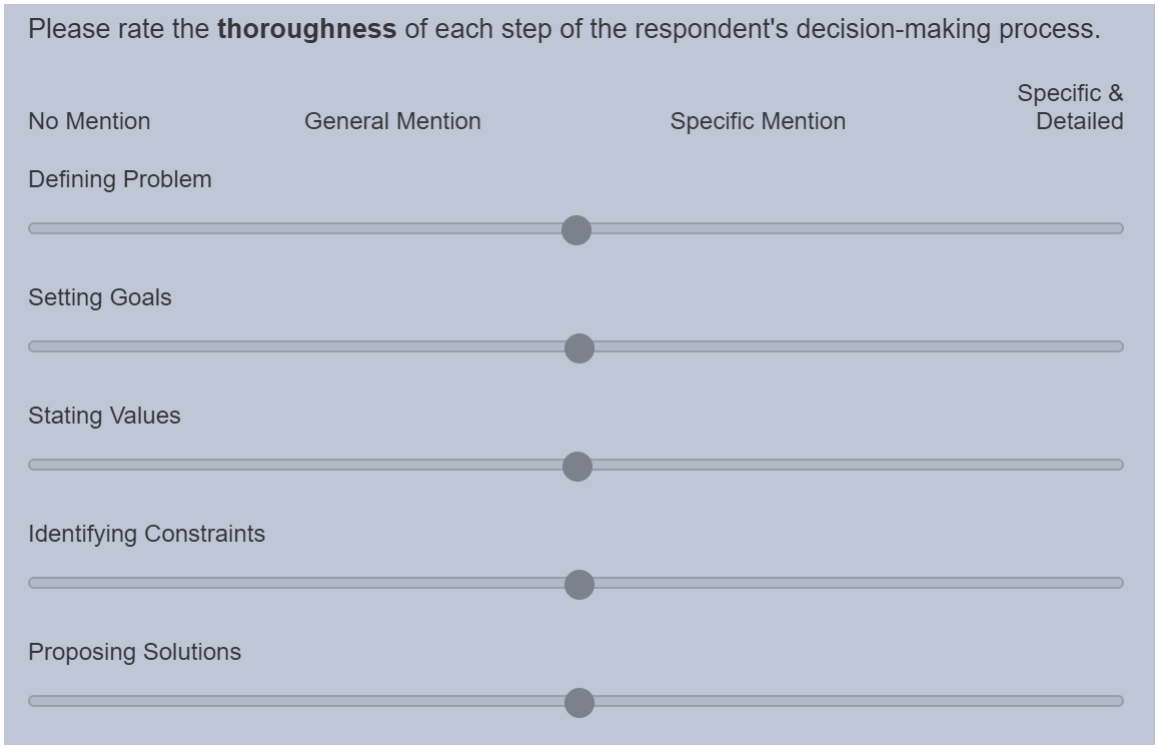
APPENDIX H

SCREENSHOTS OF JUDGES SLIDER SCALES

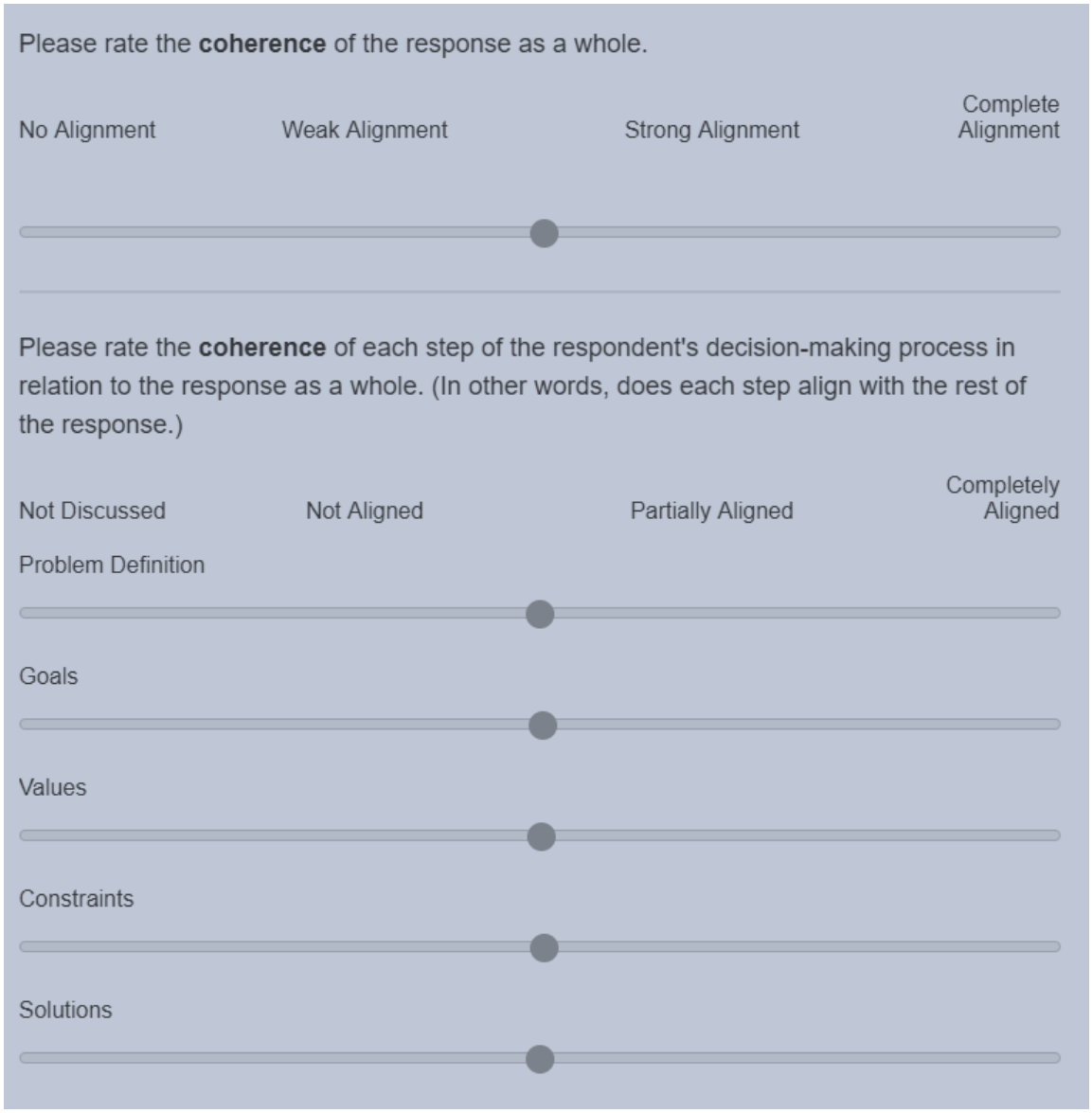
The Product Variables: Quality, Creativity, Feasibility, and Effectiveness for Pilot & Field Tests



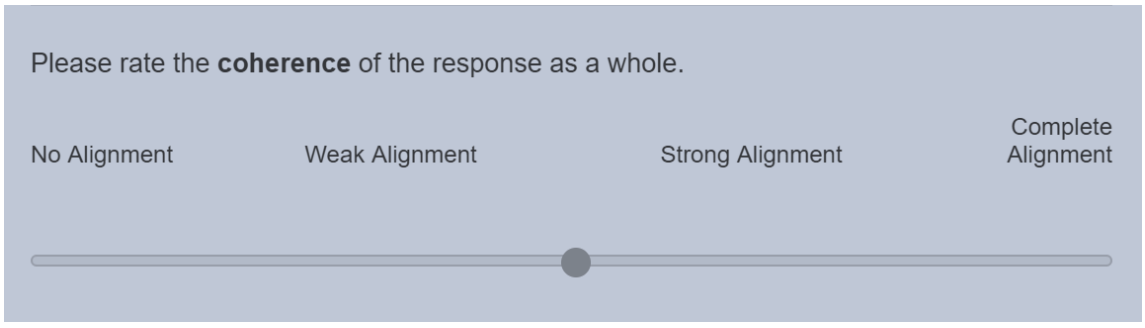
Thoroughness for Pilot & Field Tests:



Coherence for the pilot:



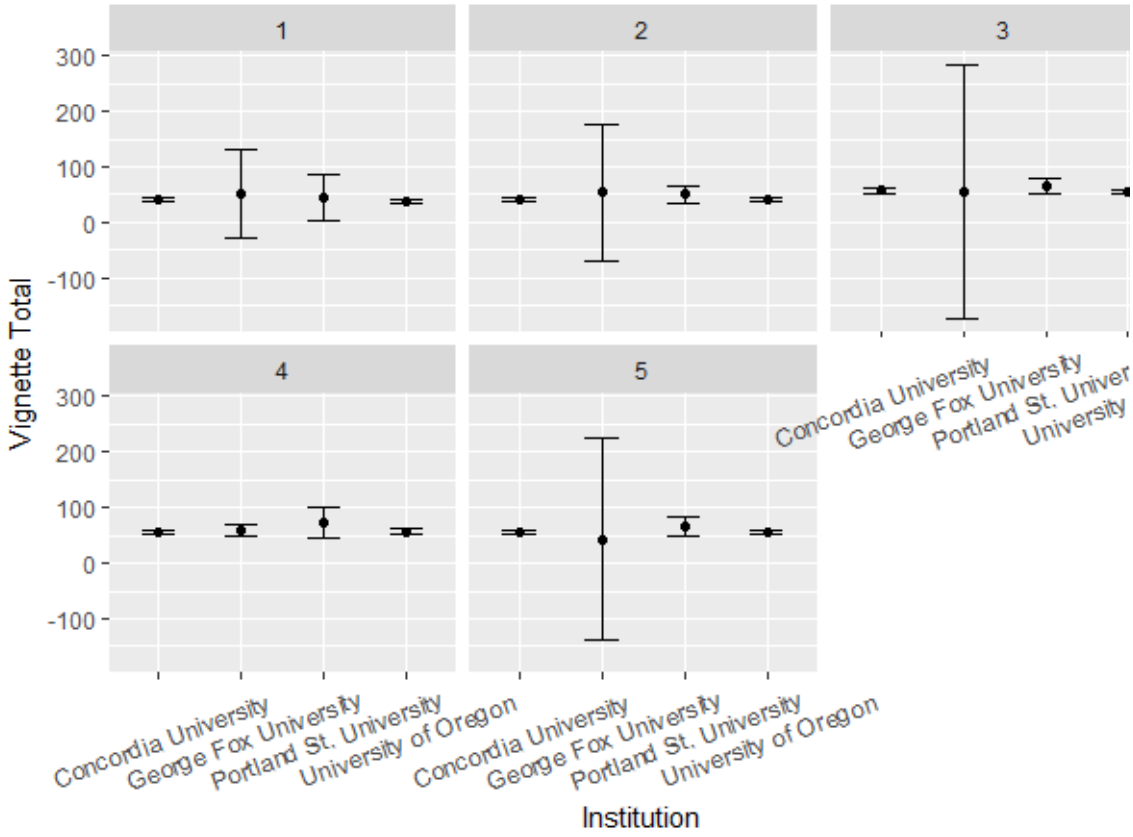
Coherence for the field test:



APPENDIX I

DIFFERENCES BETWEEN INSTITUTIONS IN FIELD TEST

Institutions by Vignette Total



APPENDIX J

ICC(C,1) FOR VARIABLES AND VIGNETTE TOTALS

Variables

	Individual Ratings Across		Average Scores Across		
	All 5 field		Unprompted	Prompted	All 8
	Mean ICC(c,1)	Range	Mean ICC(c,1)	Mean ICC(c,1)	Mean ICC(c,1)
Qual	0.41	.29 - .52	0.42	0.68	0.63
Crea	0.37	.26 - .49	0.42	0.55	0.62
Feas	0.25	.03 - .37	0.13	0.40	0.34
Effct	0.39	.26 - .53	0.29	0.61	0.61
cohOverall	0.25	.04 - .37	0.38	0.44	0.48
cohProb	0.39	.26 - .55	0.34	0.64	0.58
cohGoal	0.30	.13 - .40	0.41	0.51	0.63
cohVal	0.39	.22 - .55	0.40	0.56	0.54
cohCon	0.46	.34 - .65	0.50	0.69	0.65
cohSol	0.23	.14 - .47	0.21	0.36	0.43
thorProb	0.44	.31 - .58	0.30	0.51	0.60
thorGoal	0.34	.26 - .44	0.45	0.63	0.65
thorVal	0.47	.40 - .52	0.50	0.66	0.63
thorCon	0.46	.39 - .53	0.47	0.77	0.72
thorSol	0.32	.23 - .54	0.35	0.45	0.56
thorTot	0.71	.55 - .80	0.56	0.69	0.76
cohTot	0.67	.46 - .81	0.55	0.61	0.71
vignTot	0.71	.52 - .81	0.55	0.70	0.75

Vignettes

	Individual Ratings Across		Total Scores	
	Mean ICC(c,1)	Range	Mean ICC(c,1)	Range
Defiance	0.32	.26 - .43	0.67	.61 - .73
Online	0.3	.07 - .45	0.6	.08 - .87
Threat	0.22	.05 - .45	0.41	.12 - .90
Shooter	0.38	.09 - .59	0.69	.25 - .86
Elbow	0.45	.40 - .50	0.79	.72 - .86
GoBack	0.46	.31 - .62	0.8	.68 - .91
Homophobia	0.29	.00 - .47	0.56	-.05 - .85
Para	0.46	.35 - .60	0.79	.68 - .87

APPENDIX K

SINGLE UNIT ICCS FOR THE FIELD TEST

Variables

	Individual Ratings Across		Average Scores Across		
	All 5 field		Unprompted	Prompted	All 5
	Mean ICC(c,1)	Range	Mean ICC(c,1)	Mean ICC(c,1)	Mean ICC(c,1)
Qual	0.41	.29 - .49	0.55	0.53	0.6
Crea	0.33	.17 - .52	0.4	0.53	0.55
Feas	0.15	.00 - .22	0.19	0.25	0.3
Effct	0.36	.22 - .43	0.51	0.49	0.54
Coh	0.32	.12 - .45	0.48	0.33	0.44
thorProb	0.55	.30 - .74	0.63	0.59	0.58
thorGoal	0.51	.17 - .72	0.3	0.53	0.54
thorVal	0.51	.22 - .69	0.3	0.51	0.52
thorCon	0.53	.24 - .77	0.45	0.57	0.62
thorSol	0.45	.34 - .60	0.55	0.58	0.66
thorTot	0.62	.35 - .79	0.63	0.64	0.7
vignTot	0.57	.32 - .73	0.66	0.62	0.68

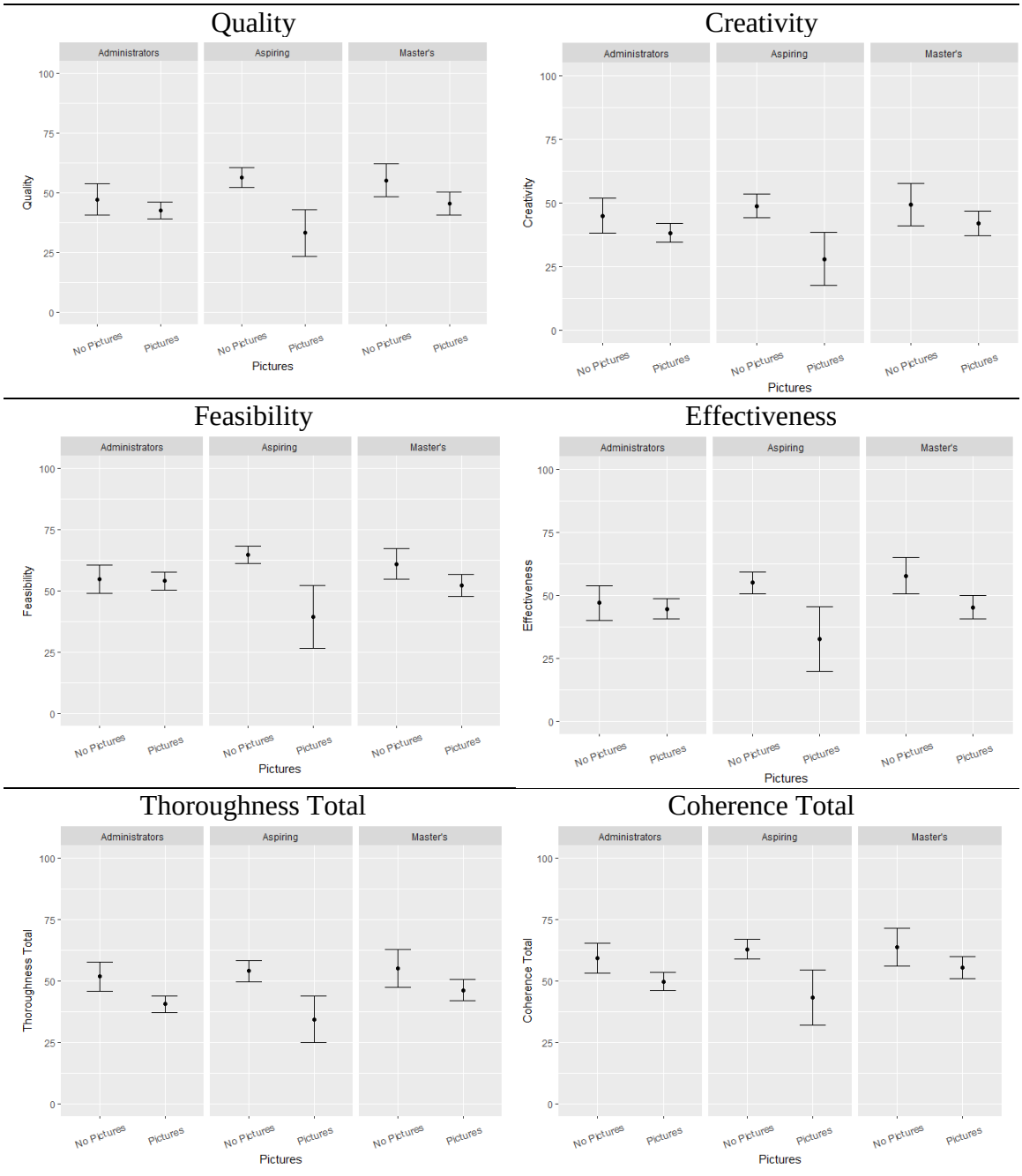
Vignettes

	Individual Ratings Across		Total Scores	
	Mean ICC(c,1)	Range	Mean ICC(c,1)	Range
Online	0.41	.32 - .57	0.62	.42 - .78
Shooter	0.43	.27 - .60	0.52	.40 - .66
GoBack	0.34	.20 - .49	0.41	.22 - .67
Homophobia	0.44	.24 - .65	0.58	.30 - .85
Para	0.41	.11 - .55	0.48	-.01 - .67

APPENDIX L

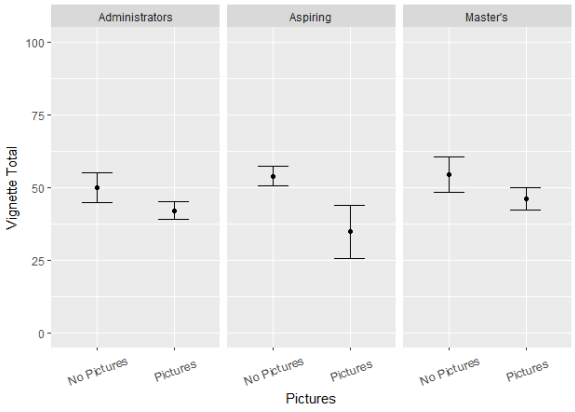
THE EFFECT OF THE PICTURES

Graphical Representation of the Effect of the Pictures in the Pilot Test



(cont'd below)

Vignette Total



APPENDIX M

PILOT TEST INTER-ITEM CORRELATION MATRIX

	qual	crea	feas	effct	Coh Ovr	Coh Prob	Coh Goal	Coh Val	Coh Con	Coh Sol	Thor Prob	Thor Goal	Thor Val	Thor Con	Thor Sol	Coh Tot	Thor Tot	Vign Tot
qual	1																	
crea	.79	1																
feas	.65	.53	1															
effct	.86	.75	.68	1														
cohOvr	.74	.65	.60	.71	1													
cohProb	.54	.44	.38	.46	.56	1												
cohGoal	.53	.45	.43	.53	.66	.58	1											
cohVal	.52	.44	.33	.45	.59	.65	.68	1										
cohCon	.52	.46	.34	.46	.55	.60	.61	.61	1									
cohSol	.54	.44	.50	.58	.67	.44	.62	.48	.48	1								
thorProb	.58	.49	.36	.50	.54	.81	.52	.60	.59	.38	1							
thorGoal	.56	.48	.40	.53	.61	.53	.75	.62	.58	.50	.62	1						
thorVal	.56	.48	.36	.50	.58	.62	.60	.83	.62	.44	.68	.68	1					
thorCon	.52	.49	.31	.46	.52	.54	.53	.57	.81	.41	.61	.58	.65	1				
thorSol	.53	.46	.44	.55	.58	.36	.55	.45	.43	.72	.41	.57	.46	.45	1			
cohTot	.65	.55	.48	.60	.74	.82	.85	.85	.82	.72	.73	.73	.77	.71	.60	1		
thorTot	.68	.59	.46	.63	.70	.71	.73	.76	.75	.60	.82	.85	.86	.82	.70	.88	1	
vignTot	.81	.72	.62	.77	.80	.76	.79	.79	.78	.70	.78	.80	.81	.76	.69	.94	.95	1

APPENDIX N

PILOT TEST ANOVA AND REGRESSION TABLES

Table 1. ANOVA Summary Table for effect of Program Enrollment on Vignette Totals

Factor	*df*	Mean Squares	*F*	*P*
Defiance				
Type	2	122.05	0.43	.650
Pictures	1	1237.54	4.41	.041
Residuals	45	280.74		
Cyber				
Type	2	203.15	1.08	.349
Pictures	1	2024.99	10.73	.002
Residuals	45	188.66		
Threat				
Type	2	61.64	0.34	.713
Pictures	1	201.74	1.12	.297
Residuals	45	180.90		
Shooter				
Type	2	553.05	2.50	.093
Pictures	1	2015.03	9.12	.004
Residuals	44	220.96		
Elbow				
Type	2	4.11	0.02	.984
Pictures	1	767.22	2.94	.093
Residuals	44	260.67		
GoBack				
Type	2	106.62	0.35	.704
Pictures	1	426.79	1.42	.240
Residuals	44	301.34		
Homophobia				
Type	2	427.51	2.21	.122
Pictures	1	485.87	2.52	.120
Residuals	43	193.17		
Para				
Type	2	860.92	2.73	.077
Pictures	1	650.33	2.06	.159
Residuals	41	315.43		

Table 2. Robust regression summary table for effect of school-based administrator status on vignette totals

Model	*Est*	SE	*T-value*	*P*	R^2
Defiance					0.06
Intercept	47.15	3.93	11.99	0.00	
SBAdmin	-0.29	5.78	-0.05	0.96	
Pictures	-8.71	5.71	-1.52	0.13	
Cyber					0.23
Intercept	51.19	3.61	14.17	0.00	
SBAdmin	-1.90	3.82	-0.5	0.62	
Pictures	-14.22	3.99	-3.56	0.00	
Threat					0.03
Intercept	37.63	2.6	14.48	0.00	
SBAdmin	0.65	3.75	0.17	0.86	
Pictures	-4.08	3.8	-1.07	0.29	
Shooter					0.21
Intercept	50.59	3.46	14.63	0.00	
SBAdmin	-1.80	5.23	-0.34	0.73	
Pictures	-15.40	5.2	-2.96	0.00	
Elbow					0.05
Intercept	63.39	3.7	17.12	0.00	
SBAdmin	-1.17	4.89	-0.24	0.81	
Pictures	-7.23	4.56	-1.59	0.12	
GoBack					0.05
Intercept	61.32	3.55	17.28	0.00	
SBAdmin	-5.30	5.86	-0.90	0.37	
Pictures	-5.10	5.21	-0.98	0.33	
Homophobia					0.13
Intercept	56.67	2.90	19.56	0.00	
SBAdmin	-5.97	4.80	-1.24	0.22	
Pictures	-8.17	4.14	-1.97	0.05	
Para					0.18
Intercept	61.72	3.67	16.80	0.00	
SBAdmin	-10.55	6.95	-1.52	0.14	
Pictures	-10.15	5.69	-1.79	0.08	

Note. The robust estimator does not provide the F-statistic for the regression; it provides a t-value instead. SBAdmin = School-based administrator status.

Table 3. Robust regression summary table for effect of self-rated expertise on vignette totals

Model	*Est*	SE	*T-value*	*P*	R^2
Defiance					0.06
Intercept	51.19	9.78	5.24	0.00	
Self-rate Exp	-0.06	0.13	-0.45	0.65	
Pictures	-9.56	5.60	-1.71	0.09	
Cyber					0.23
Intercept	56.47	6.69	8.44	0.00	
Self-rate Exp	-0.08	0.08	-0.99	0.33	
Pictures	-15.03	4.10	-3.66	0.00	
Threat					0.05
Intercept	43.41	6.54	6.64	0.00	
Self-rate Exp	-0.05	0.08	-0.66	0.51	
Pictures	-5.93	4.01	-1.48	0.15	
Shooter					0.30
Intercept	64.11	7.00	9.16	0.00	
Self-rate Exp	-0.21	0.09	-2.31	0.03	
Pictures	-18.37	4.32	-4.25	0.00	
Elbow					0.06
Intercept	66.44	7.95	8.36	0.00	
Self-rate Exp	-0.06	0.10	-0.6	0.55	
Pictures	-7.95	4.90	-1.62	0.11	
GoBack					0.05
Intercept	65.55	8.49	7.72	0.00	
Self-rate Exp	-0.08	0.11	-0.71	0.48	
Pictures	-8.05	5.23	-1.54	0.13	
Homo-phobia					0.12
Intercept	60.76	6.88	8.83	0.00	
Self-rate Exp	-0.08	0.09	-0.87	0.39	
Pictures	-10.61	4.28	-2.48	0.02	
Para					0.13
Intercept	69.03	9.11	7.57	0.00	
Self-rate Exp	-0.15	0.12	-1.33	0.19	
Pictures	-13.24	5.60	-2.36	0.02	

Note. The robust estimator does not provide the F-statistic for the regression; it provides a t-value instead. Self-rate Exp = Self-rated expertise.

Table 4. Robust regression summary table for effect of years professionally in schools on vignette totals

Model	*Est*	SE	*T-value*	*P*	R^2
Defiance					0.06
Intercept	48.88	5.02	9.73	0.00	
Years Exp	-0.15	0.37	-0.41	0.68	
Pictures	-8.71	5.49	-1.59	0.12	
Cyber					0.22
Intercept	53.05	4.43	11.98	0.00	
Years Exp	-0.20	0.25	-0.77	0.44	
Pictures	-14.02	3.96	-3.54	0.00	
Threat					0.05
Intercept	36.91	4.29	8.59	0.00	
Years Exp	0.21	0.24	0.85	0.40	
Pictures	-5.07	3.84	-1.32	0.19	
Shooter					0.25
Intercept	54.40	4.76	11.43	0.00	
Years Exp	-0.37	0.27	-1.38	0.18	
Pictures	-15.73	4.29	-3.67	0.00	
Elbow					0.05
Intercept	61.00	5.23	11.66	0.00	
Years Exp	0.09	0.30	0.30	0.76	
Pictures	-7.04	4.71	-1.49	0.14	
GoBack					0.04
Intercept	61.04	5.60	10.91	0.00	
Years Exp	-0.07	0.32	-0.21	0.83	
Pictures	-7.03	5.04	-1.39	0.17	
Homophobia					0.11
Intercept	57.38	4.65	12.34	0.00	
Years Exp	-0.16	0.27	-0.57	0.57	
Pictures	-9.71	4.12	-2.36	0.02	
Para					0.18
Intercept	67.17	5.64	11.91	0.00	
Years Exp	-0.71	0.32	-2.20	0.03	
Pictures	-11.55	5.18	-2.23	0.03	

Note. The robust estimator does not provide the F-statistic for the regression; it provides a t-value instead. Years Exp = Years of professional experience.

Table 5. ANOVA Summary Table for effect of Program Enrollment on Variables

	df	Mean Squares	*F*	*P*
Qual				
Type	2	2736.56	6.61	0.002
Pics	1	6366.16	15.38	0.000
Residuals	379	413.99		
Crea				
Type	2	1175.70	2.56	0.078
Pics	1	6419.04	14.00	0.000
Residuals	379	458.40		
Feas				
Type	2	1625.31	4.16	0.016
Pics	1	4119.10	10.53	0.001
Residuals	379	391.05		
Effct				
Type	2	1342.20	2.79	0.063
Pics	1	5515.67	11.45	0.001
Residuals	379	481.66		
cohOvr				
Type	2	1691.10	5.79	0.003
Pics	1	4256.27	14.58	0.000
Residuals	379	291.92		
cohProb				
Type	2	3294.57	4.65	0.010
Pics	1	11955.41	16.86	0.000
Residuals	379	709.00		
cohGoal				
Type	2	1784.18	3.84	0.022
Pics	1	6172.41	13.29	0.000
Residuals	379	464.41		
cohVal				
Type	2	4157.44	6.34	0.002
Pics	1	14421.52	22.01	0.000
Residuals	379	655.33		
cohCon				
Type	2	3334.80	4.44	0.012
Pics	1	10006.39	13.32	0.000
Residuals	379	751.29		
cohSol				
Type	2	247.05	0.71	0.491
Pics	1	2647.95	7.63	0.006
Residuals	379	346.85		
thorProb				
Type	2	2330.68	3.91	0.021
Pics	1	9212.18	15.47	0.000
Residuals	379	595.47		
thorGoal				
Type	2	2881.83	6.07	0.003
Pics	1	11149.29	23.47	0.000
Residuals	379	475.13		

thorVal				
Type	2	3349.03	5.68	0.004
Pics	1	15941.15	27.04	0.000
Residuals	379	589.59		
thorCon				
Type	2	2738.80	4.49	0.012
Pics	1	7734.99	12.67	0.000
Residuals	379	610.62		
thorSol				
Type	2	423.09	1.00	0.367
Pics	1	7715.23	18.31	0.000
Residuals	379	421.43		
thorTot				
Type	2	2077.94	5.62	0.004
Pics	1	10145.37	27.46	0.000
Residuals	379	369.47		
cohTot				
Type	2	2161.69	5.37	0.005
Pics	1	8444.35	20.99	0.000
Residuals	379	402.35		
vignTot				
Type	2	1685.67	5.91	0.003
Pics	1	7072.03	24.81	0.000
Residuals	379	285.04		

Table 6. Robust regression summary table for effect of school-based administrator status on the variables

	Est	SE	*T-value*	*P*	R^2
Qual					0.065
Intercept	54.43	1.76	30.96	0.00	
SBAdmin	-2.97	2.45	-1.21	0.23	
Pictures	-10.27	2.37	-4.33	0.00	
Crea					0.045
Intercept	48.68	1.89	25.69	0.00	
SBAdmin	-2.14	2.57	-0.83	0.40	
Pictures	-9.14	2.52	-3.62	0.00	
Feas					0.044
Intercept	61.58	1.49	41.38	0.00	
SBAdmin	1.87	2.50	0.75	0.46	
Pictures	-9.06	2.25	-4.03	0.00	
Effct					0.039
Intercept	53.95	1.83	29.56	0.00	
SBAdmin	-2.13	2.76	-0.77	0.44	
Pictures	-8.60	2.56	-3.36	0.00	
cohOvr					0.059
Intercept	58.00	1.42	40.84	0.00	
SBAdmin	-0.97	2.04	-0.47	0.64	
Pictures	-8.65	1.93	-4.48	0.00	
cohProb					0.069
Intercept	63.72	2.27	28.04	0.00	
SBAdmin	-6.97	3.24	-2.15	0.03	
Pictures	-11.93	3.01	-3.96	0.00	
cohGoal					0.052
Intercept	67.95	1.60	42.56	0.00	
SBAdmin	-3.73	2.49	-1.50	0.13	
Pictures	-8.15	2.25	-3.62	0.00	
cohVal					0.094
Intercept	64.74	2.11	30.73	0.00	
SBAdmin	-6.49	3.13	-2.07	0.04	
Pictures	-14.39	2.83	-5.08	0.00	
cohCon					0.051
Intercept	57.95	2.40	24.12	0.00	
SBAdmin	-6.17	3.39	-1.82	0.07	
Pictures	-10.63	3.15	-3.38	0.00	
cohSol					0.04
Intercept	68.59	1.46	47.12	0.00	
SBAdmin	-3.35	2.13	-1.58	0.12	
Pictures	-5.76	1.99	-2.89	0.00	
thorProb					0.049
Intercept	48.03	2.27	21.16	0.00	
SBAdmin	-4.64	2.85	-1.63	0.10	
Pictures	-9.68	2.89	-3.35	0.00	
thorGoal					0.088

Intercept	59.08	1.75	33.76	0.00	
SBAdmin	-5.78	2.61	-2.21	0.03	
Pictures	-11.48	2.44	-4.70	0.00	
thorVal					0.096
Intercept	53.07	2.23	23.84	0.00	
SBAdmin	-4.28	2.88	-1.49	0.14	
Pictures	-15.14	2.80	-5.41	0.00	
thorCon					0.049
Intercept	48.41	2.26	21.42	0.00	
SBAdmin	-4.53	2.98	-1.52	0.13	
Pictures	-10.08	2.88	-3.51	0.00	
thorSol					0.06
Intercept	66.50	1.85	35.92	0.00	
SBAdmin	-3.15	2.42	-1.30	0.19	
Pictures	-9.45	2.36	-4.01	0.00	
thorTot					0.093
Intercept	54.55	1.70	32.02	0.00	
SBAdmin	-4.38	2.25	-1.95	0.05	
Pictures	-11.02	2.19	-5.04	0.00	
cohTot					0.083
Intercept	63.85	1.65	38.69	0.00	
SBAdmin	-5.01	2.41	-2.08	0.04	
Pictures	-10.29	2.27	-4.54	0.00	
vignTot					0.086
Intercept	53.80	1.43	37.65	0.00	
SBAdmin	-3.52	2.01	-1.76	0.08	
Pictures	-9.44	1.92	-4.91	0.00	

Note. The robust estimator does not provide the F-statistic for the regression; it provides a t-value instead. SBAdmin = School-based administrator status.

Table 7. Robust regression summary table for effect of self-rated expertise on the variables

	Est	SE	*T-value*	*P*	R^2
Qual					0.066
Intercept	58.43	3.27	17.88	0.00	
Self-rate Exp	-0.07	0.04	-1.51	0.13	
Pictures	-11.94	2.27	-5.25	0.00	
Crea					0.062
Intercept	57.95	3.49	16.60	0.00	
Self-rate Exp	-0.14	0.05	-2.97	0.00	
Pictures	-11.55	2.40	-4.82	0.00	
Feas					0.046
Intercept	65.62	3.28	20.01	0.00	
Self-rate Exp	-0.05	0.04	-1.21	0.23	
Pictures	-9.33	2.24	-4.16	0.00	
Effct					0.047
Intercept	60.57	3.74	16.18	0.00	
Self-rate Exp	-0.10	0.05	-2.00	0.05	
Pictures	-10.59	2.49	-4.25	0.00	
cohOvr					0.062
Intercept	61.04	2.89	21.09	0.00	
Self-rate Exp	-0.05	0.04	-1.22	0.22	
Pictures	-9.55	1.92	-4.98	0.00	
cohProb					0.075
Intercept	74.20	4.19	17.70	0.00	
Self-rate Exp	-0.17	0.05	-3.11	0.00	
Pictures	-16.03	2.97	-5.40	0.00	
cohGoal					0.054
Intercept	72.95	3.42	21.33	0.00	
Self-rate Exp	-0.08	0.05	-1.76	0.08	
Pictures	-10.34	2.24	-4.61	0.00	
cohVal					0.087
Intercept	69.86	4.37	16.00	0.00	
Self-rate Exp	-0.09	0.06	-1.62	0.11	
Pictures	-17.26	2.90	-5.96	0.00	
cohCon					0.052
Intercept	66.18	4.58	14.46	0.00	
Self-rate Exp	-0.14	0.06	-2.24	0.03	
Pictures	-14.00	3.11	-4.50	0.00	
cohSol					0.035
Intercept	70.09	2.78	25.22	0.00	
Self-rate Exp	-0.03	0.04	-0.85	0.40	
Pictures	-7.12	1.93	-3.69	0.00	
thorProb					0.06
Intercept	57.65	3.81	15.13	0.00	
Self-rate Exp	-0.15	0.05	-3.06	0.00	
Pictures	-12.85	2.74	-4.69	0.00	
thorGoal					0.09

Intercept	66.46	3.62	18.37	0.00	
Self-rate Exp	-0.12	0.05	-2.53	0.01	
Pictures	-14.69	2.41	-6.09	0.00	
thorVal					0.098
Intercept	59.97	4.16	14.43	0.00	
Self-rate Exp	-0.11	0.05	-2.07	0.04	
Pictures	-17.61	2.77	-6.36	0.00	
thorCon					0.057
Intercept	57.34	4.07	14.08	0.00	
Self-rate Exp	-0.14	0.05	-2.66	0.01	
Pictures	-13.07	2.80	-4.67	0.00	
thorSol					0.067
Intercept	72.99	3.45	21.14	0.00	
Self-rate Exp	-0.10	0.04	-2.30	0.02	
Pictures	-11.66	2.32	-5.02	0.00	
thorTot					0.103
Intercept	62.45	2.96	21.12	0.00	
Self-rate Exp	-0.13	0.04	-3.21	0.00	
Pictures	-13.79	2.11	-6.54	0.00	
cohTot					0.084
Intercept	70.07	3.24	21.62	0.00	
Self-rate Exp	-0.10	0.04	-2.38	0.02	
Pictures	-13.03	2.23	-5.84	0.00	
vignTot					0.096
Intercept	60.30	2.66	22.67	0.00	
Self-rate Exp	-0.10	0.04	-2.90	0.00	
Pictures	-11.78	1.86	-6.34	0.00	

Note. The robust estimator does not provide the F-statistic for the regression; it provides a t-value instead. Self-rate Exp = Self-rated expertise.

Table 8. Robust regression summary table for effect of years of professional experience on the variables

	Est	SE	*T-value*	*P*	R^2
Qual					0.067
Intercept	56.51	2.37	23.88	0.00	
Years Exp	-0.21	0.15	-1.40	0.16	
Pictures	-11.18	2.24	-4.98	0.00	
Crea					0.048
Intercept	50.90	2.43	20.98	0.00	
Years Exp	-0.21	0.16	-1.36	0.18	
Pictures	-9.80	2.39	-4.11	0.00	
Feas					0.042
Intercept	62.54	2.11	29.64	0.00	
Years Exp	-0.05	0.13	-0.39	0.69	
Pictures	-8.57	2.09	-4.10	0.00	
Effct					0.039
Intercept	55.01	2.49	22.11	0.00	
Years Exp	-0.12	0.15	-0.77	0.44	
Pictures	-9.27	2.41	-3.85	0.00	
cohOvr					0.059
Intercept	58.06	1.87	31.03	0.00	
Years Exp	-0.02	0.12	-0.16	0.87	
Pictures	-8.91	1.84	-4.84	0.00	
cohProb					0.073
Intercept	68.38	3.05	22.40	0.00	
Years Exp	-0.48	0.19	-2.56	0.01	
Pictures	-14.01	2.89	-4.84	0.00	
cohGoal					0.049
Intercept	69.03	2.20	31.38	0.00	
Years Exp	-0.14	0.15	-0.96	0.34	
Pictures	-9.24	2.13	-4.33	0.00	
cohVal					0.084
Intercept	65.76	2.99	21.97	0.00	
Years Exp	-0.19	0.19	-1.03	0.30	
Pictures	-16.10	2.78	-5.80	0.00	
cohCon					0.047
Intercept	60.36	3.21	18.83	0.00	
Years Exp	-0.29	0.19	-1.51	0.13	
Pictures	-12.38	3.06	-4.04	0.00	
cohSol					0.033
Intercept	68.00	1.96	34.69	0.00	
Years Exp	0.00	0.12	-0.01	0.99	
Pictures	-6.71	1.87	-3.60	0.00	
thorProb					0.056
Intercept	52.11	3.05	17.11	0.00	
Years Exp	-0.40	0.18	-2.22	0.03	
Pictures	-11.16	2.76	-4.05	0.00	
thorGoal					0.077

Intercept	59.68	2.38	25.12	0.00	
Years Exp	-0.14	0.15	-0.90	0.37	
Pictures	-13.08	2.34	-5.60	0.00	
thorVal					0.093
Intercept	54.39	2.97	18.30	0.00	
Years Exp	-0.18	0.19	-0.95	0.34	
Pictures	-16.32	2.76	-5.91	0.00	
thorCon					0.045
Intercept	49.67	3.03	16.37	0.00	
Years Exp	-0.17	0.18	-0.96	0.34	
Pictures	-11.32	2.80	-4.05	0.00	
thorSol					0.056
Intercept	66.73	2.33	28.64	0.00	
Years Exp	-0.07	0.14	-0.49	0.63	
Pictures	-10.32	2.26	-4.57	0.00	
thorTot					0.088
Intercept	56.03	2.22	25.22	0.00	
Years Exp	-0.19	0.14	-1.33	0.18	
Pictures	-12.25	2.12	-5.77	0.00	
cohTot					0.077
Intercept	65.56	2.24	29.28	0.00	
Years Exp	-0.22	0.15	-1.50	0.14	
Pictures	-11.73	2.18	-5.39	0.00	
vignTot					0.084
Intercept	55.36	1.91	29.00	0.00	
Years Exp	-0.18	0.12	-1.45	0.15	
Pictures	-10.48	1.83	-5.71	0.00	

Note. The robust estimator does not provide the F-statistic for the regression; it provides a t-value instead. Years Exp = Years of professional experience in schools.

APPENDIX O

POST-PILOT FEEDBACK SURVEY RESULTS

Q2.1 Overall, the training itself was _____.	Results: N size
○ woefully insufficient	0
○ insufficient	1
○ sufficient	4
○ thorough	5
○ extremely thorough	0

Q2.2 The training slides were _____.	Results: N Size
○ not clear	0
○ a little clear	1
○ somewhat clear	3
○ extremely clear	6

Q2.3 The voice-over for the training was _____.	Results
○ not clear	0
○ a little clear	0
○ somewhat clear	1
○ extremely clear	9

Q2.4 The examples provided in the training were _____.	Results: N Size
○ Not helpful at all	0
○ A little helpful	0
○ Somewhat helpful	6
○ Very helpful	4

Q2.5 The duration of the training was _____.	Results: N Size
○ too long	0
○ somewhat long	3
○ about right	7
○ somewhat short	0
○ too short	0

Q2.6 How often did you want to ask questions but were unable to do so because the training was conducted via PowerPoint?	Results: N Size
○ No Questions	4
○ Question(s) on a few slides	3
○ Question(s) on many slides	2
○ Question(s) on every slide	1

Q3.1 How user-friendly was the Scoring Method overall?	Results: N Size
○ Very hard to use	0
○ Hard to use	3
○ Adequate	3
○ Easy	4
○ Very easy	0

Q3.2 How hard was it to pick out the pertinent information when you were reading the

	Extremely easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Extremely difficult
UNPROMPTED responses?	○	○	○ (2)	○ (7)	○ (1)
PROMPTED responses?	○ (3)	○ (6)	○ (1)	○	○

Q3.3 Did it become easier over time to pick out the pertinent information for the

	Not at all	A little easier	Somewhat easier	A lot easier
UNPROMPTED responses?	○	○ (3)	○ (6)	○ (1)
PROMPTED responses?	○ (1)	○ (1)	○ (2)	○ (6)

Q3.4 Did the variables seem appropriate for characterizing and evaluating the responses?

	Entirely INappropriate	Somewhat INappropriate	Somewhat Appropriate	Entirely Appropriate
Quality	○	○	○ (2)	○ (8)
Creativity	○	○ (1)	○ (5)	○ (4)
Feasibility	○	○	○ (4)	○ (6)
Effectiveness	○	○	○ (4)	○ (6)
Coherence	○	○ (4)	○ (4)	○ (2)
Thoroughness	○	○ (1)	○ (5)	○ (4)

Q3.6 How confident did you feel in your ability to score responses with the following variables?

	No Confidence	A little Confidence	Some Confidence	Full Confidence
Quality	○	○	○ (2)	○ (8)
Creativity	○	○	○ (5)	○ (5)
Feasibility	○	○ (1)	○ (3)	○ (6)
Effectiveness	○	○ (1)	○ (3)	○ (6)
Coherence Overall	○ (1)	○ (3)	○ (3)	○ (3)
Coherence by Component	○ (1)	○ (4)	○ (4)	○ (1)
Thoroughness	○	○ (1)	○ (5)	○ (4)

Q3.7 The directions for scoring each variable were _____.

	Not Clear At All	A Little Clear	Somewhat Clear	Quite Clear
Quality	○	○	○ (4)	○ (6)
Creativity	○	○	○ (4)	○ (6)
Feasibility	○	○ (1)	○ (2)	○ (7)
Effectiveness	○	○	○ (3)	○ (7)
Coherence Overall	○ (1)	○ (1)	○ (5)	○ (3)
Coherence by component	○ (1)	○ (2)	○ (4)	○ (3)
Thoroughness	○	○ (1)	○ (4)	○ (5)

Q3.8 Regardless of the clarity of the directions, what the directions required of you was _____.

	Quite Simple	Somewhat Simple	Somewhat Complicated	Quite Complicated
Quality	○ (3)	○ (5)	○ (2)	○
Creativity	○ (3)	○ (6)	○ (1)	○
Feasibility	○ (3)	○ (5)	○ (2)	○
Effectiveness	○ (3)	○ (6)	○ (1)	○
Coherence Overall	○ (1)	○ (2)	○ (7)	○
Coherence by component	○ (1)	○ (1)	○ (7)	○ (1)
Thoroughness	○ (2)	○ (5)	○ (3)	○

Q3.9 Did you develop a preference for scoring Coherence?	Results: N Size
○ Coherence overall	5
○ Coherence by component	0
○ No Preference	5

Were there any issues, aversive or otherwise, that you kept encountering while scoring that were not addressed in the training?

As I said before there were times that more information would have made it easier to know what to do. After doing this for so many years one can think of lots of different variables which would impact decision making.	1
I think the training was thorough but I made the mistake of taking the training and then letting time lapse before completing the scoring.	1
It was difficult to express that a suggested response to the scenario was not in line with what is expected from an administrator. Aside from marking the response as efficient (or not efficient), but ineffective, the rest of the responses created dissonance. Continuing to address alignment within the response, etc. was difficult (but not impossible). Also, I became confused as to what I should be scoring when responses to certain sections might be addressed outside of the prompted boxes. For example, if the problem description was addresses in another section (other than the problem description box), should I score that as a good description of the problem. This was probably addressed in the training, but I needed reminders as to how to score.	1
it was often difficult to differentiate between the thoroughness and coherence. It would help if you could provide popups next to each category to remind the reader of the definition of each response and definition of each category. There is so much to remember from the training that this type of reminder may produce better results.	1
None	1
None. Due to the time between scoring, I needed to review and refresh certain aspects. If I had sat down through the entire scoring session across a two day period, this would have been mitigated.	1
The definitions could have been better illustrated or described. It would have been helpful to have these definitions within the actual scoring. I had the powerpoint example also available to me and this was a bit cumbersome and I was already managing two monitors when scoring.	1
The part you had to return to the same browser. Trying to pick up the survey in another computer was problematic which slow the progress of completion. Kind of like this survey, which I started, but did not complete then I had to start over.	1
The training is good and administrators would benefit greatly from the training.	1
NA	1

Please share any other feedback about the training to help improve it.

I felt the training was sufficient	1

I think I already described my main issue with the actual scoring and the coherence piece.	1
I thought it was well done. Sorry no negative!	1
If we are talking about the training (Powerpoint) then I thought it was thorough enough so that I could attend to the scoring. Again, I completed the training and attended to the scoring at a later date when I should have done both within a reasonable time frame.	1
My memory is not precise, but in many of the scenarios, suspension seems to be overused inappropriately. It would be great to see scenarios more consistent with the new OARs surrounding disciplinary removal from instructional hours. Thanks for the opportunity to participate and sorry for the delay.	1
Potentially, there should be a cutoff for needing to score the remainder of the response. Potentially, if the response is scored as tremendously ineffective, or potentially harmful to students, then the alignment questions should not be addressed. I know that this solution would cause other problems. Just trying to provide constructive feedback. As for the second issue, I just think that periodic reminders of general scoring expectations should be provided throughout (optional, maybe).	1
The unprompted responses were more difficult to score and made scoring more subjective than I would have liked. Parcelling out the specific components was time consuming.	1
NA	3

Q3.5 Were there other variables or characteristics that you felt should have been included but weren't? If so, what were they? (See Appendix C for responses.)

Appropriate expectations of a licensed/experienced administrator	1
Closure. I would prefer this variable instead of "Thoroughness" and would make it the last component.	1
I often wished there was a comment box to explain some of the answers, as these were very subjective responses using an objective scoring system.	1
I think the way the questions were embedded in a rating scale devalued the importance of some higher value questions while also placing more value on some (creativity) that I think were of less value, in various answers. For example, a highly creative response may get high points but be completely unrealistic or effective. Obviously you must be accounting for that in the analysis. The tendency for a score however, is to not want to get very many points for creativity if the answer is completely absurd. So I wonder how that feeling will impact the overall results.	1
Legality of responses was missing. Equity and bias within responses was missing. The above answers are hard to rate without an "other" box. I am scoring them, but I don't like having just the choices provided and they are not necessarily accurate. So many of the answers provided in the scenarios were partway there and the variables are somewhat subjective. Creativity--subjective feasibility--most of the responses were feasible but may not have been very effective Effective--could have been effective but not necessarily equitable or the right thing to do. Coherence--some answers were sorely inadequate, but coherently inadequate the whole way through. High coherence? (I spoke to this in my feedback after scoring).	1
NA	5

Please share any other feedback about the scoring experience to help improve it.

Ability to write notes within the scoring.	1
See comment earlier	1
Thanks, Josh.	1
The "coherence" piece is what was confusing for me as was the "feasibility" component. Examples or prompts would have helped, i.e., An example of coherence for this vignette might look like...	1
NA	6

APPENDIX P

FIELD TEST INTER-ITEM CORRELATION MATRIX

	qual	effct	crea	feas	coh	thor Prob	thor Goal	thor Val	thor Con	thor Sol	thor Tot	vign Tot
qual	1											
effct	.81	1										
crea	.65	.55	1									
feas	.60	.64	.36	1								
coh	.68	.7	.48	.50	1							
Thor Prob	.55	.50	.39	.40	.51	1						
Thor Goal	.51	.48	.42	.32	.56	.56	1					
Thor Val	.50	.44	.39	.30	.53	.54	.71	1				
Thor Con	.49	.46	.39	.33	.53	.49	.60	.62	1			
Thor Sol	.59	.65	.47	.43	.58	.42	.50	.45	.49	1		
Thor Tot	.66	.63	.52	.45	.68	.76	.85	.84	.81	.71	1	
Vign Tot	.84	.82	.68	.64	.80	.72	.77	.75	.73	.74	.93	1

APPENDIX Q

CLASSIC ANOVA AND REGRESSION TABLES FOR THE FIELD TEST

Table 1. ANOVA Summary Table for effect of Program Enrollment on Vignette Totals

Factor	*df*	Mean Squares	*F*	*P*
Cyber				
Type	2	1.48	0.01	0.995
Residuals	115	268.21		
Shooter				
Type	2	165.18	0.84	0.435
Residuals	115	196.78		
GoBack				
Type	2	147.68	0.61	0.548
Residuals	115	243.97		
Homo-phobia				
Type	2	91.46	0.37	0.689
Residuals	115	244.62		
Para				
Type	2	494.32	1.88	0.157
Residuals	115	262.83		

Table 2. Robust regression summary table for effect of school-based administrator status on vignette totals

Model	*Est*	SE	*T-value*	*P*	R^2
Cyber					0.002
Intercept	39.15	1.80	21.70	0.00	
SBAdmin	-1.74	3.41	-0.51	0.61	
Shooter					0.018
Intercept	39.96	1.52	26.32	0.00	
SBAdmin	4.64	3.40	1.36	0.18	
GoBack					0.001
Intercept	56.11	1.80	31.13	0.00	
SBAdmin	1.35	3.37	0.40	0.69	
Homo-phobia					0.03
Intercept	55.37	1.69	32.73	0.00	
SBAdmin	6.57	3.45	1.90	0.06	
Para					0.016
Intercept	53.39	1.84	29.04	0.00	
SBAdmin	5.07	3.32	1.53	0.13	

Note. The robust estimator does not provide the F-statistic for the regression; it provides a t-value instead. SBAdmin = School-based administrator status.

Table 3. Robust regression summary table for effect of self-rated expertise on vignette totals

Model	*Est*	SE	*T-value*	*P*	R^2
Cyber					0.000
Intercept	38.88	3.14	12.37	0.00	
Self-rate Exp	0.00	0.05	-0.06	0.95	
Shooter					0.015
Intercept	37.03	2.85	13.00	0.00	
Self-rate Exp	0.08	0.05	1.65	0.10	
GoBack					0.001
Intercept	57.18	3.20	17.86	0.00	
Self-rate Exp	-0.02	0.05	-0.33	0.74	
Homophobia					0.013
Intercept	52.18	3.17	16.47	0.00	
Self-rate Exp	0.08	0.05	1.56	0.12	
Para					0.027
Intercept	48.24	3.30	14.61	0.00	
Self-rate Exp	0.11	0.05	2.05	0.04	

Note. The robust estimator does not provide the F-statistic for the regression; it provides a t-value instead. Self-rate Exp = Self-rated expertise.

Table 4. Robust regression summary table for effect of years professionally in schools on vignette totals

Model	*Est*	SE	*T-value*	*P*	R^2
Cyber					0.006
Intercept	-0.63	0.09	-6.88	0.00	
Years Exp	-0.07	0.06	-1.09	0.28	
Shooter					0.006
Intercept	-0.49	0.07	-6.51	0.00	
Years Exp	0.09	0.07	1.14	0.26	
GoBack					0.006
Intercept	0.38	0.08	4.61	0.00	
Years Exp	-0.05	0.08	-0.66	0.51	
Homophobia					0.002
Intercept	0.40	0.08	4.84	0.00	
Years Exp	-0.02	0.08	-0.19	0.85	
Para					0.000
Intercept	0.27	0.09	3.08	0.00	
Years Exp	0.02	0.09	0.20	0.84	

Note. The robust estimator does not provide the F-statistic for the regression; it provides a t-value instead. Years Exp = Years of professional experience.

Table 5. ANOVA Summary Table for effect of Program Enrollment on Variables

	df	Mean Squares	*F*	*P*

Qual				
Type	2	1456.61	3.57	0.029
Residuals	587	407.86		
Crea				
Type	2	114.56	0.24	0.788
Residuals	587	481.85		
Feas				
Type	2	1725.82	5.26	0.005
Residuals	587	327.89		
Effct				
Type	2	1592.40	3.45	0.032
Residuals	587	461.35		
Coh				
Type	2	584.26	1.80	0.166
Residuals	587	324.47		
thorProb				
Type	2	470.72	0.73	0.482
Residuals	587	643.40		
thorGoal				
Type	2	719.15	1.06	0.348
Residuals	587	680.30		
thorVal				
Type	2	47.23	0.07	0.934
Residuals	587	688.99		
thorCon				
Type	2	169.00	0.26	0.774
Residuals	587	659.26		
thorSol			3.98	0.019
Type	2	1840.59		
Residuals	587	462.10		
thorTot			0.39	0.677
Type	2	155.09		
Residuals	587	397.38		
vignTot			1.48	0.228
Type	2	439.97		
Residuals	587	296.41		

Table 6. Robust regression summary table for effect of school-based administrator status on the variables

	Est	SE	*T-value*	*P*	R^2
Qual					0.015
Intercept	42.66	1.03	41.51	0.00	
SBAdmin	6.15	2.03	3.03	0.00	
Crea					0.001
Intercept	34.50	1.17	29.5	0.00	
SBAdmin	1.37	2.10	0.65	0.52	
Feas					0.004
Intercept	61.27	0.98	62.64	0.00	
SBAdmin	2.70	1.69	1.59	0.11	
Effct					0.013
Intercept	47.23	1.10	42.97	0.00	
SBAdmin	6.21	2.14	2.9	0.00	
Coh					0.000
Intercept	51.44	0.89	57.49	0.00	
SBAdmin	0.74	1.83	0.4	0.69	
thorProb					0.003
Intercept	49.60	1.27	38.92	0.00	
SBAdmin	-3.35	2.77	-1.21	0.23	
thorGoal					0.008
Intercept	49.82	1.31	38.16	0.00	
SBAdmin	5.76	2.82	2.04	0.04	
thorVal					0.002
Intercept	43.41	1.32	33	0.00	
SBAdmin	3.20	2.82	1.14	0.26	
thorCon					0.002
Intercept	49.80	1.30	38.21	0.00	
SBAdmin	2.70	2.91	0.93	0.35	
thorSol					0.014
Intercept	62.66	1.11	56.27	0.00	
SBAdmin	6.15	2.03	3.03	0.00	
thorTot					0.004
Intercept	50.63	0.99	51.21	0.00	
SBAdmin	3.13	2.12	1.47	0.14	
vignTot					0.006
Intercept	48.80	0.86	57.05	0.00	
SBAdmin	3.41	1.78	1.91	0.06	

Note. The robust estimator does not provide the F-statistic for the regression; it provides a t-value instead. SBAdmin = School-based administrator status.

Table 7. Robust regression summary table for effect of self-rated expertise on the variables

	Est	SE	*T-value*	*P*	R^2
Qual					0.015
Intercept	38.95	2.01	19.38	0.00	
Self-rate Exp	0.10	0.03	2.89	0.00	
Crea					0.001
Intercept	33.29	2.05	16.21	0.00	
Self-rate Exp	0.03	0.03	0.85	0.39	
Feas					0.020
Intercept	56.83	1.98	28.73	0.00	
Self-rate Exp	0.09	0.03	3.09	0.00	
Effct					0.017
Intercept	42.71	2.13	20.09	0.00	
Self-rate Exp	0.11	0.04	3.15	0.00	
coh					0.004
Intercept	49.25	1.81	27.16	0.00	
Self-rate Exp	0.04	0.03	1.45	0.15	
thorProb					0.000
Intercept	48.60	2.62	18.55	0.00	
Self-rate Exp	0.00	0.04	0.10	0.92	
thorGoal					0.000
Intercept	50.00	2.57	19.47	0.00	
Self-rate Exp	0.02	0.04	0.48	0.63	
thorVal					0.001
Intercept	42.83	2.65	16.16	0.00	
Self-rate Exp	0.02	0.04	0.56	0.58	
thorCon					0.001
Intercept	51.64	2.45	21.11	0.00	
Self-rate Exp	-0.02	0.04	-0.57	0.57	
thorSol					0.009
Intercept	59.88	2.30	26.05	0.00	
Self-rate Exp	0.08	0.04	2.14	0.03	
thorTot					0.001
Intercept	50.18	1.96	25.60	0.00	
Self-rate Exp	0.02	0.03	0.66	0.51	
vignTot					0.005
Intercept	47.02	1.68	27.92	0.00	
Self-rate Exp	0.05	0.03	1.68	0.09	

Note. The robust estimator does not provide the F-statistic for the regression; it provides a t-value instead. Self-rate Exp = Self-rated expertise.

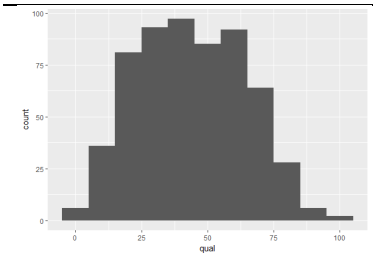
Table 8. Robust regression summary table for effect of years of professional experience on the variables

	Est	SE	*T-value*	*P*	R^2
Qual					0.000
Intercept	43.97	1.36	32.35	0.00	
Years Exp	0.01	0.10	0.13	0.90	
Crea					0.002
Intercept	35.87	1.47	24.48	0.00	
Years Exp	-0.11	0.11	-0.96	0.34	
Feas					0.000
Intercept	61.65	1.31	47.22	0.00	
Years Exp	0.03	0.09	0.29	0.77	
Effct					0.001
Intercept	47.87	1.46	32.84	0.00	
Years Exp	0.08	0.11	0.77	0.44	
Coh					0.000
Intercept	51.57	1.24	41.52	0.00	
Years Exp	0.00	0.10	0.03	0.97	
thorProb					0.003
Intercept	50.50	1.83	27.59	0.00	
Years Exp	-0.17	0.14	-1.21	0.23	
thorGoal					0.000
Intercept	51.07	1.85	27.68	0.00	
Years Exp	0.00	0.15	0.03	0.98	
thorVal					0.001
Intercept	45.25	1.83	24.78	0.00	
Years Exp	-0.12	0.14	-0.81	0.42	
thorCon					0.001
Intercept	51.58	1.73	29.79	0.00	
Years Exp	-0.13	0.14	-0.92	0.36	
thorSol					0.001
Intercept	63.42	1.49	42.58	0.00	
Years Exp	0.07	0.12	0.62	0.53	
thorTot					0.001
Intercept	52.04	1.36	38.16	0.00	
Years Exp	-0.07	0.10	-0.72	0.47	
vignTot					0.000
Intercept	49.92	1.14	43.83	0.00	
Years Exp	-0.04	0.08	-0.43	0.67	

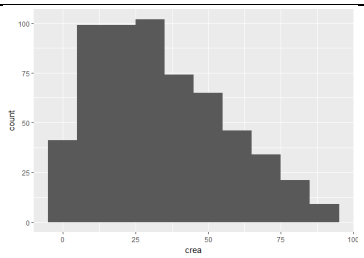
Note. The robust estimator does not provide the F-statistic for the regression; it provides a t-value instead. Years Exp = Years of professional experience in schools.

APPENDIX R

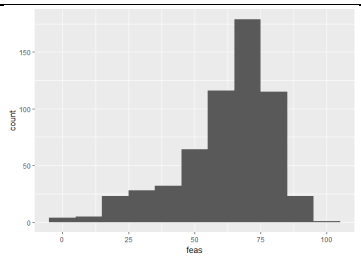
UNIVARIATE PLOTS OF FIELD TEST VARIABLES



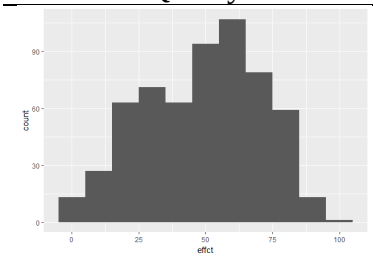
Quality



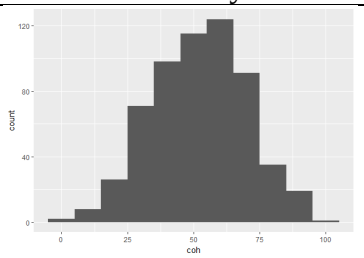
Creativity



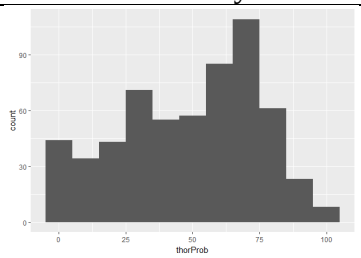
Feasibility



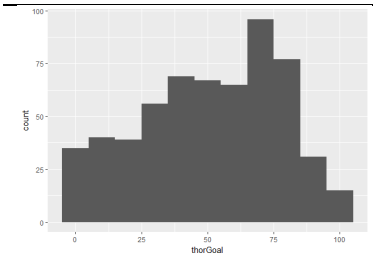
Effectiveness



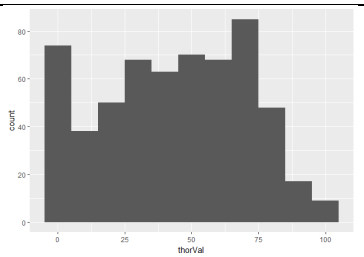
Coherence



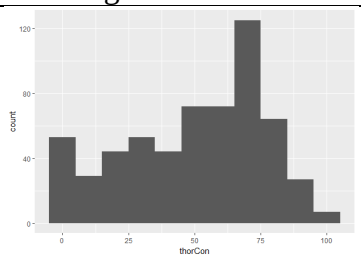
Thoroughness of Problems



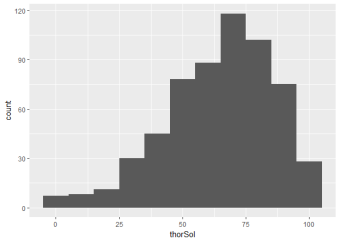
Thoroughness Goals



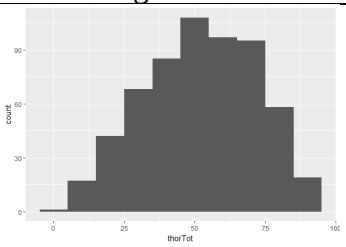
Thoroughness Values



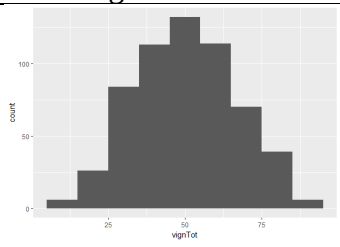
Thoroughness Constraints



Thoroughness Solutions



Thoroughness Total



Vignette Total

APPENDIX S

LEVENE'S TEST RESULTS FOR FIELD TEST VIGNETTE TOTALS & VARIABLES

WITH CATEGORICAL PROXY VARIABLES

Table 1. Levene's Test Results for Field Test Vignettes and Program Enrollment

	Df	F-statistic	p-value
Vignette 1	2, 115	0.00	1.00
Vignette 2	2, 115	0.80	0.45
Vignette 3	2, 115	0.17	0.84
Vignette 4	2, 115	0.12	0.89
Vignette 5	2, 115	0.56	0.57

Table 2. Levene's Test Results for Field Test Vignettes and School-based Admin Status

	Df	F-statistic	p-value
Vignette 1	1, 116	0.32	0.58
Vignette 2	1, 116	0.12	0.73
Vignette 3	1, 116	1.84	0.18
Vignette 4	1, 116	0.19	0.66
Vignette 5	1, 116	2.43	0.12

Table 3. Levene's Test Results for Field Test Variables and Program Enrollment

	Df	F-statistic	p-value
Qual	2, 587	0.10	0.91
Crea	2, 587	1.62	0.20
Feas	2, 587	4.40	0.01
Effct	2, 587	0.15	0.86
Coh	2, 587	0.12	0.89
thorProb	2, 587	0.42	0.66
thorGoal	2, 587	0.90	0.41
thorVal	2, 587	0.99	0.37
thorCon	2, 587	2.30	0.10
thorSol	2, 587	1.16	0.32
thorTot	2, 587	0.82	0.44
vignTot	2, 587	0.25	0.78

Table 4. Levene's Test Results for Field Test Variables and School-Based Admin Status

	Df	F-statistic	p-value
Qual	2, 588	0.96	0.33
Crea	2, 588	3.97	0.05
Feas	2, 588	1.98	0.16
Effct	2, 588	1.47	0.23
Coh	2, 588	0.48	0.49
thorProb	2, 588	1.07	0.30
thorGoal	2, 588	0.37	0.55
thorVal	2, 588	0.45	0.50
thorCon	2, 588	3.07	0.08
thorSol	2, 588	3.23	0.07
thorTot	2, 588	0.24	0.62
vignTot	2, 588	0.00	0.98

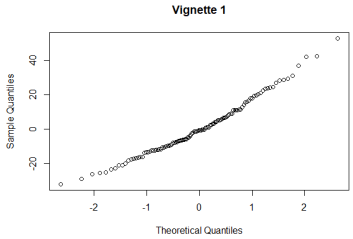
APPENDIX T

QQ PLOTS FOR VIGNETTE TOTALS AND VARIABLES REGRESSED ONTO SCHOOL-BASED ADMINISTRATOR STATUS, SELF-RATED EXPERTISE, AND YEARS PROFESSIONALLY IN SCHOOLS

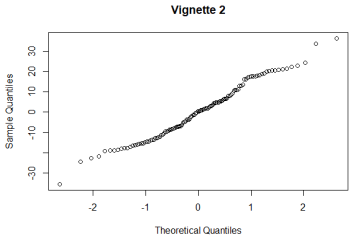
The Vignettes

School-based Administrator Status

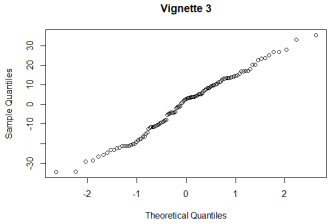
Vignette 1



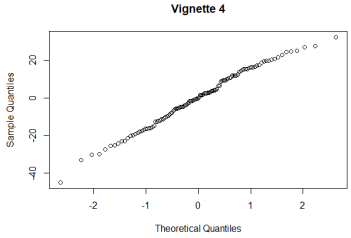
Vignette 2



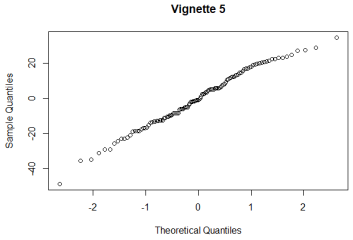
Vignette 3



Vignette 4

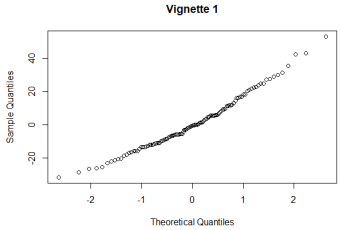


Vignette 5

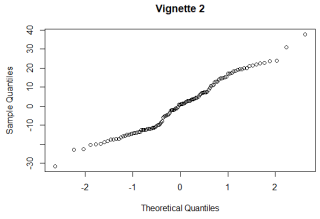


Self-rated Expertise

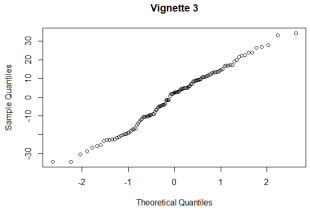
Vignette 1



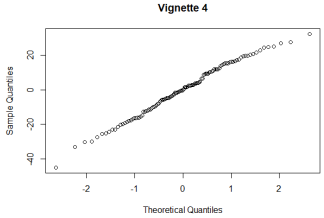
Vignette 2



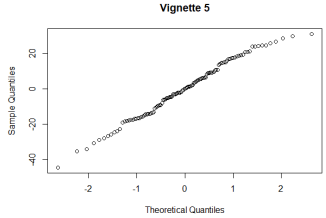
Vignette 3



Vignette 4

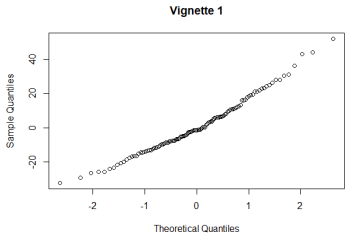


Vignette 5

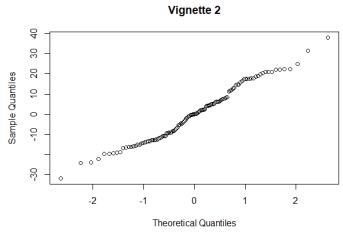


Years Professionally in Schools

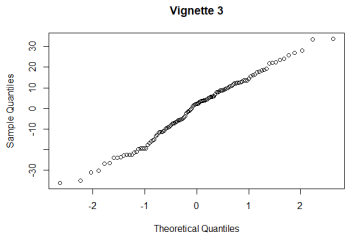
Vignette 1



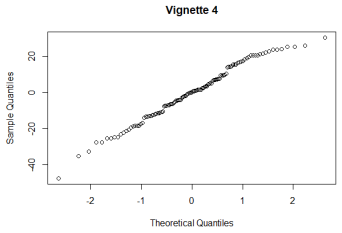
Vignette 2



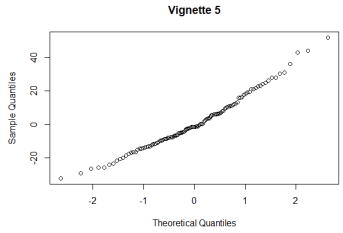
Vignette 3



Vignette 4



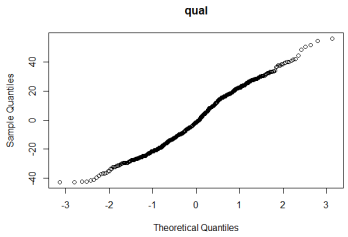
Vignette 5



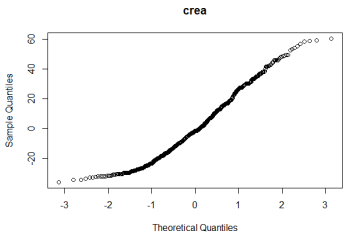
The Variables

School-based Administrator Status

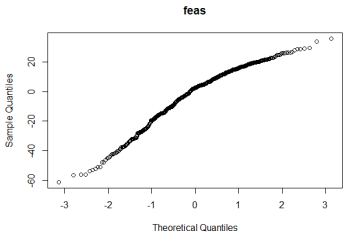
Quality



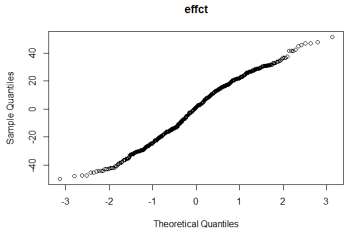
Creativity



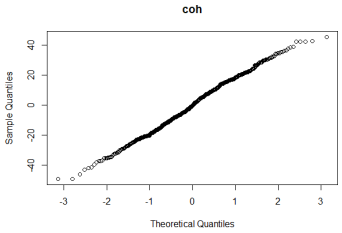
Feasibility



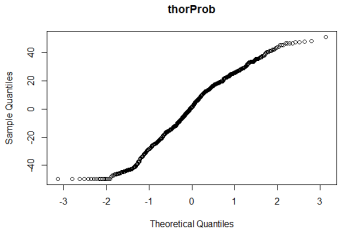
Effectiveness



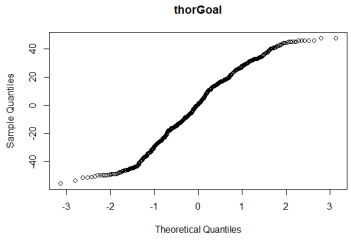
Coherence



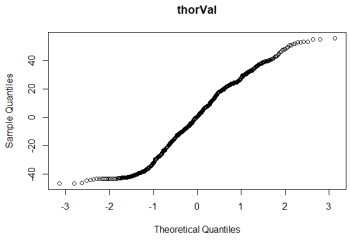
Thoroughness of Problems



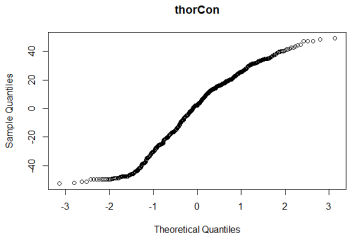
Thoroughness of Goals



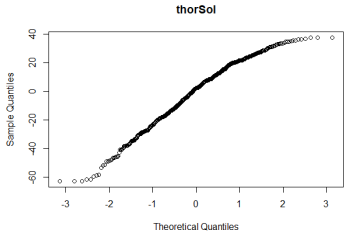
Thoroughness of Values



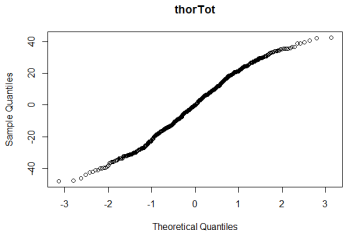
Thoroughness of Constraints



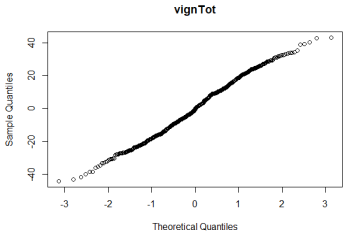
Thoroughness of Solutions



Thoroughness Total

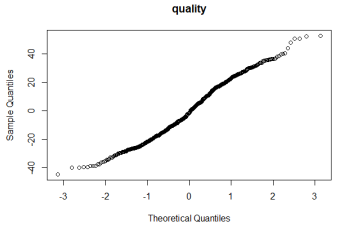


Vignette Total

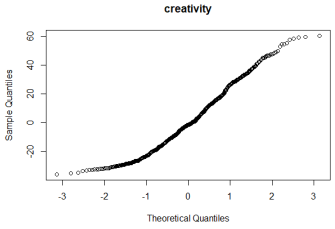


Self-rated Expertise

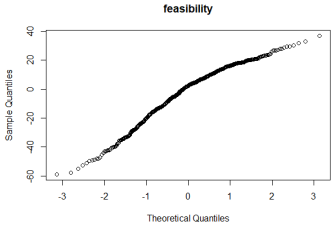
Quality



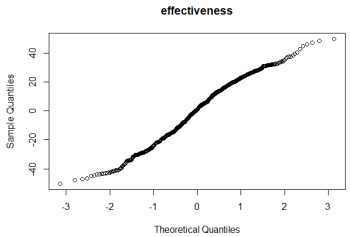
Creativity



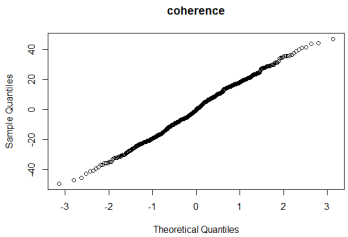
Feasibility



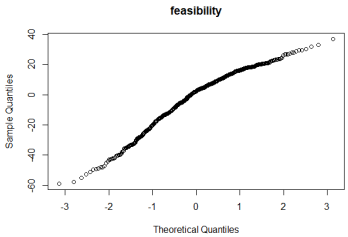
Effectiveness



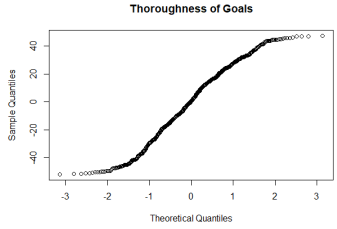
Coherence



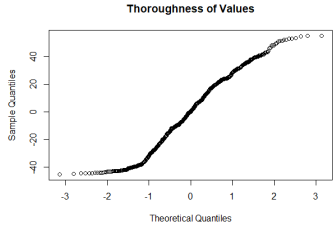
Thoroughness of Problems



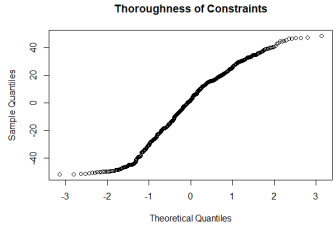
Thoroughness of Goals



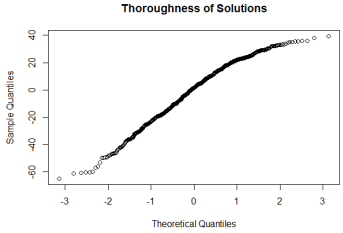
Thoroughness of Values



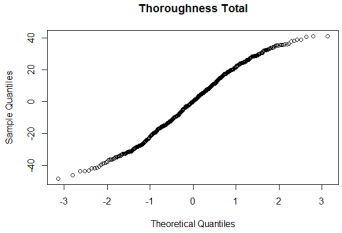
Thoroughness of Constraints



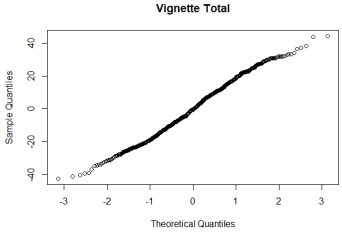
Thoroughness of Solutions



Thoroughness Total

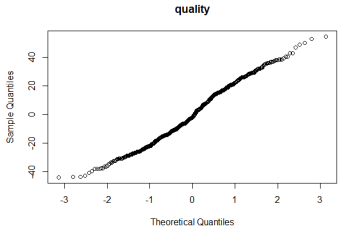


Vignette Total

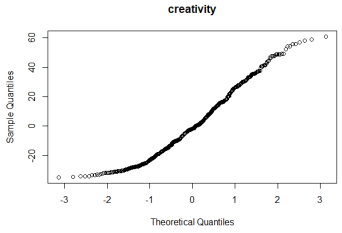


Years Professionally in Schools

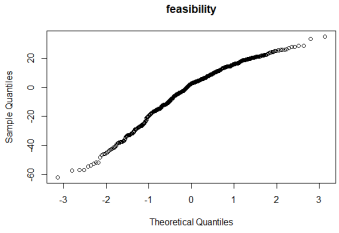
Quality



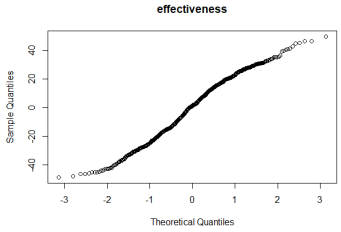
Creativity



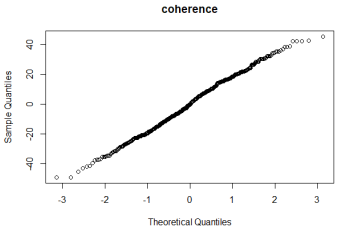
Feasibility



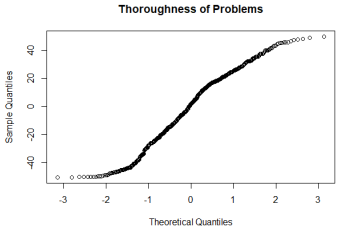
Effectiveness



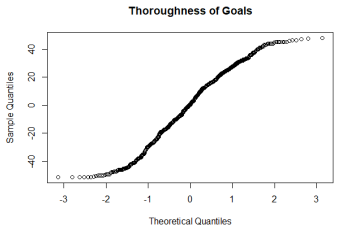
Coherence



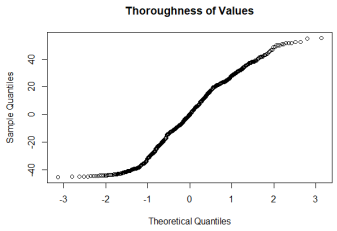
Thoroughness of Problems



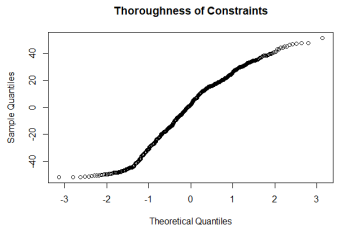
Thoroughness of Goals



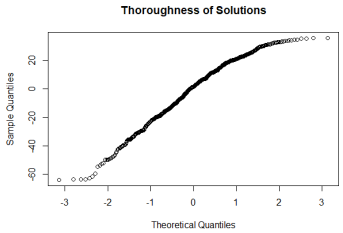
Thoroughness of Values



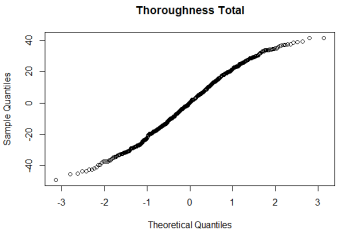
Thoroughness of Constraints



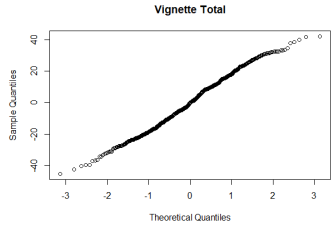
Thoroughness of Solutions



Thoroughness Total



Vignette Total



APPENDIX U

FINAL FORM RELIABILITY AND DISCRIMINANT VALIDITY COEFFICIENTS

Table 1. Reliability coefficients for final form variables

	Individual Ratings Across				Average Scores Across		
	Final Form				Unprompted	Prompted	All
	Kalpha		Mean ICC		Mean ICC	Mean ICC	Mean ICC
	Est.	95% conf. int.	(c,k)	Range	(c,k)	(c,k)	(c,k)
Qual	0.42	(0.40 - 0.45)	0.67	.51 - .78	0.76	0.69	0.79
Crea	0.23	(0.19 - 0.25)	0.56	.42 - .66	0.63	0.68	0.70
Feas	0.08	(0.04 - 0.13)	0.34	.05 - .51	0.36	0.43	0.54
Effct	0.4	(0.37 - 0.42)	0.63	.53 - .74	0.72	0.65	0.73
Coh	0.33	(0.28 - 0.36)	0.58	.26 - .78	0.67	0.45	0.61
thorProb	0.55	(0.53 - 0.57)	0.77	.60 - .90	0.8	0.78	0.79
thorGoal	0.51	(0.48 - 0.52)	0.72	.36 - .88	0.43	0.7	0.69
thorVal	0.51	(0.48 - 0.53)	0.71	.44 - .86	0.51	0.7	0.72
thorCon	0.52	(0.50 - 0.54)	0.73	.49 - .92	0.57	0.67	0.78
thorSol	0.4	(0.37 - 0.43)	0.72	.54 - .81	0.76	0.78	0.85
thorTot	0.62	(0.60 - 0.63)	0.81	.61 - .92	0.81	0.79	0.86
vignTot	0.58	(0.56 - 0.59)	0.79	.63 - .89	0.84	0.77	0.84

Table 2. ICCs for the final form vignettes

	Individual Ratings		Total Scores		Spearman-Brown	
	Mean ICC(c,k)	Range	Mean ICC(c,k)	Range	ICC if doubled	Factor for .80
Online	0.65	.58 - .80	0.81	.68 - .91	0.79	2.15
Shooter	0.66	.53 - .82	0.74	.64 - .80	0.80	2.06
Homophobia	0.67	.49 - .85	0.77	.56 - .94	0.80	1.97
Para	0.64	.20 - .79	0.66	-.02 - .86	0.78	2.25
Unprompted	0.65	.55 - .81	0.84	.76 - .91	0.79	2.15
Prompted	0.66	.47 - .75	0.77	.45 - .92	0.80	2.06
All 4	0.69	.56 - .81	0.84	.69 - .91	0.82	1.80

Note. Spearman-Brown formulae based on mean ICC for individual ratings.

Table 3. Final form variables' ability to discriminate based on program enrollment and school-based administrator status.

	Program Enrollment						School-based Admin		
	Unstandardized Mean Differences			Standardized Effect Sizes		Power	Unstandardized		Power
Variables	Ad – As	Ad – Gr	As – Gr	Eta^2	Cohen's f^2	df(2,469)	beta	(SE)	df(2,470)
Qual	0.38	4.35	3.96	.01	.10	.51	**6.53****	(2.31)	.72
Crea	-0.24	0.87	1.12	.00	.02	.07	2.91	(2.31)	.17
Feas	0.28	**4.91***	**4.62**$^{\sim}$	.02	.13	.71	2.64	(1.96)	.21
Effct	-1.54	3.63	**5.17**$^{\sim}$	.01	.10	.50	**5.71***	(2.49)	.52
Coh	-1.62	0.95	2.56	.00	.06	.18	1.64	(2.16)	.10
thorProb	-2.96	0.03	2.99	.00	.05	.15	-3.34	(3.15)	.16
thorGoal	5.14	3.85	-1.28	.01	.08	.29	**6.92***	(3.33)	.51
thorVal	1.49	0.22	-1.27	.00	.02	.07	3.26	(2.84)	.16
thorCon	1.16	1.47	0.31	.00	.02	.07	2.29	(3.31)	.10
thorSol	0.17	**5.72***	**5.55**$^{\sim}$	.02	.13	.70	**7.06****	(2.37)	.73
thorTot	1.00	2.26	1.26	.00	.05	.14	3.55	(2.45)	.27
vignTot	0.23	2.60	2.37	.01	.07	.27	**3.84**$^{\sim}$	(2.04)	.40

Note. ~ = p < .10; * = p<.05; ** = p <.001.

Table 4. Final form variables' ability to discriminate based on self-rated expertise and years professionally in schools

	Self-rated Expertise					Yrs In Schools				
	Unstandardized		Standardized		Power	Unstandardized		Standardized		Power
	beta	SE	beta	SE	df(2,470)	beta	SE	beta	SE	df(2,470)
Qual	**0.11****	(.04)	**0.15**	(.05)	.82	0.01	(.11)	0.00	(.05)	.05
Crea	0.04	(.04)	0.05	(.05)	.15	-0.09	(.13)	-0.03	(.05)	.09
Feas	**0.11****	(.03)	**0.17**	(.05)	.93	0.07	(.10)	0.03	(.04)	.08
Effct	**0.13****	(.04)	**0.16**	(.05)	.84	0.1	(.13)	0.04	(.05)	.09
Coh	0.06	(.03)	0.08	(.05)	.34	-0.01	(.11)	0.00	(.05)	.05
thorProb	0.02	(.05)	0.03	(.05)	.07	-0.08	(.16)	-0.02	(.05)	.07
thorGoal	0.04	(.05)	0.04	(.05)	.09	0.03	(.17)	0.01	(.05)	.05
thorVal	0.04	(.05)	0.05	(.05)	.12	-0.09	(.15)	-0.03	(.05)	.08
thorCon	-0.02	(.05)	-0.02	(.05)	.06	-0.12	(.16)	-0.04	(.05)	.10
thorSol	**0.10***	(.04)	**0.13**	(.05)	.66	0.12	(.13)	0.05	(.05)	.13
thorTot	0.04	(.04)	0.05	(.05)	.14	-0.03	(.11)	-0.01	(.05)	.06
vignTot	**0.06***	(.03)	**0.10**	(.05)	.44	-0.01	(.09)	-0.01	(.04)	.05

Note. ~ = p < .10; * = p < .05; ** = p < .001

Table 5. Field vignette total scores' ability to discriminate based on program enrollment and school-based administrator status

	Program Enrollment						SBAdmin		
	Unstandardized Mean Differences			Standardized Effect Sizes		Power	Unstandardized		Power
	Ad – As	Ad – Gr	As – Gr	Eta^2	Cohen's f^2	df (2,115)	beta	(SE)	df(2,116)
Online	-0.32	-0.36	-0.04	.00	.01	.05	-1.74	(3.41)	.07
Shooter	2.80	3.96	1.16	.01	.12	.20	4.64	(3.40)	.24
Harass	1.62	2.95	1.33	.01	.08	.11	**6.57***	(3.45)	.38
Para	-3.20	3.85	7.05	.03	.18	.39	5.07	(3.32)	.22

Note.

Table 6. Field test vignette total scores' ability to discriminate based on self-rated expertise and years professionally in schools.

	Self-rated Expertise					Yrs In Schools				
	Unstandardized		Standardized		Power	Unstandardized		Standardized		Power
	beta	(SE)	beta	(SE)	df(2,116)	beta	(SE)	beta	(SE)	df(2,116)
Online	0.0	(.05)	0.0	0.08	.05	-0.14	(.13)	-0.07	(.06)	.10
Shooter	**0.08***	(.05)	0.12	0.07	.30	0.18	(.16)	0.09	(.07)	.16
Harass	0.08	(.05)	0.13	0.08	.27	-0.03	(.18)	-0.02	(.08)	.05
Para	**0.11***	(.05)	0.18	0.09	.43	0.04	(.18)	0.02	(.09)	.05

Note. ~ = p < .10; * = p < .05; ** = p < .001

Final Form Analyses of Total Scores

Table 7. Final form total scores variables' ability to discriminate based on program enrollment and school-based administrator status.

	Program Enrollment						School-based Admin		
	Unstandardized Mean Differences			Standardized Effect Sizes		Power	Unstandardized		Power
Variables	Ad – As	Ad – Gr	As – Gr	Eta^2	Cohen's f^2	df(2,115)	beta	(SE)	df(2,116)
Qual	0.38	4.35	3.96	.02	.15	.29	**6.52***	(2.75)	.51
Crea	-0.25	0.87	1.12	.00	.03	.06	2.15	(3.22)	.08
Feas	0.28	**4.91***	**4.63***	.04	.20	.48	2.49	(2.36)	.13
Effct	-1.54	3.63	**5.17***	.03	.16	.32	**5.24***	(2.88)	.33
Coh	-1.62	0.95	2.56	.01	.09	.12	1.54	(2.76)	.08
thorProb	-2.96	0.03	2.99	.01	.08	.11	-1.67	(3.68)	.07
thorGoal	5.14	3.85	-1.29	.02	.15	.28	**6.27***	(3.48)	.48
thorVal	1.49	0.22	-1.27	.00	.04	.07	3.16	(2.57)	.16
thorCon	1.16	1.47	0.31	.00	.04	.07	1.86	(3.30)	.08
thorSol	0.17	**5.72***	**5.55***	.03	.19	.43	**6.93***	(2.94)	.47
thorTot	1.00	2.26	1.26	.01	.08	.11	3.51	(2.60)	.21
vignTot	0.22	2.60	2.38	.01	.11	.17	3.66	(2.46)	.25

Note. ~ = p < .10; * = p<.05; ** = p <.001.

Table 8. Final form total scores' ability to discriminate based on self-rated expertise and years professionally in schools

	Self-rated Expertise					Years In Schools				
	Unstandardized		Standardized		Power	Unstandardized		Standardized		Power
	beta	SE	beta	SE	df(2,116)	beta	SE	beta	SE	df(2,116)
Qual	**0.10***	(.05)	**0.20**	(.10)	.51	-0.01	(.13)	-0.01	(.08)	.05
Crea	0.05	(.06)	0.09	(.10)	.13	-0.09	(.18)	-0.05	(.09)	.07
Feas	**0.11***	(.04)	**0.26**	(.10)	.75	0.07	(.11)	0.05	(.08)	.07
Effct	**0.11***	(.05)	**0.20**	(.09)	.51	0.04	(.13)	0.02	(.08)	.06
Coh	0.03	(.05)	0.06	(.10)	.09	-0.07	(.13)	-0.05	(.09)	.07
thorProb	0.04	(.06)	0.06	(.10)	.09	-0.05	(.17)	-0.03	(.09)	.06
thorGoal	0.03	(.04)	0.06	(.09)	.09	-0.01	(.13)	0.00	(.08)	.05
thorVal	0.01	(.05)	0.03	(.11)	.06	-0.15	(.13)	-0.09	(.08)	.14
thorCon	-0.04	(.06)	-0.08	(.11)	.12	-0.16	(.13)	-0.09	(.07)	.13
thorSol	0.09	(.06)	0.16	(.10)	.31	0.08	(.17)	0.04	(.09)	.07
thorTot	0.02	(.04)	0.05	(.10)	.08	-0.07	(.11)	-0.05	(.08)	.07
vignTot	0.05	(.04)	0.12	(.10)	.20	-0.04	(.10)	-0.03	(.08)	.06

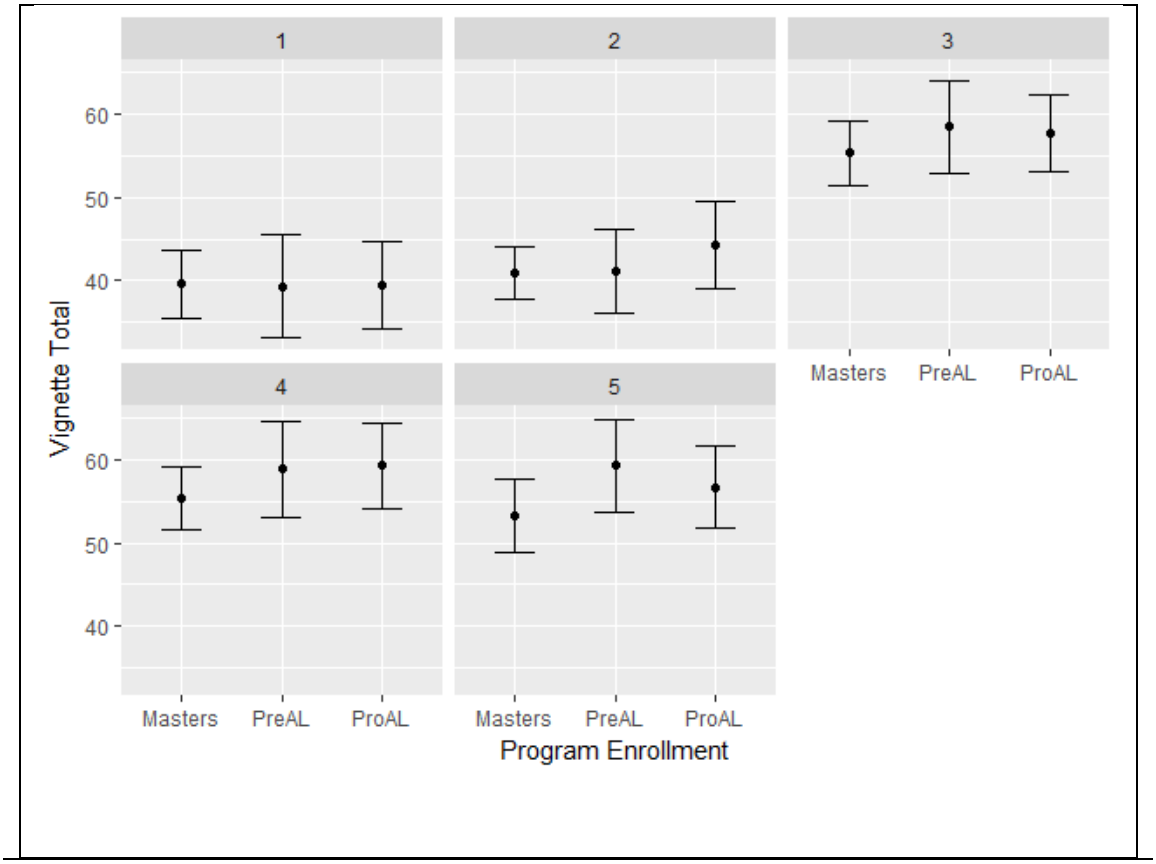
Note. ~ = p < .10; * = p < .05; ** = p < .001

APPENDIX V

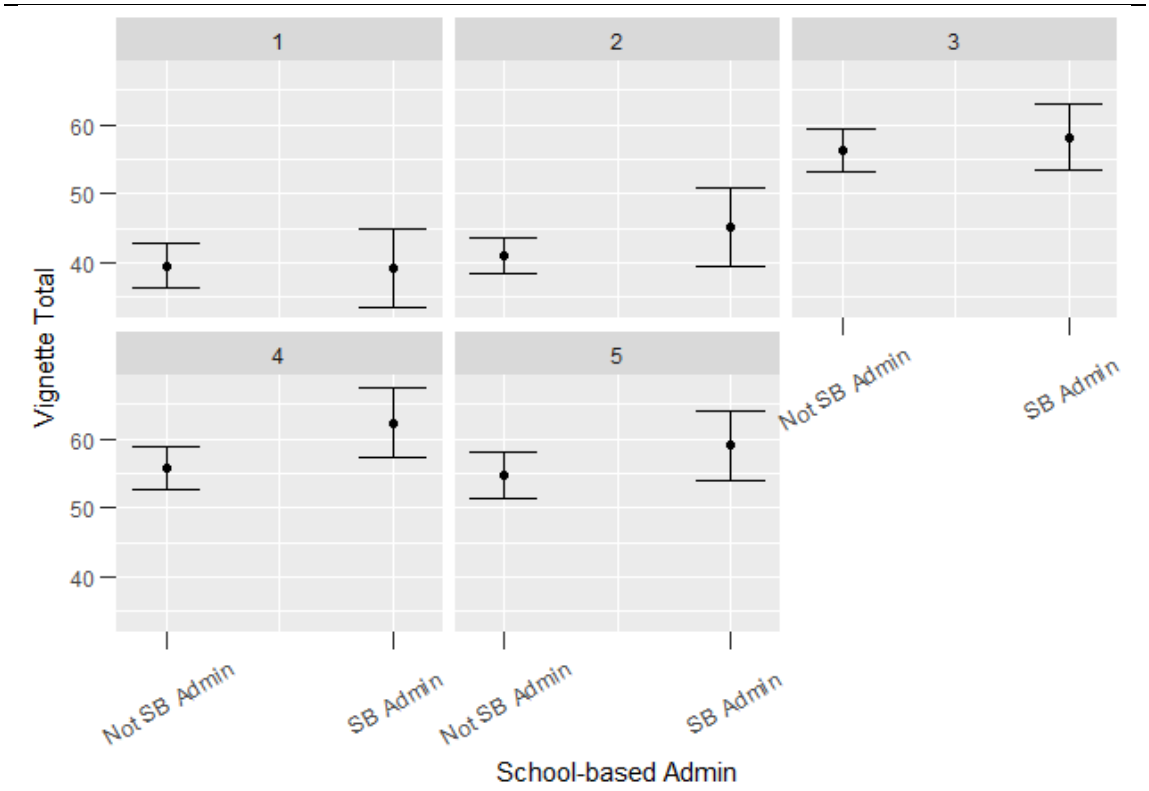
FIGURES OF THE DISCRIMINANT VALIDITY RELATIONSHIPS WITH THE CATEGORICAL PROXY VARIABLES

The Vignette Totals

Program Enrollment



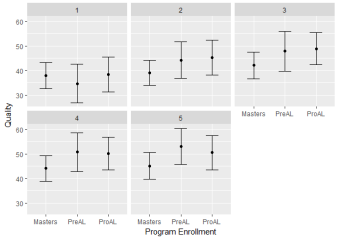
School-based Administrator Status



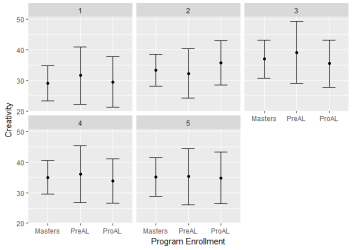
The Variables

Program Enrollment

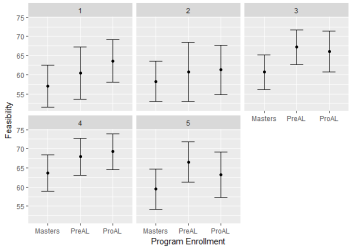
Quality



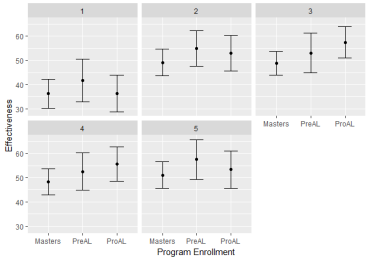
Creativity



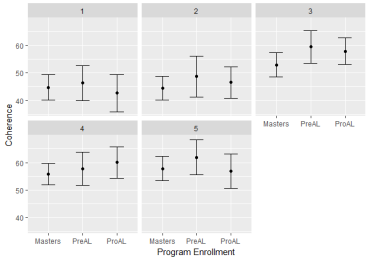
Feasibility



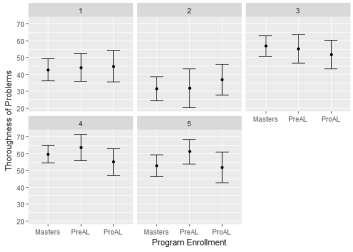
Effectiveness



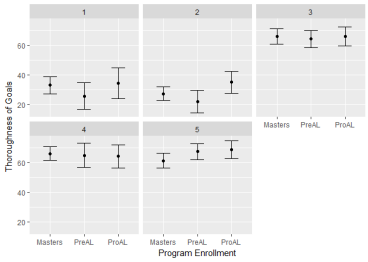
Coherence



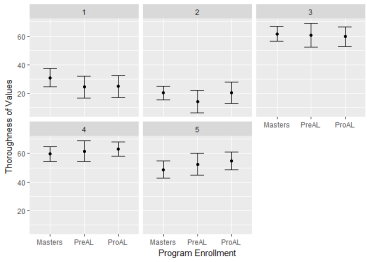
Thoroughness of Problems



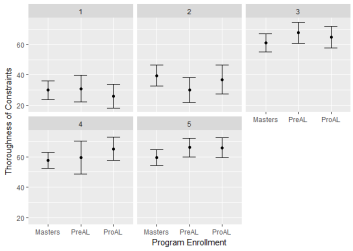
Thoroughness of Goals



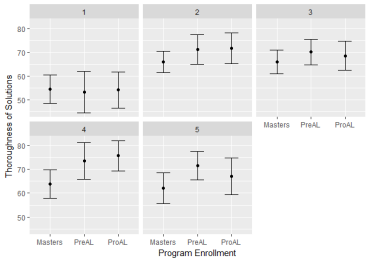
Thoroughness of Values



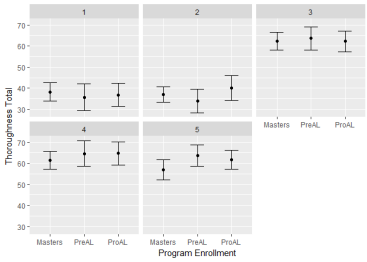
Thoroughness of Constraints



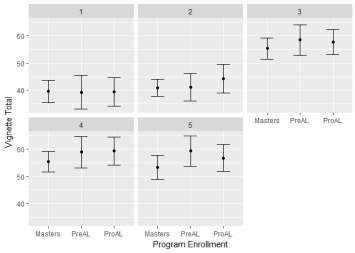
Thoroughness of Solutions



Thoroughness Total

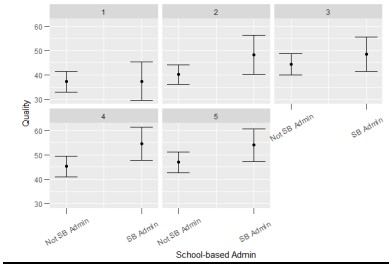


Vignette Total

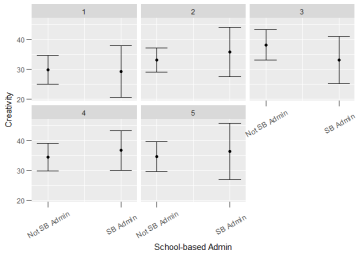


School-based Administrator Status

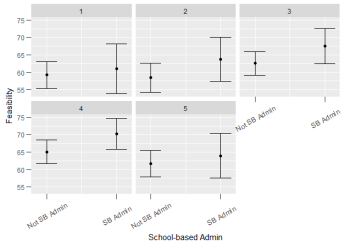
Quality



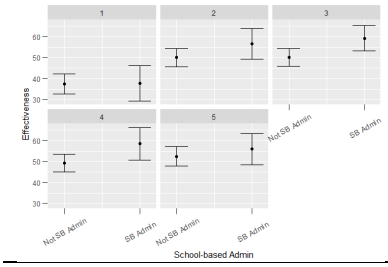
Creativity



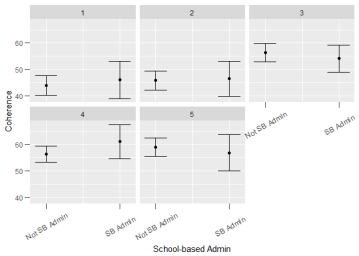
Feasibility



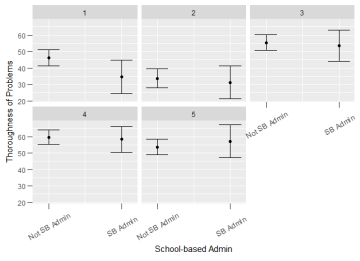
Effectiveness



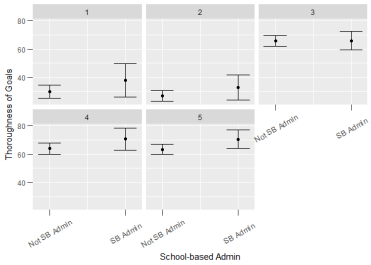
Coherence



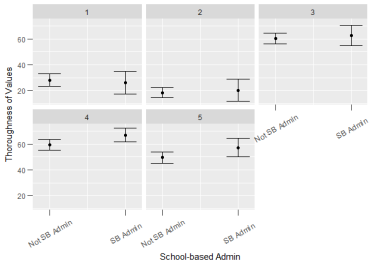
Thoroughness of Problems



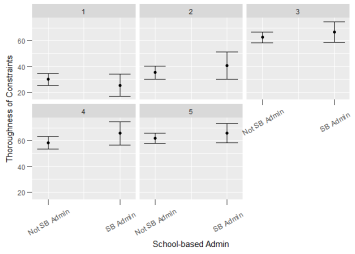
Thoroughness of Goals



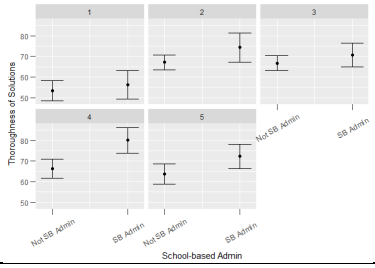
Thoroughness of Values



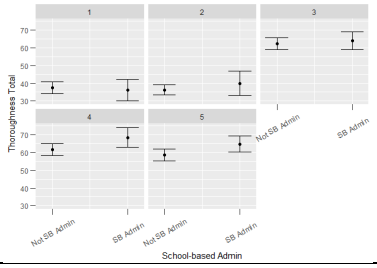
Thoroughness of Constraints



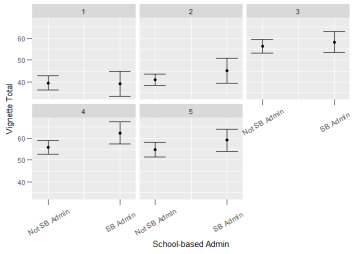
Thoroughness of Solutions



Thoroughness Total



Vignette Total

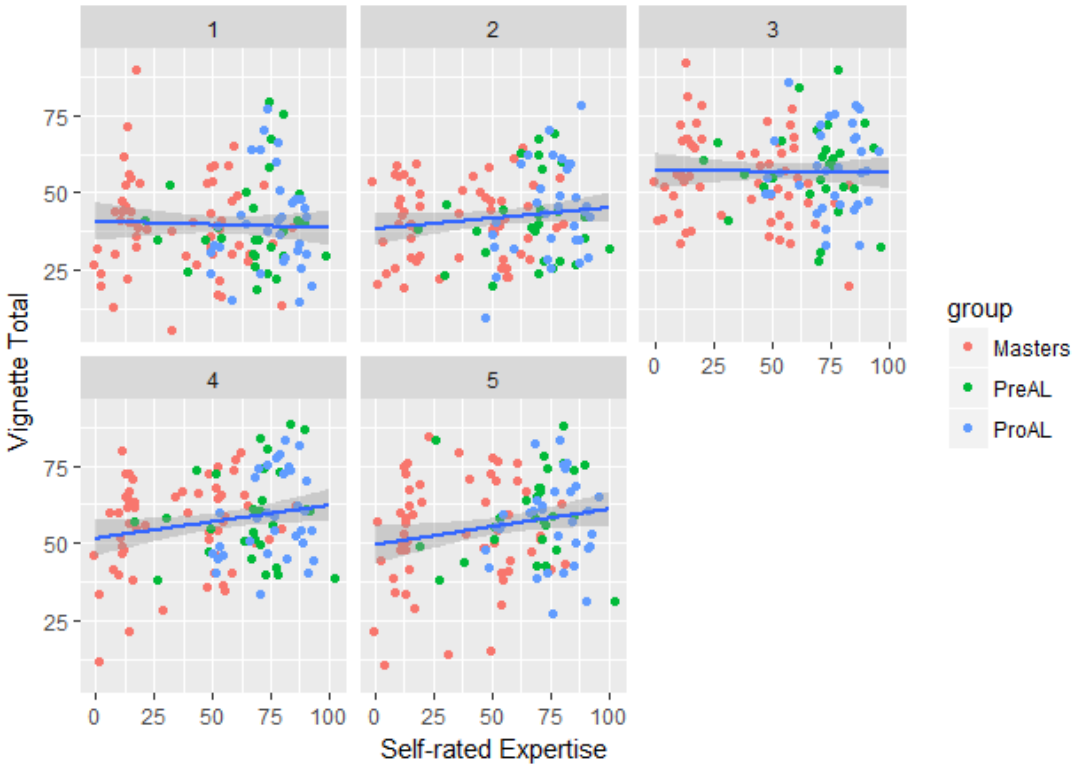


APPENDIX W

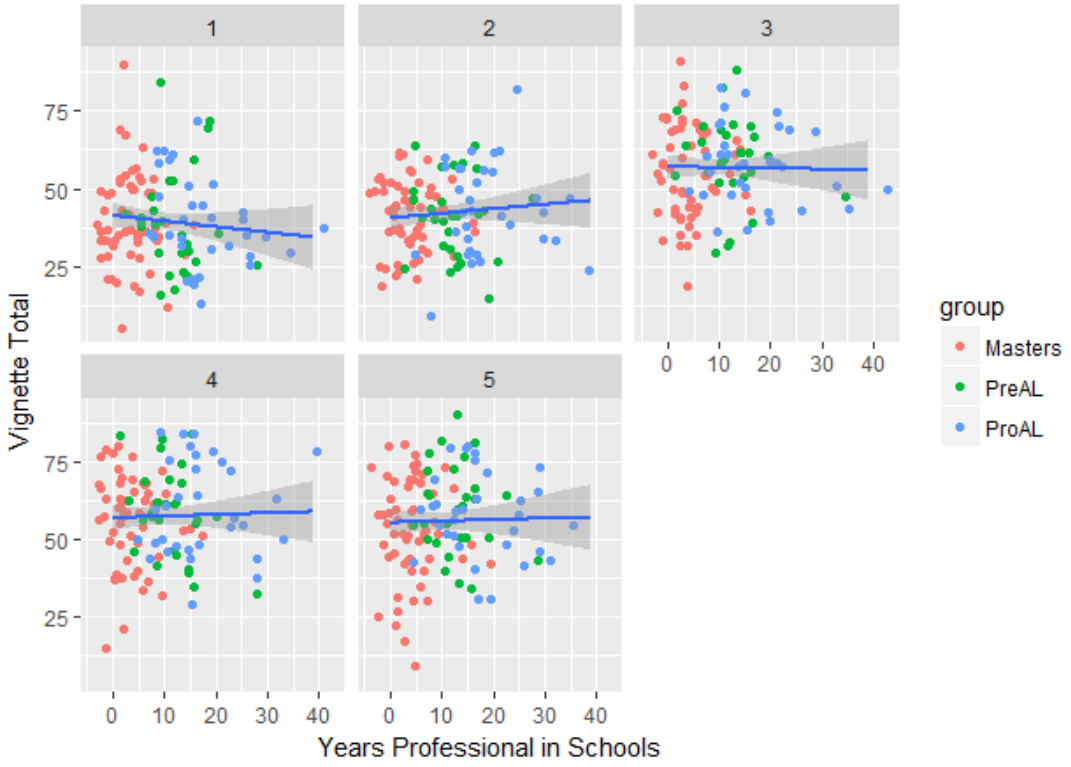
FIGURES OF THE DISCRIMINANT VALIDITY RELATIONSHIPS WITH THE CONTINUOUS PROXY VARIABLES

Vignette Totals

Self-rated Expertise

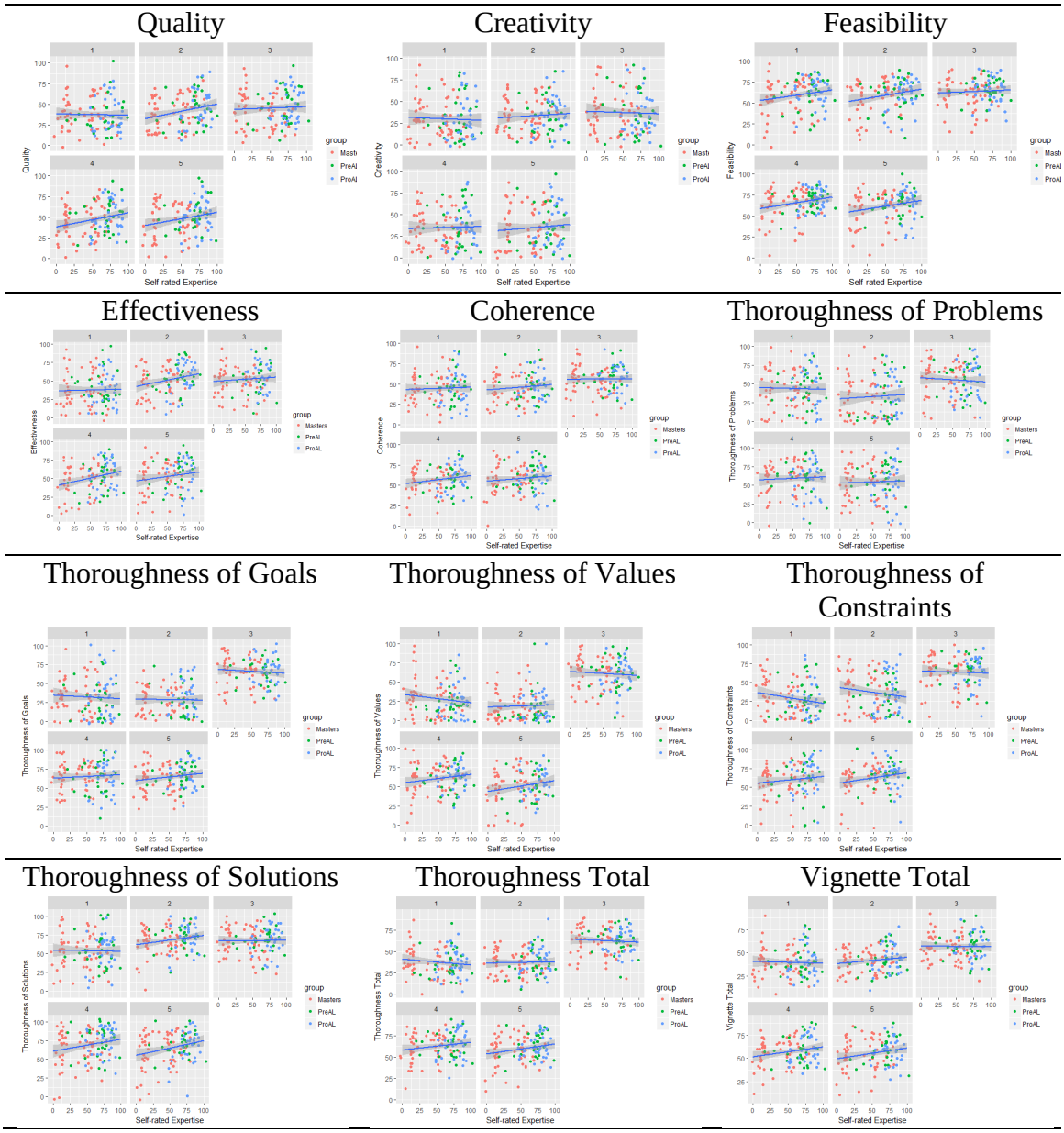


Years Professional in Schools



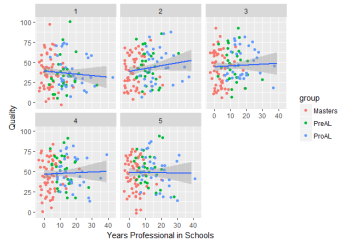
The Variables

Self-rated Expertise

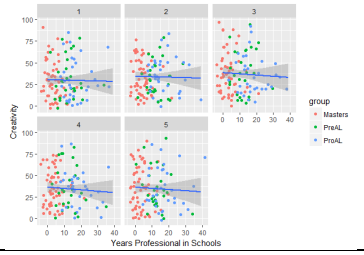


Years Professional in Schools

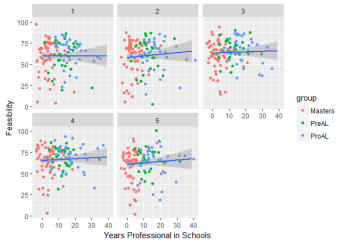
Quality



Creativity

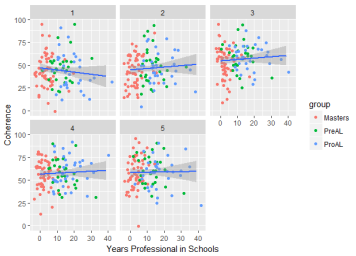


Feasibility

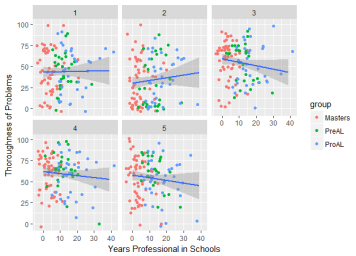


Effectiveness

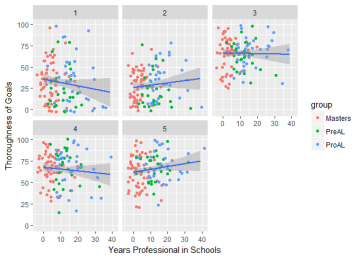
Coherence



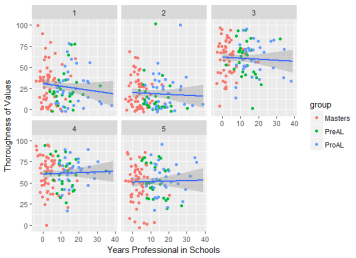
Thoroughness of Problems



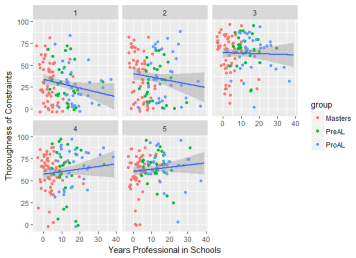
Thoroughness of Goals



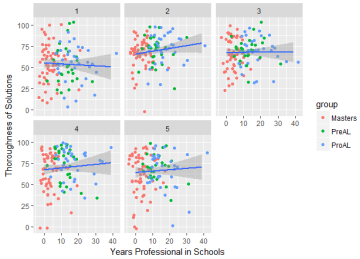
Thoroughness of Values



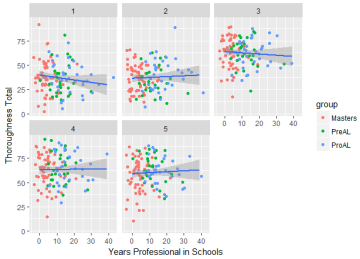
Thoroughness of Constraints



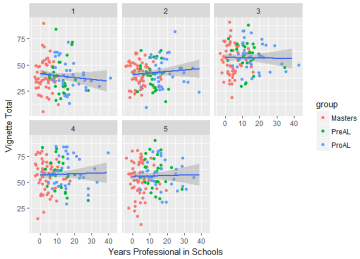
Thoroughness of Solutions



Thoroughness Total

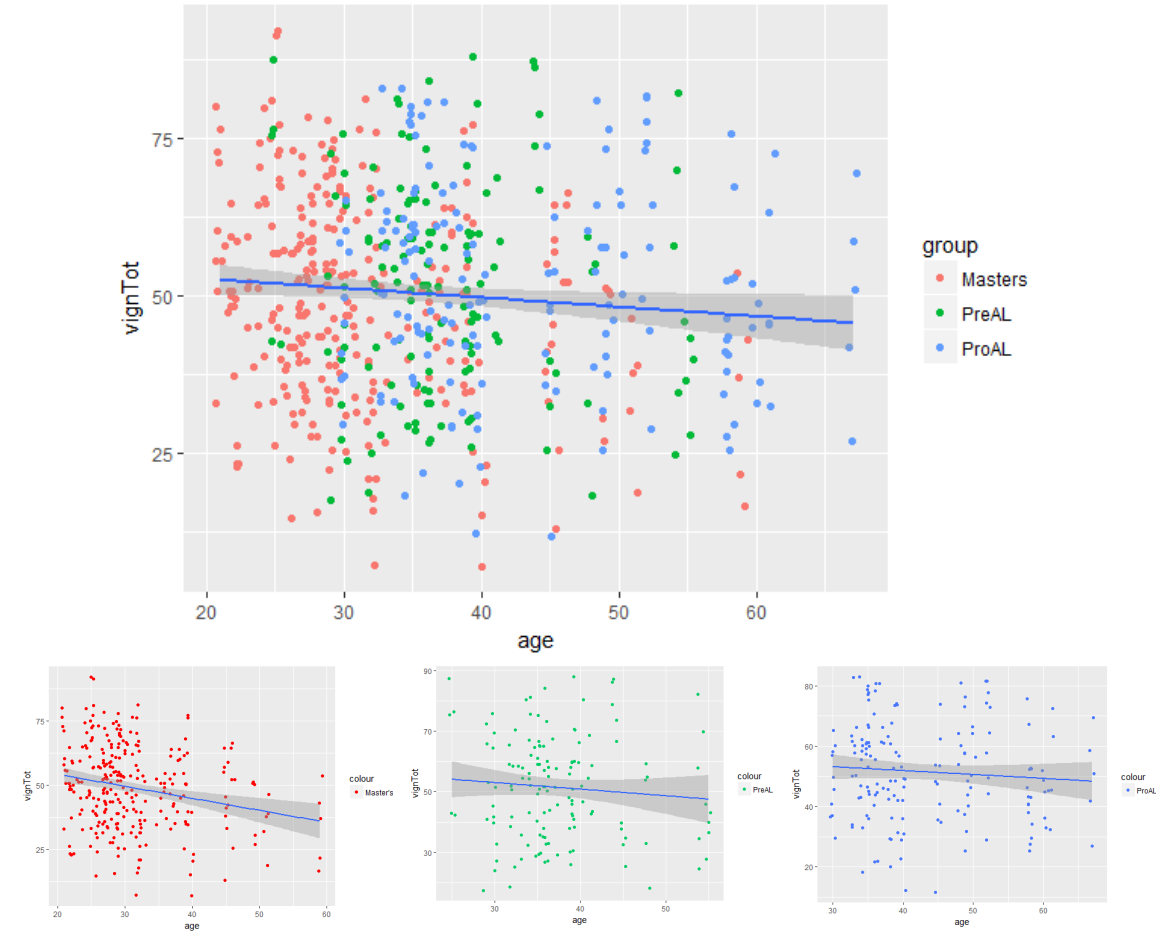


Vignette Total



APPENDIX X

RELATIONSHIP BETWEEN AGE AND VIGNETTE TOTAL



Master's	PreAL	ProAL

REFERENCES CITED

Allison, D. J. (1996). Problem Finding, Classification and Interpretation: In Search of a Theory of Administrative Problem Processing. In K. Leithwood, J. Chapman, D. Corson, P. Hallinger, & A. Hart (Eds.), *International Handbook of Educational Leadership and Administration: Part1–2* (pp. 477-549). Dordrecht, Holland: Springer Netherlands.

Allison, D. J., & Allison, P. A. (1993). Both ends of a telescope: Experience and expertise in principal problem solving. *Educational Administration Quarterly, 29*, 302-322.

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*, 997-1013. doi:10.1037/0022-3514.43.5.997

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 8*, 369-404.

Barbour, J. D. (2006). *Administration, Theories of. Encyclopedia of Educational Leadership and Administration. SAGE Publications, Inc*. Thousand Oaks, CA: SAGE Publications, Inc.

Baughman, W. A., & Mumford, M. D. (1995). Process-analytic models of creative capacities: Operations influencing the combination-and-reorganization process. *Creativity Research Journal, 8*, 37-62.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction vs. choice in cognitive measurement*. Hillsdale, New Jersey: Erlbaum.

Bennett, R. E., Jenkins, F., Persky, H., & Weiss, A. (2003). Assessing complex problem solving performances. *Assessment in Education: Principles, Policy & Practice, 10*, 347-359.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*, 77-92.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review, 94*(4), 991-1013.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*, 385-395.

Branch, G. F., Hanushek, E. A., & Rivkin, S. G. (2013). School leaders matter. *Education Next, 13*(1). Retrieved from http://libproxy.uoregon.edu/login?url=http://search.proquest.com.libproxy.uoregon.edu/docview/1238139538?accountid=14698

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement, 27*(2), 93-108.

Brenninkmeyer, L. D., & Spillane, J. P. (2008). Problem-solving processes of expert and typical school principals: A quantitative look. *School Leadership and Management, 28*, 435-468.

Bullis, M., Bull, B., Johnson, P., & Johnson, B. (1994). Identifying and assessing community-based social behavior of adolescents and young adults with EBD. *Journal of Emotional and Behavioral Disorders, 2*, 173-188.

Bullock, K., James, C., & Jamieson, I. (1995). An exploratory study of novices and experts in educational management. *Educational Management Administration & Leadership, 23*, 197-205.

Chan, T. C., & Pool, H. (2002). *Principals' Priorities versus Their Realities: Reducing the Gap*. Paper presented at the Association, American Educational Research, New Orleans, LA.

Chi, M. T., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. New York, NY: Psychology Press.

Chi, M. T. (2006). Two approaches to the study of experts' characteristics. In K. A. Ericsson, N. Charness, R. R. Hoffman, & P. J. Feltovich (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 21-30). New York, NY: Cambridge University Press.

Davis, G. A. (1966). Current status of research and theory in human problem solving. *Psychological Bulletin, 66*, 36-54.

Derr, C. B., & Gabarro, J. J. (1972). An organizational contingency theory for education. *Educational Administration Quarterly, 8*(2), 26-43.

Donaldson, L. (2001). *The Contingency Theory of Organizations*. Thousand Oaks, CA: Sage.

D'Zurilla, T. J., & Maydeu-Olivares, A. (1995). Conceptual and methodological issues in social problem-solving assessment. *Behavior Therapy, 26*, 409-432. doi:10.1016/S0005-7894(05)80091-7

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*, 215-251.

Frederiksen, N. (1983). Implications of cognitive theory for instruction in problem solving. *ETS Research Report Series, 1983*(1), 363-407.

Glasman, N. S. (1995). Generating information for the evaluation of school principals' engagement in problem solving. *Studies in Educational Evaluation, 21*, 401-410. doi:10.1016/0191-491X(95)00022-M

Goldfried, M. R., & D'Zurilla, T. J. (1969). A behavior analytic model for assessing competence. *Current Topics in Clinical and Community Psychology, 1*, 151-195.

Goldring, E., Huff, J., Spillane, J. P., & Barnes, C. (2009). Measuring the learning-centered leadership expertise of school principals. *Leadership and Policy in Schools, 8*, 197-228.

Godsil, R. D., Tropp, L., Goff, P. A., & powell, j. a. (2014). *Addressing implicit bias, racial anxiety, and stereotype threat in education and health care*. Retrieved from http://perception.org/app/uploads/2014/11/Science-of-Equality-111214_web.pdf

Hallinger, P., Leithwood, K., & Murphy, J. (1993). *Cognitive Perspectives on Educational Leadership.* . New York, NY: Teachers College Press.

Hemphill, J. K. (1958). Administration as problem-solving. In A. W. Halpin (Ed.), *Administrative theory in education* (pp. 89-118). New York, NY: Macmillan.

Hennessey, B. A., Amabile, T. M., & Mueller, J. S. (2011). Consensual Assessment. In M. A. Runco & S. R. Pritzker (Eds.), *Encyclopedia of creativity* (2nd ed., pp. 253-260). London, UK: Academic Press.

Horng, E. L., Klasik, D., & Loeb, S. (2010). Principal's time use and school effectiveness. *American Journal of Education, 116*, 491-523.

Horvath, J. A. (1999). *Tacit Knowledge in the Professions*. Mahwah, NJ: Lawrence Erlbaum.

Hoy, W. K., & Miskel, C. G. (1987). *Educational administration: Theory, research, and practice*. New York, NY: McGraw Hill.

Hoy, W. K., & Tarter, C. J. (2008). *Administrators solving the problems of practice: Decision-making concepts, cases, and consequences*. New York, NY: Pearson.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus, & Giroux.

Kennedy, M. M. (1987). Inexact Sciences: Professional education and the development of expertise. In E. Z. Rothkopf (Ed.), *Review of research in education* (pp. 133-167). Washington, D. C.: American Educational Research Association.

Klein, G., & Weitzenfeld, J. (1978). Improvement of skills for solving ill-defined problems. *Educational psychologist, 13*, 31-41.

Lazaridou, A. (2007). How effective principals think while solving problems. *International Electronic Journal for Learning in Leadership, 10*, 1-16.

Lazaridou, A. (2007). Values in principals' thinking when solving problems. *International Journal of Leadership in Education, 10*, 339-356.

Lazaridou, A. (2009). The kinds of knowledge principals use: Implications for training. *International Journal of Education Policy and Leadership, 4*, 1-15.

Leithwood, K. A. (1987). Using the principal profile to assess performance. *Educational Leadership, 45*, 63-66.

Leithwood, K., Cousins, J., & Smith, G. (1990). Principals' problem solving: Types of problems encountered. *Canadian School Executive, 9*(7), 9-12.

Leithwood, K. A., & Stager, M. (1989). Expertise in principals' problem solving. *Educational Administration Quarterly, 25*(2), 126-161.

Leithwood, K., & Steinbach, R. (1991). Indicators of transformational leadership in the everyday problem solving of school administrators. *Journal of Personnel Evaluation in Education, 4*, 221-244.

Leithwood, K., & Steinbach, R. (1992). Improving the problem-solving expertise of school administrators: Theory and practice. *Education and Urban Society, 24*, 317-345.

Leithwood, K., & Steinbach, R. (1993). Total quality leadership: Expert thinking plus transformational practice. *Journal of Personnel Evaluation in Education, 7*, 311-337.

Leithwood, K., & Steinbach, R. (1995). *Expert problem solving: Evidence from school and district leaders*. Albany, NY: SUNY Press.

Mayer, R. E. (1992). *Thinking, Problem Solving, Cognition*. New York, NY: WH Freeman/Times Books/Henry Holt & Co.

Mayer, R. E. (2013). Problem Solving. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology* (pp. 769-778). New York, NY: Oxford University Press.

Meacham, J.A., & Emont, N.C. (1989). The interpersonal basis of everyday problem solving. In J.D. Sinnott (Ed.), *Everyday problem solving: Theory and applications* (pp. 7-23). New York: Praeger.

Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

McCall, M. W., & Kaplan, R. E. (1985). *Whatever it takes: Decision makers at work.* Englewood Cliffs, NJ: Prentice-Hall.

McFall, R. M. (1982). A review and reformulation of the concept of social skills. *Behavioral Assessment, 4*, 1-33.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i-29.

Mumford, M. D., Zaccaro, S. J., Harding, F. D., Jacobs, T. O., & Fleishman, E. A. (2000). Leadership skills for a changing world: Solving complex social problems. *The Leadership Quarterly, 11*, 11-35.

Newmann, F. M., Smith, B., Allensworth, E., & Bryk, A. S. (2001). Instructional program coherence: What it is and why it should guide school improvement policy. *Educational evaluation and policy analysis, 23*, 297-321.

The Organisation for Economic Co-operation and Development (OECD; 2010). *PISA 2012 Field Trial Problem-solving Framework.* Retrieved from http://www.oecd.org.libproxy. uoregon.edu/dataoecd/8/42/46962005.pdf on 12/14/2015.

Ohde, K. L., & Murphy, J. (1993). The development of expertise: Implications for school administrators. In P. Hallinger, K. Leithwood, & J. Murphy (Eds.), *Cognitive Perspectives on Educational Leadership* (pp. 75-87). New York, NY: Teachers College Press.

O'Neil, H. F., & Schacter, J. (1997). Test specifications for problem-solving assessment (CSE Tech. Rep. No. 463). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Quellmalz, E. S. (1989). Needed: Better methods of testing higher-order thinking skills. In A. Costa (Ed.), *Developing Minds: A Resource Book for Teaching Thinking* (pp. 338). Alexandria, Virginia: Association for Supervision and Curriculum Development.

Reyes, A. (2006). Contingency theories. *Encyclopedia of Educational Leadership and Administration.* Thousand Oaks, CA: SAGE Publications, Inc.

Robinson, V., Meyer, F., Le Fevre, D., & Sinnema, C. (2015). *Leaders' Problem-Solving Capabilities: Exploring the "Quick Fix" Mentality*. Paper presented at the American Education Research Association, Chicago, Il.

Scott, T. M., & Barrett, S. B. (2004). Using staff and student time engaged in disciplinary procedures to evaluate the impact of school-wide PBS. *Journal of Positive Behavior Interventions, 6*, 21-27.

Simon, H. A. (1977). The structure of ill-structured problems. In R. S. Cohen & M. W. Wartofsky (Eds.), *Models of discovery* (pp. 304-325). Dordrecht, Holland: D. Reidel Publishing Company.

Sleegers, P., Wassink, H., Van Veen, K., & Imants, J. (2009). School leaders' problem framing: a sense-making approach to problem-solving processes of beginning school leaders. *Leadership and Policy in Schools, 8*, 152-172.

Skiba, R. J., Michael, R. S., Nardo, A. C., & Peterson, R. L. (2002). The color of discipline: Sources of racial and gender disproportionality in school punishment. *The Urban Review, 34*, 317-342.

Skiba, R. J., Horner, R. H., Chung, C.-G., Rausch, M. K., May, S. L., & Tobin, T. (2011). Race is not neutral: A national investigation of African American and Latino disproportionality in school discipline. *School Psychology Review, 40*, 85-107.

Spillane, J. P., Halverson, R., & Diamond, J. B. (2004). Towards a theory of leadership practice: A distributed perspective. *Journal of curriculum studies, 36*, 3-34.

Spillane, J. P., White, K. W., & Stephan, J. L. (2009). School principal expertise: Putting expert-aspiring principal differences in problem solving processes to the test. *Leadership and Policy in Schools, 8*(2), 128-151.

St. Germain, L., & Quinn, D. M. (2005). Investigation of tacit knowledge in principal leadership. *The Educational Forum, 70*, 75-90.

Sugrue, B. (1995). A Theory-Based Framework for Assessing Domainl-Specific Problem-Solving Ability. *Educational Measurement: Issues and Practice, 14*(3), 29-35.

Tarter, C. J., & Hoy, W. K. (1998). Toward a contingency theory of decision making. *Journal of Educational Administration, 36*, 212-228.

Todd, A., Horner, R. H., Tobin, T., Eliason, B., & Conley, K. (2013). Referral Form Definitions, version 5.0. Retrieved from www.pbis.org.

Voss, J. F., Greene, T. R., Post, T. A., & Penner, B. C. (1983). Problem-solving skill in the social sciences. *The Psychology of Learning and Motivation, 17*, 165-213.

Voss, J. F., & Post, T. A. (1988). On the solving of ill-structured problems. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The Nature of Expertise* (pp. 261-286). Hillsdale, NJ: Erlbaum.

Waters, T., Marzano, R. J., & McNulty, B. (2003). Balanced Leadership: What 30 Years of Research Tells Us about the Effect of Leadership on Student Achievement. A Working Paper.

Wiggins, G. (1989). A true test. *Phi Delta Kappan, 70*, 703-713.

Zaccaro, S. J., Mumford, M. D., Connelly, M. S., Marks, M. A., & Gilbert, J. A. (2000). Assessment of leader problem-solving capabilities. *Leadership Quarterly, 11*, 37-64.