KNOWING "HOW" IS MORE THAN KNOWING "THAT": A STUDY OF EDUCATIONAL LEADERSHIP EXPERTISE

By

Jason Taylor Huff

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Leadership and Policy Studies

May, 2009

Nashville, Tennessee

Approved:

Professor Ellen B. Goldring

Professor David S. Cordray

Professor Joseph Murphy

Professor James P. Spillane

DEDICATION

To my wife Emily: you are endlessly creative and generous, and you have loved me and shared so much to help me on this path.

To my daughter Anna and son Taylor: your shouts of "Daddy!" after a long day, drawings, ballet demonstrations, sword fights, and Lego creations have brought energy I've needed many, many times.

To Mom and Dad and Jane and Brevard: you've constantly offered support and encouragement.

You've all reminded me that there are many, many things that matter more than this.

Thank you.

I love you all very much.

ACKNOWLEDGEMENTS

I wish to thank my family for all their love and support as I've researched and written this dissertation.

My advisor Ellen Goldring has been a remarkable mentor and guide over the past four and a half years, and I thank her for all the challenges, support, trust, and opportunities that she has extended to me. Above all, her energy, focus, and discipline as a researcher are traits I hope to learn beyond all the academic training. David Cordray has offered helpful insights and guidance on the statistical analyses for this, and I am grateful. Joe Murphy frequently offered guiding comments that pushed me to look at this study's contribution to the field, and Jim Spillane never let me forget that leadership can be best understood only when we recognize its distribution throughout a school.

Finally, this research has been supported by a pre-doctoral grant from the Institute for Educational Sciences through the Department of Education. I am grateful for the ample resources that IES has provided for my academic training and my work on this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER I

INTRODUCTION

As more research highlights the important role that principals play in improving achievement in schools, there is limited work on just what expertise guides their behavior (Goldring, et al., 2009; Leithwood, 2004). While much of the recent literature has focused on principals' actions and roles and their possible connections to school conditions or student achievement (Hallinger and Heck, 1996), far fewer efforts have been made to measure exactly what principals know or need to know to enact certain practices, or how they think about what they do (Stein and Nelson, 2003). Smylie and Bennett (2005) contend that for the field "knowledge of effective leadership practices is not the same thing as knowledge of the capacities required for enactment" (p. 141). Despite substantial improvements in our understanding of effective leadership practices, the field nonetheless lags behind in its understanding of the expertise crucial to leaders' effectiveness.

Leadership expertise goes beyond knowledge of *what* a leader does to *how* to do it in a given situation (I distinguish between the two later). This marks a critical gap in the school leadership literature, because expertise informs and guides individuals' actions, and their practices in turn help shape their exeprtise—those with greater expertise typically possess more elaborate and interconnected cognitive schemata of their conditions or challenges (Borko and Shavelson, 1990). Such information in turn shapes individuals' perceptions and helps them select problem-solving strategies in a given situation. Levels of expertise differ across individuals, and people use expertise more or less

efficiently to respond to, act within, or change their environments (Allard, Graham, and Paarsalu, 1980; Anderson, Reder, and Simon, 1996 and 1997; Borko and Livingston, 1989; Lampert and Ball, 1998). As in other fields, "what administrators do depends on what they think—their overt behaviors are the result of covert thought processes" (Leithwood and Steinbach, 1995, p. 7). Principals' efforts and actions to improve student achievement, for example, are the result of the expertise they possess that guides their related leadership functions and roles. Principals' pre-existing expertise is also an integral mediating factor between their professional training and practice; what they learn and use from professional development depends in part on what prior knowledge they already possess (Bransford, 2000; Bransford and Schwartz, 1999). As researchers examine leadership practices and their influences within schools, a deeper understanding of the expertise that lies behind those practices can help guide efforts to equip individuals for leadership as well as evaluate their readiness to lead schools. Alongside analyses of effective leaders' practices it is important to explore the question, what exactly do leaders *need to know* to enact these practices?

Scholarship in the area of principal expertise has made mixed progress in two primary areas: 1) the content or domains that comprise key leadership expertise, and 2) the availability of reliable and valid instruments that measure this expertise. Recent efforts to define key areas of expertise in educational leadership emerged first from Leithwood and colleagues (1989, 1992, 1993, and 1995) who examined the problem-solving skills of "expert" versus nonexpert principals. In a second line of thinking Stein and Nelson (2003) defined school leaders' expertise as "leadership content knowledge" which consists of

knowledge of subject matter content and pedagogical content knowledge that guides how teachers best present the subject matter to promote learning, and strategies to promote teachers' improvement of their knowledge and skills. Finally, Murphy, et al., (2006), and Goldring, et al., (2009) have described yet another definition of leadership expertise by focusing on "learning-centered leadership." They argue this leadership is associated with a principal's efforts to improve teacher instruction and student achievement, and they include areas such as data-based decision making, monitoring teachers' instructional improvement, and principals of effective teaching and learning. These domains extend outside of problem-solving skills or subject matter to broader organizational knowledge that a principal uses to organize his or her school around the goal of improving teaching and learning.

While Leithwood and others have reported most extensively on research analyzing experts' problem-solving skills (in part because these efforts precede the other approaches discussed here—see for example Leithwood and Stager, 1986 and 1989 and Leithwood and Steinbach, 1989), all these lines of investigation comprise limited conceptions and measures of the expertise that school leaders employ in their work. Leithwood and Stager (1986 and 1989) and Leithwood and Steinbach (1989 and 1995) have thus far produced the most extensive work in studying principals' scenario responses and demonstrating that differences exist between expert and nonexpert leaders. The other two domains have only recently provided measures of leadership expertise. Work flowing from Stein and Nelson (2003) regarding leadership content knowledge has focused primarily on measuring two dimensions in mathematics content knowledge through survey and scenario responses of leaders: 1) knowledge of

mathematics content, and 2) beliefs about mathematics learning and teaching (Nelson, Benson, and Reed, 2004; Nelson, Goldsmith, Johnson, and Reed, 2005; Nelson, Stimpson, and Jordan, 2007). Goldring, et al., in 2008 and 2009 employed teacher and principal surveys, principal scenarios, and principal daily logs in their efforts to tap learning-centered leadership expertise. Given the importance of understanding what school leaders know and how they use this information to act within their schools, there is ample need for a deeper examination of the concepts and measures of principal expertise, particularly the relationships of these conceptions to one another.

This dissertation measures the three primary domains of principal expertise based on the lines of research cited above and examines their relationships to one another. The project summarizes these three broad domains and identifies their central subdomains (for example, the leadership content knowledge domain includes subject matter, pedagogical content knowledge, and knowledge of teachers as learners) around which specific measures are developed. I use these subdomain measures to score scenario responses from principals, and I then use the resulting scores and other methods of measurement to evaluate the measures' construct and criterion validity. The dissertation consists of three studies:

Study 1 evaluates the content validity of the proposed subdomain measures by soliciting reviews from a panel of experts.

Study 2 examines the construct validity of these subdomain measures by asking, do these measures of leadership expertise relate to each other as predicted by theory?

Study 3 explores the criterion validity of the expertise measures by asking, how do these relate to other measures of principals' expertise and practice?

This study first discusses the need for measures of school leaders' expertise followed by an examination of efforts outside the field to define and measure expertise. It next traces and critiques the literature surrounding the three primary definitions of school leader expertise to provide a conceptual grounding for the methodology and proposed measures employed in the study. Finally, the dissertation uses a three-step strategy to examine the validity of the proposed subdomain measures and their possible relationships in comprising the larger three domains of leadership expertise.

The Need for Measures of Leadership Expertise

There is substantial support from cognitive, educational leadership, and other disciplines to suggest that ideas and information influence practice. Through expertise, knowledge, memories, and perspectives leaders filter their actions; they provide the context for their decisions to act. Researchers have wrestled not so much with the issue of whether expertise matters but rather *what exact expertise* influences leadership practices, *how* it influences practice, and how to measure it. Those interested in improving school leadership therefore can learn much from studying not only what principal practices are essential but also the expertise that lies behind those practices.

As educational researchers and practitioners search for programs and/or strategies to improve school leadership, measures of expertise not only can help them identify more closely what principals need to know to improve their practices, but they can also help them trace the effects those interventions have on what principals know and learn and how they employ such expertise.

Recent criticism of principal preparation programs across the country helps to illustrate the need for measures of leadership expertise. There is widespread skepticism about the effectiveness of existing certification and development programs to prepare new principals for their work, and many have called for changes in these programs (Levine, 2005; Elmore, 2000; Hess, 2003; Tucker & Codding, 2002; Hess, 2007). They have questioned these programs' relevance to school leaders' work, their admissions standards, and their academic rigor (Hale & Moorman, 2003; Jackson & Kelley, 2002; McCarthy, 1999a; McCarthy 1999b). However, researchers have little evidence to inform these debates (Hess, 2003); there are few if any instruments to measure just what graduates learn from their programs and how they use this expertise. Much of the criticism has focused instead on descriptions of the programs themselves or the theoretical assumptions behind them (Miklos, 1983; McCarthy, 1999a) rather than on changes in what graduates know or do. Respected research on the outcomes of principal certification, training, and professional development programs is scant at best (Smylie and Bennett, 2006; McCarthy, 1999b; Copland, 2000). Evaluations of such programs have consisted primarily of participants' self-reports of the usefulness of or satisfaction with their training—measures that rely too heavily on biased perceptions and offer few insights into what participants do with the knowledge and training they receive. Few make an effort to examine programs' impacts on participants' expertise (for an exception see Copland's 2000 study of leaders' problem-framing skills). As policymakers and trainers debate just how to equip school leaders for their jobs, they lack measures of the expertise that mediates training's influence on leaders' practices,

making it all the more difficult to answer not only what to teach but what program structures are most successful in improving leaders' expertise to act.

Before discussing the specific concepts and measures of educational leadership expertise at the heart of this dissertation, it is important to review their theoretical and empirical roots to provide a context from which to evaluate the measures and analyses employed in this study. I next provide a broader survey of the literature that distinguishes between expertise and other forms of knowledge before moving to a tighter focus on the literature regarding educational leadership expertise.

CHAPTER II

PAST EFFORTS TO DEFINE AND MEASURE LEADERSHIP EXPERTISE

> *"There is no contradiction, or even paradox, in describing someone as bad at practicing what he is good at preaching." (Gilbert Ryle, 1949, p. 49)*

Conceptual Background

Research on expertise from outside education long precedes the study of educational leadership expertise. Such previous work has set the theoretical groundwork and central definitions on which educational researchers have based their studies. As I discuss later, developing constructs of expertise poses a difficult challenge empirically; these previous bodies of work not only provide guidance for such development but also important criteria by which to evaluate the measures and results of the current study. This section discusses a number of previous theorists' works to lay out the broader theoretical context in which the current work resides as references by which to evaluate this work. It begins with some of the earlier distinctions in types of knowledge that inform actions and proceeds to a more specific focus on studies that attempt to define and measure expertise.

Efforts to define what specific expertise influences practice have generated multiple constructs and definitions. As demonstrated in Ryle's quote above, researchers have often noted differences between that which we can say we know and that which actually guides our actions. The concepts of "leadership expertise" used in this study draw from previous researchers' efforts to define and measure these distinctions.

As early as 1949 Gilbert Ryle proposed differences between "knowing how and knowing that" (p. 25). He drew these distinctions as a challenge to many researchers' image of knowledge as sets of ideas and constructs that individuals simply learned, understood, and remembered—for Ryle they missed the important component of just how people *used* these ideas: "Theorists have been so preoccupied with the task of investigating the nature, the source and the credentials of the theories that we adopt that they have for the most part ignored the question what it is for someone to know how to perform tasks (p. 28)." He argued there was a great difference between someone's knowledge of certain truths or ideas and his or her ability to *do* things with that knowledge. For Ryle simple possession or understanding of an idea is no guarantee of exactly how someone might act on it.

In an effort to distinguish between these two concepts Ryle discussed just how people learned these different types of knowledge. He offered the example of someone learning to play chess. In initial games a player often reviews the rules and asks how they apply to particular situations. Over the course of many games he or she increasingly internalizes these explicit rules and less frequently recites them during the course of play. As these concepts become more automatic the player may have trouble reciting them—just as he or she becomes more proficient in using these rules so also may the player lose more explicit knowledge of them. His or her knowledge of the explicit chess rules may not indicate just how well the player can use them, and vice versa. Ryle also theorized that a player may learn the chess moves by simply watching other players; this educative process may not involve much hearing or reading of the rules. Such learning would only exaggerate the conditions in which a player can

execute the chess rules during a game without being able to recite or state them. It also demonstrated the distinctions between understanding a concept and knowing how to use it. Such knowledge does not come through traditional memorization of regulations: "we learn *how* by practice, schooled indeed by criticism and example, but often quite unaided by any lessons in the theory" (p. 41).

For Ryle knowing "how" and knowing "that" were not entirely separate or distinct, but they were quite different from each other. Knowledge of ideas such as chess game rules (knowing "that") does not guarantee that someone knows "how" to perform the ideas or use the chess rules well. Just as well, observing someone complete a particular skill (like actually playing chess) would offer evidence that he or she knows "how" to do something, but it does not insure that the person can verbally state those rules or concepts which inform the action.

While Ryle proposed and developed these initial distinctions, others have developed the ideas further. Contemporary researchers refer most often to Michael Polanyi's works published over a number of years (see 1962, 1966, and 1975) in which he developed and discussed the use of "tacit" knowledge and information that people employ but can't easily identify. For Polanyi tacit knowledge rests on "the fact that we can know more than we can tell" (1966, p. 4). This concept assumes that individuals carry more knowledge in the form of different skills, mental models, or intuitive responses than what they can describe or communicate. He offers the illustrations of riding a bike or recognizing a person's face: a person who can successfully do these things often has difficulty describing exactly how he or she accomplishes these tasks. People

acting with tacit knowledge often cannot explain the decision rules that direct their actions; with this concept "the aim of a skillful performance is achieved by the observance of a set of rules which are not known as such to the person following them" (Polanyi, 1962). A person riding a bike may not have the slightest idea how she does this, yet she can happily go on riding. Just as well, one who can explain the dynamics that make cycling possible may not be able to employ those concepts in actually riding a bike.

For Polanyi there are two components to tacit knowing, proximal and distal. We possess "proximal" information so far as we can specify what we know—in short we can identify and describe it. The physical dynamics of riding a bike that one might describe fall into this category. On the other hand, we know other types of information "only by relying on our awareness of it for attending to the second" or "distal" function: "in an act of tacit knowing we attend from something for attending to something else; namely, from the first term to the second term of the tacit relation." Polanyi returns to the acts of identifying a face or performing a physical skill. We attend *from* a person's facial characteristics to the final act of recognizing her or him. Because we only use these smaller details for the larger purpose of identifying someone we are often unable to specify that person's unique characteristics. Likewise, we rely "on our awareness of a combination of muscular acts for attending to the performance of a skill (such as riding a bike). We are attending *from* these elementary movements *to* the achievement of the joint purpose, and hence are usually unable to specify these elementary acts" (Polanyi, 1966, pages 9-10). The way in which we use knowledge influences our awareness of it along with our ability to describe or explain it; we are aware of certain proximal details only in the

appearance or performance of a larger distal term. Thus when pressed for the proximal information separate from the distal goal we often cannot identify or explain it.

Polanyi also explored briefly how we learn these different types of knowledge. For him tacit knowledge was the outcome of "active shaping of experience performed in the pursuit of knowledge" (1966, p. 16). He focused on our bodies as the "ultimate instruments" by which we take in external knowledge:

> Our own body is the only thing in the world which we normally never experience as an object, but experience always in terms of the world to which we are attending from our body…Whenever we use certain things for attending from them to other things, in the way in which we always use our own body, these things change their appearance. They appear to us now in terms of the entities to which we are attending from them, just as we feel our own body in terms of the things outside to which we are attending from our body. (1966, p. 16)

For Polanyi, true understanding for an individual would emerge as he or she applied or practiced knowledge.

More recently Richards and Busch (2005) have contributed to these distinctions in tacit knowledge by studying responses to scenarios that describe practical working situations. They propose the concept of articulable Tacit Knowledge (aTK) as compared to explicit knowledge. Explicit knowledge is a technical, academic type of knowledge that is easily described in formal language. "Explicit knowledge is technical and requires a level of academic knowledge or understanding that is gained through formal education, or structured study" (Smith, 2001, p. 315). Tacit knowledge on the other hand is "non-codified, disembodied know how that is acquired in the informal take–up of learned behavior and procedures" (Howells, 1995, pg. 2). Articulable Tacit

Knowledge is knowledge that can be "articulated for practical and competitive reasons" within an organization (Richards & Busch, pg. 1). In a number of articles (see for example Busch and Richards 2001, and Busch, Richards, and Dampney, 2001b) they present exploratory results of strategies to map such organizational knowledge.

In his ACT (Adaptive Control of Thought) theory of human cognition John Anderson has employed similar distinctions in types of knowledge by exploring interactions between what he has termed "declarative" and "procedural" knowledge (Anderson, 1983; Anderson, Reder, and Simon, 1997). He argues that "one of the key factors in human intelligence is the ability to identify and to utilize the knowledge that is relevant to a particular problem" (Anderson, 1983, p. 86) or, to paraphrase, it is not just what you know but how you use it. Cognition involves selecting what knowledge to process; a person must choose between any number of alternative ways to organize and analyze information available to him or her. Anderson argues that knowledge initially comes in "chunks or cognitive units" that "encode a set of elements in a particular relationship" (p. 23). These declarative representations or knowledge comprise units for the mind to store information; however, on their own they do not influence a person's actions.

According to Anderson's ACT theory "productions" comprise a second type of knowledge that connect declarative knowledge and behavior: "The productions themselves are not part of the fixed architecture of the system; rather, they are a second kind of knowledge that complements the declarative knowledge contained in long-term memory" (p. 215). These productions form the links between declarative concepts that in turn inform one's actions. They

comprise procedural knowledge, or knowledge about how to do something. To match these to Ryle's previous distinctions, declarative knowledge best parallels "knowing that," and procedural knowledge best parallels "knowing how."

Anderson draws deep distinctions not only in the natures of these two types but in how we learn declarative and procedural knowledge. While we may learn declarative knowledge through reading, watching, or listening to someone, procedural learning occurs only when we execute a skill; in short, we learn by doing. It is a much more gradual learning process than declarative learning. Anderson posits that we first use declarative representations of skills (e.g. an elaborate list of steps to perform under specific conditions) to practice them before our minds first "compile" this information into productions and then "proceduralize" (his terms, p. 235) these productions so that they depend less on the elaborate steps under specific conditions which we've initially learned to use the skill. This process of compilation takes much longer than simple memorization of declarative knowledge, but it ultimately speeds an individuals' ability to choose and pursue particular behaviors in various conditions.

A final major contributor to these distinctions in knowledge is Sternberg, who in work with numerous colleagues has helped to develop the concepts of "tacit" versus "explicit" knowledge. Sternberg borrows from Anderson in defining his tacit knowledge: "procedural knowledge that guides behavior but that is not readily available for introspection" (Sternberg and Horvath, 1999, p. 231). Like procedural knowledge, tacit knowledge is practical, intimately tied to our actions, and acquired primarily through experiences. It often consists of general preferences or rules of thumb for what to do under particular circumstances (Wagner, Sujan, Sujan, Rashotte, and Sternberg, 1999). Tacit

knowledge helps individuals deal with more practical problems which are often poorly formulated, in need of re-evaluation, lacking information necessary for a solution, or poorly defined (Sternberg, Wagner, Williams, and Horvath, 1995). Explicit knowledge, on the other hand, consists of those ideas and information that we can state verbally, describe, or identify as knowing. Sternberg, et al. (1995) write that while someone may be able to tell you what explicit knowledge they have by describing various concepts or steps, tacit knowledge must often be *inferred* from an individual's statements or actions (p. 916). It is also tied to particular uses or conditions, unlike explicit knowledge, which is often nonspecific to a particular use. Finally, tacit knowledge is usually acquired on one's own or with minimal support (for example, much of this type of knowledge may be obtained through "on the job experience" rather than formal classroom instruction).

Sternberg and others have presented research to argue that tacit knowledge inventories are distinct from traditional intelligence tests. For example, Wagner and Sternberg (1985) found only small correlations (.16, $p>.05$) between tacit knowledge and verbal reasoning tests administered to undergraduates. Eddy (1988) reported small correlations between a tacit knowledge test and scores on the Armed Services Vocational Aptitude Battery results. However, when Wagner and Sternberg (1985) looked at the relationships of tacit knowledge scores to merit-based salary increases and average performance ratings they found correlations ranging from .48 to . 56 ($p<.05$). With such findings they have asserted that the tacit knowledge inventories do indeed measure another form of knowledge beyond that tapped in standard

intelligence tests, and that these measures may tap a construct more relevant to job performance than traditionally defined intelligence.

Expertise As A Form of Knowledge

At the core of the works discussed thus far lies a distinction between types of knowledge that are more or less relevant to practice. The study of "expertise" has borrowed heavily from these distinctions in an effort to identify holders of the knowledge that influences behavior and actions. Ericsson, Charness, Feltovich, and Hoffman's (2006) definitions of "expert" and "expertise" illustrate the conceptual overlaps with the work just reviewed. They describe an expert as

> one who is very skillful and well-informed in some special field…someone widely recognized as a reliable source of knowledge, technique, or skill whose judgment is accorded authority and status by the public or his or her peers. Experts have prolonged or intense experience through practice and education in a particular field. (p. 3)

Expertise refers to "the characteristics, skills, and knowledge that distinguish experts from novice and less experienced people" (p. 3). Implicit in these definitions are the tacit or procedural knowledge discussed above: experts possess a larger amount of more practical, field-specific knowledge that enables them to complete their work skillfully. They not only possess specialized knowledge and skills, but they can also use or apply them successfully in their work. If works from Anderson and Sternberg and others have drawn distinctions between the types of knowledge we possess and use, research on expertise has also worked to identify the holders of such specialized knowledge. Primary foci in research on expertise have been not only to identify and measure what exactly these skills and knowledge are but also to identify those people who possess them.

A significant body of work in expertise has also targeted the different "cognitive strategies" that experts possess or use (Brenninkmeyer and Spillane, 2008). Chi, Feltovich, and Glaser (1981) and Glaser and Chi (1988) offer some of the most recent examples of this; they examined differences in individuals' problem categorizations and representations (1981) and problem-solving skills (1988). In these cases they found differences between experts and novices in how they 1) conceptualized and analyzed physics problems and their solutions, and 2) represented and evaluated problems they faced and then retrieved relevant memories for use in a solution (1988). The authors reported that experts not only looked at problems in significantly different ways but they also accessed information more efficiently as they solved a problem. Chi, et al. (1981) argued the findings suggested that experts were able to see underlying conceptual similarities between problems while novices perceived differences based on surface features in the problems. The two respondent groups also differed in their prescriptions for how to solve the problems; experts used their perceptions of deeper principles in the problems to recommend solutions while novices paid more attention to the surface problems. The authors offered these findings as evidence that experts "have a great deal of tacit knowledge that can be used to make inferences and derivations from the situation described by the problem statement" (p. 149). Borrowing from Anderson's terms they concluded that "declarative knowledge contained in the [experts'] schema generates potential problem configurations and conditions of applicability for procedures," while "procedural knowledge in the schema generates potential solution methods that can be used on the problem" (p. 150). Thus experts reference a greater amount of declarative knowledge that in turn helps them choose more practical solutions.

Just as the works summarized here have increasingly distinguished between types of knowledge that are more and less relevant to practice, so also has the educational leadership research on expertise focused on the practical knowledge that guides leaders' actions. In Chapter III, I connect this previous research to the three primary areas of expertise that I use in this dissertation. While Leithwood and Steinbach's (1995) work provides some of the clearest connections, researchers in the other two areas also stress the importance of understanding the specific expertise that informs school leaders' practices.

Before discussing these ties, however, I first review those methods from outside the field of educational leadership that have been used to measure expertise. Just as previous research has offered a conceptual base for educational researchers to build upon, so also has it provided a guide to the methods most helpful in capturing educational leadership expertise. The next section provides a context from which to understand the methods of measurement I employ in this study.

Methods of Measurement in the Study of Expertise

Here I focus on methods that have most influenced expertise research in educational leadership. While Anderson's work has focused on developing computer simulations that mimic actual cognitive processes, and it has informed the discussion of different types of knowledge, his research offers limited insight into exactly how to measure expertise. Chi and Glaser, et al., and Sternberg, et al., have wrestled more directly with this question of measurement, and I consider their work before tracing its connections to studies in educational leadership.

In their studies of tacit knowledge, Sternberg and colleagues used instruments that fell into one of three categories: 1) respondents rated a series of possible responses to a work-related situation (Wagner and Sternberg, 1985), 2) participants rated different pre-set action statements according to how well they described their own work behavior, and 3) participants wrote plans of action to describe how they would respond to complex, open-ended work scenarios (Sternberg, et al., 1995).

Scoring varied across these instruments. With the first type of instrument respondents' ratings were compared to their group membership (e.g. experienced manager, business school student, undergraduate) to determine relationships that might exist between the ratings and group membership (see Wagner & Sternberg, 1985). Researchers later comprised a profile of responses from a group of nominated "experts" and then compared them to participants' selection of different responses (Wagner, 1987). In this case they again sought relationships between scores and participants' expertise designations. Finally, in Wagner, Rashotte, and Sternberg's (1992) study of tacit knowledge of sales the authors generated "rules of thumb" from interviews and industry texts that summarized broad principles to distinguish between expert and novice reactions (e.g. "in evaluating your success think in terms of tasks accomplished rather than hours spent working"). They then used these as pre-set responses to scenarios describing actual working conditions, and participants had to prioritize the different responses in deciding how to address each situation.

In Wagner and Sternberg (1985) and Wagner (1987) the researchers found significant differences in the scenario responses according to group membership (business manager, business graduate student, and undergraduate). Differences

here included individuals' application of strategies to job-related conditions such as improving customer relations or navigating sales agreements. Later, Williams and Sternberg (unpublished) scored responses to work scenarios and were able to identify differences in individuals' tacit knowledge according to membership in low-, middle-, and upper-management (for example, middle managers demonstrated significantly greater knowledge of working effectively within the work environment than lower-level supervisors, and upper managers showed greater knowledge in influencing and controlling others). These findings offered support for differences in the practical or tacit knowledge that individuals at different levels of experience and expertise possess.

Chi and Glaser and others focused on differences in two types of cognitive strategies, knowledge representation and problem-solving skills. Chi, et al. (1981) first asked undergraduate (novice) and graduate (expert) students in physics to categorize different physics problems according to similarities between the solutions and then explain their reasons for the groupings and the categories they chose. Next they chose problems in which "surface structures" (objects or literal physics terms described in the texts) roughly crossed with "deep structures" (the major physics concepts applicable to the solution of a problem) (p. 125). They found that novices did indeed categorize problems according to these surface structures while expert students grouped the same problems by their deep structures. They reported significant differences in the problem categorization and structures employed by the two groups—while novices paid more attention to the conditions described in the problems, experts considered the physical concepts at play in the problems.

These methods of measurement share a number of characteristics that have influenced measures of expertise in the field of educational leadership (these will be discussed later). First, the measures all place participants in practical situations. As much as possible, all the measures create conditions that respondents might actually encounter. Wagner and Sternberg (1985) Wagner et al. (1999) and Richards and Busch (2005) created numerous scenarios from real-life situations to which participants had to respond. Second, they require respondents to *use* rather than simply state certain knowledge. For example with Chi, et al. (1981) participants did not simply describe what they knew about the physics concepts but rather how they would categorize and solve the problems. Likewise Wagner et al. (1999) asked participants to rank or explain responses to actual work conditions instead of identifying different concepts they knew about work. Finally, these analyses examined responses to identify differences according to individuals' varied levels of experience and expertise. As I will discuss later, various efforts to measure educational leadership expertise have employed these same strategies. As one evaluates the measures used within educational leadership research (including those that I propose here) this literature provides valuable examples against which to compare them.

This review also helps to highlight questions that persist in the research of expertise. For example, while these studies purport to measure expertise that influences individuals' behavior, little of the work above has demonstrated empirically if expertise relates to specific practices or actions. For example, Wagner and Sternberg (1985) related differences in scenario responses only to group membership, not the work-related actions taken by the participants. There is much room for researchers to investigate whether and how expertise is related

to *actual* practices. In addition, many of these measures lack more rigorous validation efforts through such construct validation strategies as examining their relationship to other constructs, or through criterion validity analyses to compare the results of more than one method used to measure the same constructs. In this project I discuss how the findings relate not only to the specific field of educational leadership expertise but also to these issues in other fields of expertise.

CHAPTER III

DOMAINS AND MEASURES OF EDUCATIONAL LEADERSHIP EXPERTISE

As summarized in the introduction, for this dissertation I focus on three primary conceptions of leadership expertise: expert problem-solving skills as developed by Leithwood and Stager (1986 and 1989) and Leithwood and Steinbach (1995), leadership content knowledge as proposed by Stein and Nelson (2003), and learning-centered leadership as developed by Murphy, et al. (2006) and Goldring, et al. (2007 and 2009). Like the work reviewed in the previous sections, studies in these areas all attempt to identify and measure the practical expertise that school leaders employ in their work. Leithwood and Steinbach's (1995) definition of expertise demonstrates the connections to Ryle's, Anderson's, and Sternberg's works:

> a) the possession of complex knowledge and skill,
> b) its reliable application in actions intended to accomplish generally endorsed goal states, and
> c) a record of goal accomplishment as judged by others in the field (p. 13).

Just as Ericcson, et al. (2006) focused on practical knowledge and its use, Leithwood and Steinbach (1995) examined leaders' possession of practical knowledge as well as their capacity to apply it successfully. The studies in this section also use measures similar in structure to those I have just summarized. In this section I first explain my selection of the three areas of expertise that I include in this study—I defend why I have focused on these particular domains over others. I then provide a theoretical review of each domain of educational leadership expertise along with its measures before critiquing these efforts and

discussing how the measures I employ in this study make a contribution in light of these limitations. For each of the three main domains of expertise I also summarize some of the key theoretical components or *subdomains* that comprise them. For example, Stein and Nelson (2003) lay out three key areas of expertise that comprise "leadership content knowledge": subject matter, pedagogical content knowledge, and a knowledge of teachers as learners. The measures that I use in this dissertation are based in part on these subdomains, and I explain these in more detail for each area of expertise in Chapter 5.

Selection of Domains of Expertise For This Study

The three domains of expertise that I have included in this study are by no means an exhaustive group. There are multiple other forms of expertise that guide principals' practices (such communication, collaboration, problem-framing, human relationships, and building and resource management). Such a breadth necessitates a prioritization of particular areas to focus this study, and I explain my selection of three primary domains in this section.

The two main conversations that guided my selection of domains originated largely with the effective schools research in the 1970's and 80's. The domains in this study focus on a) expertise in problem-solving, and b) expertise in instructional leadership. Multiple studies from this research reported that effective schools often contained strong leaders who were integral to their successful organization and who focused on rigorous teaching and learning. As researchers identified conditions that were key to effective schools and leaders they often broadened these findings into recommendations for other practitioners to heed. However, this literature often ignored the more complex

contexts of schools and how leaders addressed these conditions (Hallinger & Murphy, 1986b; Hallinger, Leithwood, & Murphy, 1993). Such omissions generated prescriptions for leaders' practices that were limited both practically and theoretically, and Hallinger, et al. (1993) and others doubted that such research on leaders' actions could generate adequate understandings of how successful leaders really succeeded: "…studies of principal behaviors and practices would never provide the type of information needed to understand how leaders adapted to the complex contexts in which they worked. Such understanding would result only from investigations that incorporated explorations of the thinking that accompanied such practices or behaviors" (p. xiii). In short, they argued that only research that examined both what leaders *do* as well as *how they think about what they do* would yield more valuable insights for researchers and practitioners alike.

Building on cognitive work in other areas of management, researchers began to examine what cognitive processes and skills school leaders used to understand and address the conditions they faced. Two areas that received particular attention consisted of problem-solving skills (Leithwood, et al., 1986, 1989, 1993, and 1995) and problem-framing strategies (Bolman & Deal, 1993). These initiatives were some of the first to respond to criticism of the more prescriptive research for effective leaders, and they stepped away from education-specific expertise to understand how leaders made sense of and responded to their complex environments. A number of the studies were able to identify specific differences between expert and non-expert principals (for example, see Leithwood & Stager, 1989). These findings illustrated the

importance of understanding leaders' problem-solving skills, and this evidence led me to include this area of expertise in the study.

However, while such studies helped to understand some of the more complex expertise that leaders used to analyze their environments, they did not address the more "content-specific" questions that the effective schools research raised for leaders. The effective schools studies also identified effective school leaders as those more actively focused on improving teaching and learning in their schools, and this call for "instructional leaders" was the second consideration that guided my selection of domains of expertise. Multiple studies from the research described effective leaders as those "hip-deep in curriculum and instruction and unafraid of working directly with teachers on the improvement of teaching and learning" (Hallinger, 2005, p. 224; see also Murphy, 2006). This interest in instructional leadership has only intensified with the recent focus on performance standards; Hallinger writes that principals now find themselves "at the nexus of accountability and school improvement with an increasingly explicit expectation that they will function as 'instructional leaders'" (2005, p. 222). Thus recent research and the current policy environment have prioritized such expertise and skill sets for principals. As research has increasingly identified it as key to effective schools, the policy focus on student achievement has called for greater principal engagement in improving teaching and learning.

The question remains, however: just what is expertise in instructional leadership? What exactly do successful principals need to know? Multiple definitions and responsibilities have been associated with this concept (e.g. Hallinger & Murphy, 1985; Bossert, et al., 1982; Leithwood, Begley, & Cousins,

1990; Stein & Nelson, 2003). A review of two differing models demonstrates how this instructional leadership includes much more than expertise in curriculum and instruction. Stein & Nelson (2003) have recently advocated for principals to have a deep understanding of the curriculum, pedagogical content knowledge, and professional development strategies for their teachers' different subject areas: "…as demands increase for them to improve teaching and learning in their schools, administrators must be able to know strong instruction when they see it, to encourage it when they don't, and to set the conditions for continuous academic learning among their professional staffs" (2003, p. 424). Their work argues for expertise deeply anchored in teachers' specific content and subject matter. This contrasts with previous work from Hallinger and Murphy (1985) who identified the following broader dimensions of instructional leadership: a) defining the school's mission, b) managing the instructional program, and c) promoting a positive school climate. Such factors included expertise in curriculum and pedagogy along with other areas such as establishing a mission statement to guide efforts within the school, organizing different programs and resources according to this mission, and evaluating teachers' instruction and students' academic progress (Hallinger, 2001; Hallinger & Murphy, 1985). According to this second model, "instructional leadership" can entail not only engaging in and analyzing curriculum and instruction but also analyzing and organizing broader school conditions that support improved teaching and learning. This means that principals must not only understand what good teaching and learning look like in the classroom, but they must also understand how to analyze and align a school's components as a larger whole according to these goals. These differing lines of research illustrate how complex a concept

like instructional leadership can be to define and how much researchers (and practitioners) can differ in what expertise they believe school leaders need to succeed.

My selection of three domains of expertise for this study derives primarily from the above studies of principals in effective schools. As researchers have examined the expertise and knowledge that informs these leaders' practices, they have used definitions that are more "content-rich" (such as Stein & Nelson's (2003) "leadership content knowledge"), "content-independent" (such as problem-solving expertise (Leithwood, et al., 1986, 1989, 1993, and 1995) and problem-framing strategies (Bolman & Deal, 1993)), and more organizational in scope (such as Hallinger & Murphy, 1985). In an effort to examine how principals differ in expertise as defined across this spectrum, I have chosen three bodies of work that best represent these different areas.

Leithwood, et al., (1986, 1989, 1993, and 1995) have thus far produced the most extensive research on content-independent expertise. This work has not only developed key definitions for problems-solving, but it has demonstrated significant differences in these areas for expert and non-expert principals. Finally, others such as Brenninkmeyer and Spillane (2008) have published results that support their findings.

Murphy, Elliott, Goldring, and Porter (2006) as well as with Goldring, Spillane, Huff, Barnes, and Supovitz (2008) have offered the most extensive definitions for a broader form of expertise with "learning-centered leadership." These pieces include knowledge of curriculum and teaching along with more organizational perspectives such as data-based decision making and standards-

based reform. Goldring, et al. (2008 and 2009) have presented initial findings about these levels of expertise.

Stein and Nelson's (2003) "leadership content knowledge" represents the most content-rich definition for expertise of successful leaders. They advocate for principals' deeper understanding of the content and pedagogical strategies that teachers use in their classrooms, and Nelson and Sassi (2005) and Nelson, Goldsmith, Johnson, and Reed (2005) have presented work that examines the nature of this content knowledge.

In the next sections I elaborate on each of these domains by first explaining their different definitions and then discussing existing measures for each of them.

The Domain of Problem-Solving Expertise

In the introductory summary to their 1995 book "Expert Problem Solving: Evidence from Schools and District Leaders" Leithwood and Steinbach argue (as in the introduction to this paper) that "what principals do depends on what they think" (p. 7)—their actions as leaders result from the cognitive processes they employ to analyze conditions and formulate effective responses to them. For Leithwood and colleagues, previous studies from the "effective schools" movement called for effective principals in schools but offered only limited examinations of and prescriptions for particular behaviors (Bloomberg and Greenfield, 1980; Martin & Willower, 1981). Leithwood and Steinbach (1995), for example, argue that much of this existing research from inner-city, low socioeconomic status, small elementary schools implicitly assumed there was little variation in principals' context—such studies often over-generalized reports

of principal behaviors outside of these limited contexts. They called for a look behind the behaviors at the decisions and processes that drive their practices.

Leithwood and Stager (1989) contend that too many of the effective principal descriptions left the administrator's mind as a conceptual "black box" (p. 127) with little understanding of how leaders chose their actions and reactions within school environments. A cognitive approach to principals' problem-solving processes, they propose, would offer a look at the mental strategies and decisions that drove these practices—mental processes that were more consistent than contextually contingent practices from previous work. Leithwood and Steinbach (1995) assert that the first dimensions of leadership effectiveness to measure are not necessarily a principal's external actions to address a situation but rather the cognitive processes that inform his or her actions. While principals' actions may be highly contingent on their school contexts and conditions, their problem solving strategies are more likely dependent on what expertise they possess (Leithwood & Stager, 1995). Only when we understand these cognitive dimensions of problem-solving, they argue, can we provide leaders with the learning and thinking skills to respond to the unique conditions of their schools instead of prescribing a static list of behaviors that often cannot account for the complexity and variety of administrators' worlds.

Leithwood and Stager (1989) cite Schon (1987) in explaining the cognitive foundations of their work:

> When practitioners respond to the indeterminate zones of practice by holding a reflective conversation with the materials of their situations, they remake a part of their practice world and thereby reveal the usual tacit processes of worldmaking that underlie all their practice (p. 6).

To study the different cognitive strategies that administrators use in their work, Leithwood and Stager employ Schon's "indeterminate zones of practice" to create practical "unstructured problems" to which participants responded. They theorize that such open-ended situations would prompt school leaders to demonstrate what tacit knowledge they used in their reactions to work conditions. They choose to focus on respondents' problem-solving skills because while principals routinely solve fairly well-structured problems they must also address much more complex conditions in which they first need to identify the problem and its stakes before pursuing a resolution. Thus using only well-structured problems (with clear solutions) would also limit the generalizability of their findings to only those more simplistic conditions in which administrators operate. With their unstructured problems Leithwood and Stager tap not only the surface decisions that principals make but also the way they analyze the conditions and framed possible solutions.

Colleagues found that, compared to nonexpert principals, expert principals

1. are better able to regulate their problem-solving processes through reflection,
2. possess more information relevant to the problem and are able to access it more quickly and extend it to new situations,
3. recognize patterns in problems faster and sense deeper themes or concepts in the problems they encounter (as discussed with Chi, et. al, above),
4. identify and possess more complex goals for problem solving and goals related to action,
5. spend more time in their initial response planning overall strategies and are more flexible planners while addressing the actual problem,
6. are more sensitive to the social contexts in which problems are solved (Leithwood & Stager, 1989, p. 130).
7.

Measures of Problem-Solving Expertise

Thus far Leithwood and various colleagues have provided the bulk of the research measuring expert problem solving skills for school leaders. Leithwood and Stager (1986) first asked elementary school principals using a "think-aloud protocol"[1] interview to perform a problem-sorting task in which they selected which problems they would solve alone, with one person, or with a group. They then asked them how they would prioritize the problems they would address or leave alone. Finally, they asked respondents to discuss how they would address each problem based on their own experiences. They found differences between principal responses to these problems and used the findings in the problems they used in their next study (reported in 1989).

In their 1989 study they presented principals with brief hypothetical case problems ranging from resource staff decisions to program evaluation strategies and asked them to 1) rank the problems according to how clear to them the course of action was, 2) to present with as much detail as possible their solutions to the most and least clear problems they ranked, and 3) describe from their own experience situations that had similar degrees of clarity. They first identified a series of grounded categories that captured all statements in the responses. With these categories they examined differences between individuals across scenarios of different clarity (Leithwood & Stager, 1989, p. 133). Through qualitative and quantitative analyses they found differences between novice and expert principals.[2] Most notably, those problems that principals identified as the least

[1] All of the studies summarized for Leithwood, et al., used this protocol in which participants verbally shared their responses and researchers taped the discussions.

[2] Leithwood and Stager used two criteria to identify experts. First they asked two central office administrators to identify those principals from their sample group who were experts. They only used those

structured generated the greatest differences between experts and nonexpert principals. Examples of differences included the following.

1. Experts perceived difficult problems as manageable with careful thinking, while typical principals viewed difficult problems as stressful or frightening.
2. Experts discussed the need to collect information before solving a problem while nonexperts made assumptions in lieu of collecting information.
3. Experts focused on implications for students and program quality while typical principals more often mentioned staff-oriented goals.
4. Expert principals saw few constraints to solutions while nonexperts discussed multiple constraints.
5. Experts emphasized detailed prior planning while nonexperts (1989, p. 139) gave little attention to prior planning.

These findings have influenced not only the measures used by other researchers (for example, Brenninkmeyer and Spillane (2008) and Goldring, et al. (2009) used low-structure problems) but also the terms for which researchers analyzed differences between expert and nonexpert prinicipals (e.g. Brenninkmeyer and Spillane (2008) examine principal responses to scenarios for evidence of prior planning and strategies to collect information before addressing a situation).

Leithwood and others extended these same types of measures and analyses to other groups (for example Leithwood and Steinbach focused on secondary school principals in 1990 and on superintendents in 1991). Other researchers took these measures of low-structured problems and notions of "expert" versus "nonexpert" principals into additional areas. Most notably Johnson (2003) developed a list of problem-solving components for use in conflict scenarios, and Brenninkmeyer and Spillane (2008) and Brenninkmeyer and Weitz White (2005) created coding definitions to analyze instructional

---

individuals identified by both administrators. Second, they interviewed all those who passed the first screen using a framework of principal growth from *The Principal Profile* (Leithwood & Montgomery, 1986).

scenarios (Leithwood and Steinbach's 1990 and 1991 prompts consisted of primarily administrative conditions). Brenninkmeyer and Spillane (2008) and Brenninkmeyer and Weitz White (2005) presented principals with six scenarios during one-hour interviews and asked them to respond; these scenarios focused on instructional conditions that included two subject areas (math and reading—see examples in Appendix B). Like Leithwood they argued that "ill-structured" problems were well-suited to examine cognitive strategies because they required a respondent "to structure the problem before being able to answer it, as well as come up with a solution based on that structure" (Brenninkmeyer & Spillane, 2008, p. 444). They analyzed principal responses using a coding scheme that closely followed the qualitative findings from Leithwood and Stager's (1989) piece as well as findings reported in Leithwood, Steinbach, and Raun (1995). They scored responses to the problems according to codes similar to work by Leithwood, et al. (for example, identification of constraints, facing conflict, focusing on student program quality versus staff goals, and gathering information and data before pursuing a solution). Brenninkmeyer and Spillane (2008) found differences between novices and experts in certain areas such as their identification of constraints, their discussion of planning before pursuing a solution, and their discussion of delegating responsibilities to others.

The Domain of Leadership Content Knowledge

Borrowing from Shulman's (1986) work on pedagogical content knowledge which distinguished between teachers' knowledge of subject areas such as math or history and the knowledge to help students learn those subjects, Stein and Nelson (2003) and colleagues set out to define and understand the

knowledge of subject matter and teaching that leaders need to know to be effective instructional leaders. They argue that as supporters and evaluators of teaching in their schools, administrators comprise key leverage points through which to improve instruction systemically (Stein and Nelson, 2003, p. 425). Their paper elaborates on how this expertise relates to school district leaders and those in charge of principal supervision or professional development, but I focus here on their discussion of what principals in particular need to know for their work. They briefly define their concept as "that knowledge of academic subjects that is used by administrators when they function as instructional leaders" (p. 423). Stein and D'Amico (2000) had previously provided a longer elaboration and purpose for this concept:

> In order to provide intellectual leadership for instruction, principals and superintendents must understand the manner in which classroom practices and curricular programming differ in mathematics vs. literacy, as well as the different needs that teachers have with respect to each subject area. Only then will they be able to wisely select among the plethora of professional development programs, to evaluate the quality of instructional programs and practices, to validly select and interpret the results of student assessments, and to steer building-wide reforms that span various grades" (p. 10).

Acknowledging that a school principal cannot be expected to know as much about a subject (or its pedagogical content knowledge) as teachers or specialists in different subject areas, they nonetheless argue that school administrators need to understand to some degree the different subject areas taught in their schools and their respective differences. For example, "school mathematics is comprised of a definable body of knowledge, a structure of interrelated concepts, a symbol system, and vocabulary…derived from the discipline of mathematics" while "literacy has much less of a delineated knowledge base through which to proceed" (p. 4)—such subject differences influence discussions on what

curricula, the specific content teachers choose to use in their classrooms, and strategies for best presenting the materials for students to learn. In addition, Stein and Nelson (2003) also argue that school leaders must also understand how teachers themselves best learn subject matter and the skills with which to teach them.

Leadership content knowledge stands at an intersection of subject matter knowledge and knowledge of leadership practices--principals' knowledge of subjects can inform their efforts as instructional leaders (e.g. it may influence details on which they focus in observing and evaluating a math versus a reading teacher), and their actions to provide such leadership may transform the subject matter knowledge they use (e.g. principals may have to distill this knowledge from a more complex set of ideas about a subject into a list of areas on which their teachers must focus to pursue professional development in their subject areas).

Stein and Nelson (2003) divide this knowledge as a whole into two different categories: knowledge of the substance (or *what* the work is about) and knowledge of *how* to facilitate the learning at different levels (classroom, school, and district). Leaders at progressively higher levels of administration need to know something about the learning required for each level they supervise. For principals specifically, knowledge of the substance involves three areas: 1) the subject matter that is taught in their schools' classrooms, 2) pedagogical content knowledge that helps to explain how students learn different subjects, and 3) an understanding of teachers as learners and effective ways to teach teachers or support their learning (p. 426). Their knowledge of how to promote learning focuses on 1) understanding the learning needs of individuals (in this case, their

teachers), 2) arranging conditions such as professional development which provide appropriate expertise and tasks to promote teacher learning, 3) offering incentives to motivate individuals to learn, and 4) providing adequate resources to support the learning (p. 424).

To summarize, principals not only need to know differences in subject matter and understand how students learn different subjects, they also need to know how *their teachers* learn different subjects and their pedagogies best. Issues relating to this last aspect may include a teacher's past training in a subject, his/her views and strategies of how to teach the subject, and how best to provide support for that individual to learn new subject matter and/or pedagogical strategies on his or her own or through cooperation with colleagues. For Stein and Nelson, all of these components are integral to a principal's provision of instructional leadership: only when a principal is able to employ knowledge of a subject matter and its pedagogical content to guide teachers in learning new information and strategies to improve their craft can he or she help to improve the instruction that is vital to raising student achievement.

Measures of Leadership Content Knowledge

Stein and Nelson offered case study examples of the concepts they proposed, but Nelson and colleagues in subsequent work have developed a number of measures to evaluate leaders' content knowledge in mathematics. The bulk of Nelson and her colleagues' work (primarily Nelson, et al., 2003; Nelson, Benson, and Reed, 2004; Nelson, Goldsmith, Johnson, and Reed, 2005; and Nelson and Sassi, 2005) focuses on two subdomains of leadership content knowledge in mathematics: 1) knowledge of the subject itself and 2) beliefs

about mathematics learning and teaching. Two of these pieces demonstrate best the measures they have developed and employed.

Nelson, et al. (2004) write that "mathematics knowledge is quite complex, consisting of a strong conceptual understanding of mathematical ideas interwoven with knowledge of algorithms, mastery of computational procedure and mathematical facts, and mathematical 'habits of mind,' or ways of approaching mathematical problems, including skill at choosing representations for numerical situations, mathematical reasoning, problem-solving, and proof" (p. 3)." Given the complexity of this knowledge and its use in teaching, Nelson et al. (2004) used a nontraditional measure of mathematics knowledge from the Study for Instructional Improvement (SII) (Ball, Hill, and Bass, 2002; Hill, Schilling, and Ball, 2003) in their study of 14 elementary school principals. This collection of items was designed to measure mathematics "knowledge for teaching"—that content knowledge which is specific to the math that elementary teachers teach and use as opposed to math used in engineering, statistics, accounting, or other fields. These items evaluate an individual's understanding of the methods, concepts, and problem-solving strategies used in elementary mathematics, and they frequently use a student's learning as the context for the question. For example, one item describes that the respondent is working with a class on multiplication and notices a specific pattern in the way many of them are displaying their work. Thus the content and the design of the questions focus on tapping an individual's understanding of math concepts that she or he will use frequently in an elementary math class. In their 2004 piece Nelson, et al. found relative differences in the mathematical knowledge of their participants, and the SII allowed them to identify particular relationships between principals'

understanding of different math content areas. For example, while principals were often able to correctly compute answers to problems they had more difficulty identifying the concepts under the problems. Thus Nelson and her colleagues concluded that simple ability to compute an answer did not always mean a principal would understand the conceptual foundations for those problems. They argued that the SII measures they employed were important in distinguishing between computational and conceptual understandings.

With their mathematics epistemology instrument these researchers also focused on measuring principals' beliefs about math teaching and learning. It consisted of two parts: 1) a section of nineteen likert items that asked participants to rate how much they agreed or disagreed with statements about math teaching and learning, and 2) an open response section with a scenario that describes interactions during a math lesson between the teacher and her students (p. 22). These components allowed researchers to rate how much principals agreed with a "constructivist" approach to teaching and learning math versus a "direct instruction" approach and how these beliefs related to their analyses of classroom conditions. For example, they found that most principals in their sample agreed with constructivist instructional methods to support learning, but there was less agreement about just how mathematical concepts were learned. They also found that most of the principals in their sample focused on behavioral, surface-level features in a lesson plan (such as the manipulatives or activities used or the format of discussion) rather than how these different aspects of teaching support students' efforts to understand and employ the ideas in the lesson (p. 23, 25).

Nelson, et al., (2005) modified these measures before employing them in a broader study of 96 elementary and middle school principals. They continued to use the SII measure of Mathematics Content Knowledge as well as their epistemology instrument with the survey and the scenario, but they also employed an additional piece to this instrument. This component consisted of five statements from fictitious teachers about their teaching philosophies of math instruction. Respondents had to recognize and order these statements from most traditional to most "reform-based" (p. 8). This additional piece provided the researchers with another perspective into principals' beliefs about teaching mathematics.

While Nelson and colleagues have tightly focused their efforts on mathematics subject matter and pedagogical beliefs this work provides helpful guidance for the measures used in this dissertation that tap subject matter expertise in mathematics and/or reading. I develop measures for subject matter and pedagogical content knowledge as well as the additional subdomain of "knowledge of teachers as learners" as discussed in Stein and Nelson (2003).

The Domain of Learning-Centered Leadership

A final dimension of leadership expertise is the content of school leadership for propelling student learning, often referred to as "learning-centered leadership" (Murphy et al., 2006). This includes expertise in areas such as standards-based reform, monitoring instruction for improvement, data-based decision-making and others--knowledge not isolated to any specific subject matter taught in schools but essential for leaders to improve teacher instruction and student achievement in their schools (see Murphy, et al., 2006; Goldring and

Berends, 2008; Leithwood and Jantzi, 2005; Eubanks and Levine, 1983; Heck, 1992). With this expertise a principal examines the larger school conditions and their alignment according to specific goals and strategies. Learning-centered leadership expertise steps beyond subject matter content and problem-solving skills to encompass the broader organizational expertise that a leader possesses and employs to organize a school around the goal of improving instruction and student achievement. Such expertise relates to conditions not only in the classroom (e.g. effective teaching strategies) but also in the school as a whole (e.g. the process for establishing a school-wide vision) (Murphy, et al., 2006). Because of a principal's unique position to influence multiple areas of a school, measures of his or her expertise must encompass actions across the broader organization.

Goldring, Spillane, Huff, Barnes, and Supovitz (2009) argue that even with subject matter knowledge, it is important to measure and understand what principals *do with* this knowledge to focus the school organization on improved instruction and learning--the question remains "*what are the mechanism, or how do* school leaders work to establish the communities of practice that can impact school climate, instructional organization and ultimately student learning" (p. 29). Only by looking at these areas of principal expertise can we understand the knowledge that informs principals' efforts to lead their schools. Murphy, et al. (2006) noted the slim body of empirical work on the areas of expertise that comprise learning-centered leadership. They also commented that findings in relevant literature were uneven, with more robust findings coming from those dimensions that influenced "the most powerful variables in the equation of student learning (e.g., quality instruction, curriculum alignment)" (p. 8).

Murphy, et al. (2006) proposed eight major dimensions or subdomains:

vision for learning, instructional program, curricular program, assessment program, communities of learning, resource acquisition and use, organizational culture, and social advocacy. Goldring, et al. (2009) offered a similar set of subdomains (e.g. standards-based reform, data-based decision making, coaching, professional development, school learning environment), though they have focused on only a few in their exploratory analyses thus far (standards-based reform, data-based decision-making and monitoring classroom instruction in 2006, along with effective teaching and learning in 2008). These areas cut across subject areas and grade levels as principals use such expertise to examine how well their organizations align according to broader school goals. For example, with "standards-based reform" expertise principals analyze curriculum changes and teaching initiatives to examine whether or not they align with the broader standards for their schools. Leaders use their "data-based decision making" expertise to evaluate student achievement *throughout* their schools and identify those issues that need the most attention. Finally, they "monitor classroom instruction" to determine if new instructional strategies support the school's larger instructional improvement efforts. Unlike leadership content knowledge, this larger domain of expertise focuses on principals' understanding of and attention to the school's organization as a larger whole.

Measures of Learning-Centered Leadership

Efforts to measure learning-centered leadership have thus far proceeded along two primary fronts. First, Murphy, Goldring, Porter, Elliott, and Cravens have written multiple pieces in their efforts to develop an assessment of leadership for school principals (see 2006 and 2007). They focused this

assessment around a theory of action that effective leadership consists of core components enacted through key processes (Goldring, et al., 2007), but these measures focus ultimately on principals' behaviors as reported through surveys of their staffs. Their 2006 model includes knowledge and skills as precursors to leaders' actions and practices, but they do not provide measures for the domains of expertise that inform principal behaviors.

The more direct measures of expertise in this area have come from Goldring, et al. (2009) in a separate study of principal professional development. Measures of principal expertise in learning-centered leadership have consisted of principal and teacher surveys about principals' expertise and practices and scenarios that present hypothetical situations to which principals must respond. I summarize each of those here.

First, the principal survey items were based on a revised and adapted version of *The School Leadership Self Inventory* (National Policy Board for Educational Administration, 2000), a self-reporting inventory consisting of Likert scale items based on the ISLLC standards for school leadership. These items asked principals to self-report the extent to which they possessed "personal mastery (knowledge and understanding)" of different areas related to domains within learning-centered leadership (such as data-based decision making, standards-based reform, and monitoring instructional improvement).

Next Goldring and her colleagues designed a set of six scenarios to which principals responded by reporting how they would address the conditions in each vignette. The scenarios were modeled after Leithwood and Stager's (1989) scenarios and Brenninkmeyer, Sherin, and Spillane's (2004) scenarios. They were all designed to be ill-structured problems to take advantage of Leithwood and

Stager's (1989) finding that ill-structured problems differentiated experts from nonexpert administrators. The team designed the scenarios to be as open as possible to increase the opportunities for the principals to detail the expertise that they might use in addressing the question posed in the scenario (these will be described in more depth later in this piece). In their analyses of the responses they first developed a set of definitions for the domains of learning-centered leadership. They then applied these definitions in two different analyses. In their 2006 paper, they scored responses according to the frequency with which principals mentioned different domains. Scorers simply marked those sections of the text where principals employed particular domains in their answers. Scores were awarded to respondents according to the number of times they had mentioned a particular domain in their answers. In their 2008 paper, they scored the responses according to the "quality of response" a principal demonstrated. For a principal to score higher with an answer, he or she had to mention a particular aspect and then demonstrate a deeper understanding of it through a longer discussion (see Appendix D for examples; these contain the final scenario scoring rubrics of the four subdomains of analysis used for learning-centered leadership). Thus, while they counted mere mentions of the topics, only deeper discussions of the different aspects actually scored higher on these rubrics (I will discuss these analyses in more detail in the methods section that follows).

Goldring, et al., have presented two papers that examine the results and validity issues of these different measures (2008 and 2009). Correlations between the frequency and quality of response scores of the scenarios are significant but vary in size across domains (results to be published). While the two different scoring strategies overlap in some of what they capture, it appears that the

second method is indeed tapping more than just frequency of mention. Furthermore, scoring results for both the simple frequency of mention and the quality of response analyses of the scenarios have shown only limited correlations with the principal and teacher surveys. Goldring, et al. (2008) have begun to question if the scenarios and self-report meaures are really measuring different constructs. For example, perhaps the scenario method is more dependent on a respondent's written communication skills than on his or her leadership expertise, or maybe the self-report measures are not reliable as each principal has a different metric as to what he or she considers expertise. The authors question if the scenario quality of response scores may tap more tacit, practical knowledge while the principal survey may measure school leaders' more declarative knowledge as far as what principals say that they know.

A Critique of Existing Measures

While the works cited in this section provide the empirical framework to guide this study, each of them contains limitations on which the field can improve. Here I offer a short critique of the lines of research summarized above and discuss how various features of this dissertation have addressed some of these gaps.

While work by Leithwood and Steinbach (see 1989, 1991, and 1993b for examples) has provided the foundation for much recent research on school leaders' problem-solving expertise (e.g. Brenninkmeyer & Spillane, 2008), their work has nonetheless relied heavily on principals' self-reports as provided in interviews. As Corrigan (1995) writes, such "responses are open to question because what principals say they will do and what they actually do in practice

may be quite different" (p. 650). Beyond Leithwood and Steinbach's (1993b) use of staff surveys, Leithwood and others have employed few other methods to capture problem-solving expertise. Such reliance on principal self-reports has left researchers with few opportunities to validate their expertise measures by examining their relationships to other variables. Furthermore, these measures of expertise have thus far relied primarily on identifying binary differences in principal responses (for example, do principals discuss planning strategies or not?). These methods leave open the question of whether or not principals also differ in the degree to which they possess differing levels of expertise, and what those different variations might be.

This study has addressed these issues in a number of ways. First, the scenario scoring rubrics at the center of this research focused on capturing differences in the quality of principals' responses, not simply the presence or absence of expertise in what they write. In contrast to Leithwood and Steinbach's (1993) scoring, these protocols helped identify more nuanced differences in principals' levels of expertise. For example I scored principals' level of expertise in data-based decision making on a score of 0 (no mention) to 5 (the principal provides at least two more extensive discussions of a concept and links those discussions conceptually). I also provided both numeric scores and qualitative examples to demonstrate differences between individuals—such information helps to show more effectively just *how* leaders vary in their levels of expertise (for example, what is the qualitative difference in a response that scores a "2" versus a "3"?).

On the other hand, Nelson, et al. (2004 and 2005) have developed measures that capture varying degrees of leadership content knowledge (as

discussed before, their mathematics epistemology instrument in the 2005 piece identified principals' theoretical beliefs for mathematics instruction on a spectrum ranging from "constructivist" to "teacher-centered"). However, their measures thus far focus exclusively on mathematics subject matter and pedagogical content knowledge.

In this dissertation I used measures that relate to literacy and reading/language arts. I also proposed a measure with which to capture principals' expertise in "teachers as learners" (how to direct or help facilitate teachers' professional development), heretofore an area that Nelson and Stein (2003) propose as a third domain of leadership content knowledge for principals but have not yet measured.

Finally, none of the studies discussed in this section has examined extensively how different measures of expertise relate to one another. These limitations in the literature derive in part from the origins of the different bodies of research. While Leithwood and colleagues used an emergent design to identify key problem-solving strategies that leaders possessed and employed in responding to situations, the literature for "leadership content knowledge" and "learning-centered leadership" consist primarily of pieces that have *prescribed* what types of expertise successful school leaders need to have. Limited research has been done to examine whether or not leaders actually possess these different types of expertise, and at what different levels.

Does educational leadership expertise consist of theoretically distinct areas (or domains) as defined in the three areas I have reviewed, or do they possess expertise across these areas? Do principals more often possess expertise in just one of the domains summarized above (but not in others) to suggest that

these areas are separate? Or do principals' scores across the different domains indicate that these relationships are more complex? I used the scores from the different scoring rubrics to examine the possible relationships among domains of expertise.

Implications of this Study for Distributed Expertise

After discussing the theoretical bases for expertise one further qualification for this study helps to locate it in the literature. While much of the work cited in this review has thus far focused on individuals' expertise, more recent studies have examined expertise as it is distributed across individuals within an organization. This research builds upon concepts such as Simon's (1975) "bounded rationality" that emphasizes individuals' cognitive limitations, and it stresses the need to understand just how people share their skills and expertise when working together. Supporters of situated learning (Greeno, 1989; Brown, Collins, & Duguid, 1989, for example) have similarly argued that investigations of cognition must account for both individual actions as well as interpersonal social interactions. Pea (1993) writes that distributed intelligence is "commonly socially constructed, through collaborative efforts toward shared objective or by dialogues and challenges brought about by differences in persons' perspectives" (p. 48). For Pea such distributed views of cognition, knowledge and expertise stand "in sharp contrast to the common focus on 'intelligence' as an attribute of individuals, carried primarily in internal transformations of mental representations, of symbols for goals, objects, and relations" (p. 49).

Particular attention to "distributed expertise" has more recently come from business management and organizational circles as companies consider

how best to use and protect their employees' expertise (Prahalad and Hamel, 1990; Nanda, 1991; Nonaka, 1991). Business ventures such as "knowledge management" (as described in Smith, 2001; and McCune, 1999) view expertise and other resources as inherently stretched across the multiple employees in an organization.

In education circles, widening views of leadership in schools have prompted researchers to look beyond the principal to identify additional key sources of instructional leadership. While leadership studies still focused on the individual principal through the mid 1980's (Bridges, 1982), since then researchers have begun to look also to teachers and external change agents as additional sources of guidance for schools (Spillane, Halverson, and Diamond, 2001; Camburn, Rowan, Taylor, 2003). Definitions of "distributed leadership" differ (for example, Firestone and Corbett, 1988, and Firestone, 1989 defined such leadership according to individuals' various organizational functions while Spillane, et al., 2001, offered a more practice-based definition), but these broader views of school leadership have nonetheless pushed the question about whose expertise it is important to measure. Today researchers certainly recognize the importance of examining more than just the principal's leadership expertise.

And yet this broader scope of study by educational researchers cannot overlook the key roles that principals play in their schools and the importance of understanding just what expertise informs their practice. Ample research still points to the vital influence that principals can have on conditions in their schools both directly and indirectly (Dwyer, 1985; Hallinger and Murphy, 1986; Hallinger and Heck, 1996a and b; Leithwood and Jantzi, 1999; Griffith, 2003).

More recent research also offers evidence that principals provide different forms of leadership than others in their schools. Camburn, Rowan, and Taylor (2003) found that in implementing comprehensive school reforms (CSR's) principals tended to engage in higher, more general levels of leadership than other leadership team members:

> We see that after controlling for all of the other variables in the analysis, principals generally report engaging in higher levels of leadership…than incumbents in any other position…Though they are members of a team, principals…clearly stand out. On average, they are generalists, performing a broader range of leadership functions than other leaders, and usually at higher levels (p. 366).

Other leadership team members such as CSR coaches focused more on such areas as instructional leadership or developing instructional capacity, but overall other team members engaged less in boundary spanning or broader management activities. Camburn, et al.'s findings not only underline principals' important roles in their teams, but they point to the unique leadership functions that principals fill in their schools. These differences show up even when comparing their practice to those of other leadership team members. Such differences emphasize the importance of examining principals' individual expertise to understand what informs their practice.

Based on these more recent findings this study has focused on principals' individual expertise while recognizing that many others play leadership roles in schools and possess expertise crucial to guiding a school. Principals' unique roles and practice in schools justifies a close look at what expertise they bring to their positions, but such an examination in no way captures the expertise of all important school leaders. Analyses of other members' expertise or how expertise is distributed throughout schools are beyond the scope of this study; however,

such studies are essential to understanding the full resources available to staff and how different members share such resources.

While this dissertation focuses on measuring the expertise of individual school principals, the concepts and measures developed here can certainly be used to measure the leadership expertise in other school members as well. In the final conclusion I discuss further what implications the results have for conceptualizing and measuring leadership expertise that is distributed across different roles and practices in schools.

CHAPTER IV

METHODOLOGY

Overall Objectives

Cronbach (1971) describes validation as the process by which a test developer or user collects evidence to support the types of inferences that are drawn from test scores. One must scrutinize the definitions and scoring criteria a test uses—it has validity only to the extent that it measures what it purports to measure. If a test measures leadership content knowledge, it is imperative to examine not only the consistency of the scores it provides but whether or not the measure uses plausible, accurate definitions of such expertise. There are multiple strategies by which researchers compile such evidence to defend their tests as accurate measures of the constructs they purport to tap, and I used three of these strategies in this dissertation.

Researchers of the three domains and measures of educational leadership expertise I have reviewed thus far offer arguments that the measures define and capture the constructs of expertise they describe. Accordingly, measures of these three domains of expertise should discriminate between principals based on their different levels of expertise—they should be able to generate scores that represent meaningful differences in individual school leaders' expertise. However, because I proposed new measures of expertise in this study I could not rely on the previous validation studies that earlier researchers have reported. Rather, I addressed questions of validity anew for these measures.

In this section I further detail the three studies I used to answer each of the research questions above. With study one I examined the content validity of the

measures that I developed by asking a panel of experts to review them and recommend any changes to the definitions and scoring guides. In study two I analyzed the construct validity of these measures by looking at whether or not their resulting scores behave according to theory (for example, do scores for the subdomain of data-based decision making correlate highly with each other to offer evidence that they tap the same theoretical construct?). With the third study I used principal and teacher reports from surveys of principal expertise and practice as criterion variables by which to assess the criterion validity of these measures. I hypothesized that principals who demonstrate higher or lower levels of expertise in particular subdomains would self-report corresponding levels of or practice and/or have teachers who reported higher levels of expertise or practice for them as well.

To summarize, this study developed measures for the three different domains of principal expertise based on the literature I reviewed above and examined the validity of these measures and their relationships to one another. The three primary studies focused on the following goals.

Study 1 evaluated the content validity of the proposed subdomain measures by soliciting reviews and feedback from a panel of content experts.

Study 2 examined the construct validity of these subdomain measures by asking, do these measures of leadership expertise relate to each other as predicted by theory?

Study 3 explored the criterion validity of the expertise measures by asking, how do these relate to other measures of principals' expertise and practice?

Setting and Subjects

This study is part of an ongoing research project evaluation of a professional development program for school principals in one southeastern school district in the United States. The program was a district-level strategy that was designed to improve student achievement by arming principals with the knowledge and skills needed to lead instructional improvement efforts in their schools. A total of 48 principals were included in the study. This sample included all principals in the district except principals who were members of the district leadership team who were declared to be ineligible for the study since they were delivering the principal professional development to other principals in the district. Teachers from all participating principals' schools in the district were included in the study as well (n=2070).

Among the 48 principals, 28 were in elementary schools, 10 in middle schools, 6 in high schools, and 4 were in alternative/special education schools. I view the fact that all the principals in this study come from one district as a strength in that it held the district context and district-level policy context constant, though I acknowledge that the ability to generalize from these data is limited.

As can be seen in Table 1, even though all schools were located in the same urban district, there was substantial variation in their demographic characteristics. The average student enrollment for the schools of the 48 principals was 644, though the standard deviation of 301 indicated a substantial range across schools. On average, the schools of principals had an African-American enrollment of 67 percent, although the standard deviation of 26 percent indicates a broad range of student ethnicity in schools.

[Insert Table 1 Here]

Data Collection

Scenarios: Measures of Leadership Expertise

All principals in the sample responded to five written scenarios and one video simulation. All of these scenarios were "ill-structured" or open-ended--they consisted of complex school situations or problems with no clear solution implied. The scenarios were modeled after Leithwood and Stager's (1989) and Brenninkmeyer, Sherin, and Spillane's (2004) respective scenarios, which were designed to take advantage of Leithwood and Stager's (1989) finding that ill-structured problems differentiated expert from nonexpert administrators. Each scenario ended with an open-ended question to increase the opportunities for the principals to detail the expertise that they might use in addressing the problem. As Brenninkmeyer, Sherin, and Spillane write, such unstructured problems provide insight into a principal's thinking for two reasons: "they force one to structure the problem before being able to answer it, as well as come up with a solution based on that structure" (2004, p. 8). Furthermore, the scenarios mostly focused on instructional improvement situations and in some cases were subject matter specific. The first scenario was a video that asked participants to evaluate a brief snippet of a teacher's reading and writing lesson and summarize what feedback they would give the teacher in the video (it asked "what did you notice as you watched this video clip" and "what guidance, if any, would you give to this teacher"). The five others were written vignettes, asking principals how they would respond to school-related problems (see Appendix A for exact texts.)

Principals wrote narrative responses to the scenario problems on laptop

computers. They responded in an open-ended format and had 45 minutes to respond to all six scenarios. While principals could vary the amount of time they devoted to any one scenario, a proctor reminded them every 9 to 10 minutes that so much time had elapsed and they should be moving to the next scenario. Overall, the average number of words written per scenario was 84.8, ranging from 115.7 for scenario 1 to 71.9 for scenario 6, though length or response was not correlated with placing of scenario – response to prompt 2 of the simulation which came first generated the shortest response with an average word count of 63.7.

Principal Survey

The third study in this dissertation used data from a self-report principal survey to assess the criterion validity of the scenario results. This study compared principals' demonstrations of expertise in the scenarios to their self-reports of expertise and practice in the same areas. The principal survey items focusing on expertise were based on a revised and adapted version of *The School Leadership Self Inventory* (National Policy Board for Educational Administration, 2000), a self-reporting inventory consisting of Likert scale items based on the ISLLC standards for school leadership. The original inventory included items relating to the content of each of the six ISLLC standards (e.g. Articulates a vision of student learning for the school community, Supports a school culture focused on student learning). The items used in this study read as follows: "This question asks about your knowledge in a variety of areas of school leadership. For each area please indicate the degree to which you believe your current knowledge reflects personal mastery (knowledge and understanding of the

area)." The stem then read, "To what extent do you currently have personal mastery (knowledge and understanding) of the following:" The choices were a 5 point scale, "a little, some, sufficient, quite a bit, a great deal." This instrument was used in another study (Goldring & Vye, 2005) to study changes in principal knowledge after completion of a professional development program for school leaders. In that study this instrument was pilot tested and revised after extensive psychometric considerations, including factor analyses and reliability analyses; all of the original subscales yielded Cronbach's alpha reliability measures of .73 to .86.

For the third study additional measures were developed to capture leaders' self-reports of their expertise in particular areas. I hypothesized that principals who demonstrated higher expertise in the scenarios would also self-report having higher expertise in the same area. Higher correlations between the scenario scores and the principal self-reports would provide supporting evidence of these relationships. I examined the correlations between a number of the scenario scores and relevant scales on the self-reports. These scales included principals' self-reports of their expertise in data-based decision making, effective teaching and learning, monitoring instructional improvement, standards and systems thinking, subject matter, creating school learning cultures, data collection and analysis, and planning. Table 2 below summarizes the items in these scales and their respective alpha reliability scores, and Table 3 at the end of this paper specifies the actual items included in the measures for each primary domain of expertise.

Table 2. Principal Self-report Measures of Their Expertise		
Name	**Number of Items**	**Alpha Coefficient/Correlation**
Data-based Decision Making	3	α=0.82
Effective Teaching and Learning	6	α=0.84
Monitoring Instructional Improvement	2	r=0.82
Standards and Systems Thinking	2	r=0.68
Subject Matter	2	r=0.77
Creating School Learning Cultures	1	.57 (test-retest reliability)*
Data Collection and Analysis	1	.73 (test-retest reliability)
Planning	4	α=0.86

** The test-retest reliability calculations above used data from a principal survey given previously to the same participants with the same questions on it.*

The principal survey also included a group of items that asked school leaders about the frequency with which they engaged in particular actions that related to the areas of expertise. I hypothesized that principals who were higher in different areas of expertise would engage in related activities more frequently. Therefore, those individuals who demonstrated higher expertise in scenarios would also self-report engaging in relevant practices more frequently. The stem for these questions read, "During the current school year, how often did you do any of the following?" The choices were a 5 point scale with the following responses: "never, a few times throughout the year, a few times per month, 1-2 days per week, more than 2 days per week." Reliability measures across these scales ranged from .72 to .87, with one additional two-item scale having a correlation of .5 (the principal's practice of support staff development). These

measures included principals' participation in practices of data-based decision making, examining and discussing student work, monitoring instructional improvement, staff development, and planning. Table 4 summarizes these scales, and Table 3 at the end of this paper specifies the actual items used in these measures.

Table 4. Principal Self-report Measures of Their Practices

Name	Number of Items	Alpha Coefficient/Correlation
Data-based Decision Making	11	A=0.82
Examining and Discussing Student Work	3	A=0.72
Monitoring Instructional Improvement	4	A=0.80
Engaging in Staff Development	2	R=0.50
Planning	5	A=0.87

Teacher Survey

I also hypothesized that teachers would observe their principals' expertise in different interactions with them. The third study used teacher survey measures to examine whether or not principals who showed higher expertise in the scenarios also had teachers who reported that they had greater expertise in those areas or engaged more frequently in related practices. I examined correlations between the scenario and teacher surveys for evidence of these relationships between the measures. All teachers and other professional staff in the principals' schools responded to these surveys at the same time that the principals responded to their surveys. The response rate for the school staff

survey was 87% (N=2070).

Teachers first answered a number of questions about their principals' understanding of different areas. These items contained the following stem: "Please mark the extent to which you disagree or agree which each of the following: The principal at this school has a strong understanding of…" Teachers answered using a 4-point Likert scale that included responses of "strongly disagree, disagree, agree, strongly agree." Multiple item scales in this survey included the following: teachers' reports of their principal's understanding of principles of pedagogical content knowledge and their principal's understanding of how to support professional development (despite this item's low test-retest reliability it was nonetheless included to explore the relationship between the scenarios and survey measures). Table 5 summarizes these scales, and Table 3 at the end of the paper specifies those items that were included for each scale.

Table 5. Teacher Survey Reports of Their Principals' Expertise

Name	Number of Items	Alpha Coefficient/Correlation
Principal Pedagogical Content Knowledge	3	α=0.92
How to Support Teacher Professional Development	1	.57 (test-retest reliability)*

** The test-retest reliability calculation above used data from a teacher survey given previously to the same participants with the same questions on it.*

Teachers also answered questions about their principals' practices. The stems for these questions were the following: "Please mark the extent to which you disagree or agree which each of the following: The principal at this

school…" These questions included a number of activities and practices with 5-point Likert scale responses of "not applicable, strongly disagree, disagree, agree, strongly agree." Cronbach's alpha reliability cofficients for these constructs ranged from .75 to .93, with some correlations between two-item scales and test-retest reliability values for single-items being lower (.5 to .7). Teachers reported the extent to which principals monitored instructional improvement, evaluated instruction, developed teacher capacity, encouraged teachers' improvement in learning, encouraged teachers to take responsibility, and developed/planned/communicated instructional goals. The survey questions also asked teachers the extent to which their principals encouraged them to improve their teaching), took interest in teachers' professional development, and were open to discussing worries and frustrations. Table 6 summarizes these measures, and Table 3 at the end of this study details the items included in each.

Table 6. Teacher Survey Reports of Their Principals' Practice

Name	Number of Items	Alpha Coefficient/Correlation
Monitor Instructional Improvement	3	α=0.84
Evaluate Instruction	2	r=.92
Develop Teacher Capacity	3	α=0.81
Encourage Teachers' Improvement in Learning	2	r=.60
Encourage Teachers to Take Responsibility	2	r=.70
Develop/Plan/Communicate Instructional Goals	6	α=0.93
Encourage Teaches to Improve Their Teaching	2	r=.60
Interact with Teachers Regarding Instruction	5	α=.75
Demonstrate Interest in Teachers' Professional Development	1	.55 (test-retest reliability)*

Open to Discussion Worries and Frustrations with Teachers	1	.50 (test-retest reliability)

** The test-retest reliability calculation above used data from a teacher survey given previously to the same participants with the same questions on it.*

Before discussing the objectives in this paper it is important to emphasize that the reliabilities and correlations for these scales ranged greatly (two-item and single-item values were often lower). Many of these lower values derived from the fact that they included fewer items (Gliem & Gliem, 2003), and the criterion validity results in the third study are therefore limited by these low values—a number of these scales are marginal in their reliability. Nonetheless, I have included these items because of the exploratory nature of the criterion validity study: correlations between the scenarios and "low-reliability" measures offer initial support for examining these relationships in future studies.

Methodology

Study 1. Content Validation through Expert Panel Feedback on Expertise Measures

This first study focused on the content validity of scoring rubrics that I developed for each of the subdomains of leadership expertise. While I lay out the integral components of each subdomain of expertise below and explain why the proposed measures for this study adequately capture each area, such theoretical arguments to support these measures offer only a starting point for examining their validity. A crucial next step involved soliciting experts' reviews of the measures.

Evaluating measures of expertise involves a closer look at their content—how much does each of them cover the range of meanings included under each

domain of expertise? Content validity examines the degree to which a measure reflects the content of a particular construct; the goal of a content validation study is to assess whether the measure's items represent the intended construct. Under this broader umbrella of content validation, logical validity includes the specific strategy of asking a group of experts separate from the researcher to review a test and determine if it taps those concepts that it claims to tap. Because the proposed scoring rubrics in these three areas of expertise have not been used before, I asked a group of content experts to review them in an effort to evaluate their content validity. Allen and Yen (1979) comment that through expert review "a person examines the test and concludes that it measures the relevant trait" (p. 96). Nunnally and Bernstein (1994) similarly state that such a review asks others (in this particular study, a group of experts) if they "feel the instrument measures what it is intended to measure" (p. 110).

I solicited expert feedback on the rubrics and then used their comments to modify the rubrics. I used two groups in a two-step process to evaluate the content validity of the measures, principal experts, and content experts. First, I asked a group of principals identified as experts to respond to the scenarios summarized above. Second, I asked content experts to review the proposed rubrics, score the expert principals' responses with the rubrics, and then recommend changes based on their experiences and content knowledge.

I identified the first group of principal experts by contacting university and state department of education officials who have worked closely with

principals on an ongoing basis.[3] In a form letter I summarized the study and specified the three areas of leadership expertise under development, and I asked them to nominate principals whom they identified as experts in one of the three areas (Appendix B contains this letter). I also specified that they nominate individuals (if possible) from elementary, middle/junior high, and high school levels. With these references I identified a total of fifteen expert principals, three for each area of expertise (learning content knowledge, learning-centered leadership, problem-solving expertise) with two acting as potential back-ups. For each of the three areas of expertise I chose one principal from each school level. I then contacted these principals and asked them to complete the scenarios according to the same guidelines and restrictions that principals in the sample were given as they responded (e.g. they read each of the scenarios and wrote how they would respond to the situation, and they had a total of one hour to complete all of the scenarios). Of the nine I first contacted two did not respond, and I used the back-up candidates to reach the goal of three expert principals for each domain (Appendix C summarizes the participants and their school levels).

In the second step I identified a group of content experts according to their research in one of the three domains of expertise. I looked particularly at their publications, presentations, and the roles they have played in educational leadership research in the three domains of leadership expertise: leadership content knowledge, learning-centered leadership, and problem-solving expertise. I selected a group of individuals who a) have conducted research and published

[3] I identified these university and state employees through recommendations from my dissertation committee members and additional colleagues at three universities in the midwest and northwest United States.

in one of the three areas of expertise, b) are currently engaged in research in one of the three areas, or c) have conducted research using scenario measures that I asked them to evaluate.

The literature on how many experts to include ranges significantly. While Lynn (1986) recommends a minimum of 3, others specify a range of 2 to 20 (Gable & Wolf, 1993; Walz, Strickland, & Lenz, 1991). As with the principal experts I chose a total of nine individuals with three providing feedback in each of the areas of expertise. Appendix D lists the experts I contacted; it includes each person's current position and work responsibilities, research and scholarship relevant to the selected area of expertise, and a brief rationale for including him or her in the group. Because of the highly theoretical nature of this part of the content validity study I did not solicit content expert feedback from practicing educational leaders such as principals or district officials.

These content experts received an email with two documents for their respective domain of leadership expertise. The first offered a theoretical summary of the domain and its subdomains, the actual scoring rubrics for that domain, a series of questions about the scoring rubrics, and instructions for them to complete the evaluation. The second document included the scenario responses from the principal experts in their respective domain (see Appendix E for an example of these feedback documents and instructions focused on "Leadership Content Knowledge").

The instructions guided content experts to read the domain summary and scoring rubrics along with the principal experts' responses. Using the scoring rubrics the content experts then scored the responses according to each of the different subdomains in each domain (for example, "subject matter,"

"pedagogical content knowledge," and "teachers as learners" fall under the expertise domain of "leadership content knowledge"). Upon completion of the scoring they were asked to complete a set of questions about each of the scoring rubrics. These questions consisted of the following items.

1. "The scoring rubric offers a clear definition for this subdomain."
2. "The definition for this subcategory needs to include additional dimensions for it to be more complete."
3. "The definition for this subcategory needs to include fewer dimensions for it to be more accurate."
4. "The directions provide clear guidance about how to use the rubric to score the text."
5. "The scoring guide provides clear explanation of what response qualifies for each level of expertise."
6. "The scoring guide provides clear examples of responses that qualify for each level of expertise."

Each of these items was followed by a series of 5-point Likert scale responses: "completely disagree (1), mostly agree (2), neutral (neither agree nor disagree) (3), mostly agree (4), completely agree (5)." Respondents were asked to explain further their responses in more detail below each of the items. In a final section, after completing the scoring and the comments each expert reported the extent to which each scenario prompted a principal to demonstrate his or her expertise in each subdomain. For each scenario experts were asked "to what extent do you think each scenario prompts principals to demonstrate expertise for each subdomain?" They could choose one of the following responses: "a great deal," "somewhat," "a little bit," or "none at all." Appendix E includes a full example of the questions for the subdomain of "subject matter" under leadership content knowledge.

I analyzed content experts' responses in three areas: a) feedback and comments regarding the scoring protocols and the definitions used in each rubric, b) feedback and comments about the directions used in each scoring

rubric, and c) feedback and comments about the examples and explanations included in each scoring rubric. In these areas, experts' answers on the Likert scale items were analyzed descriptively to compare their responses to one another. Their more detailed written comments were analyzed according to one of the three areas listed above, and the type of comment they provided (e.g. did they recommend adding texts or changing existing texts).

Finally, the content experts' actual scoring of the three acting principals' expert responses were analyzed for consistency—how much did experts' scores of the responses agree with one another? This last analysis provided evidence of how well the measures produced consistent scores from the experts. I viewed consistency or agreement between experts on the scores as evidence that the measures clearly defined and demonstrated the areas of expertise such that the experts agreed on their definitions and use, while disagreement between experts provided at least some evidence of the need to clarify the rubrics further. While other factors (such as experts' different initial assumptions about these areas of expertise) may certainly have contributed to any disagreement that existed, I used the scores as a further guide to revisit and clarify the rubrics.

All small recommendations regarding directions or examples were documented from the feedback data and used to make edits or corrections. More substantial recommendations were considered against others' comments and feedback (i.e. did other experts recommend the same changes to the materials? Did experts contradict one another in the changes they recommended or their evaluations of the example responses)?

In Chapter 5 I first summarize the scoring rubric guides that I asked content experts to evaluate, and I then report my findings from their feedback for each subdomain and discuss how I used the data to modify the rubrics.

Study 2. Examination of Measures' Construct Validity

I used the revised rubrics for Study 1 to score the responses from the 48 principals in the sample described above. Before coding this principal data, another graduate student and I scored data from five trial cases (from another round of scenarios administered to principals) and compared our scores to check for agreement. Once we reached a satisfactory level of agreement I completed coding of all the principals' responses to the scenarios on my own. This second study examined relationships between the resulting scores, and I organized this section into three substudies.

As explained extensively by Cronbach and Meehl (1955), construct validity refers to the degree to which a measure captures the construct or "postulated attribute of people" (p. 283) it was designed to measure. With this form of validity, a construct's meaningfulness or importance is made explicit first through its definition (as I have laid out earlier in this paper) and then by an examination of how it relates to other variables. With a given construct and its definition or measure, one can hypothesize about its behavior under certain circumstances, and confirmation of these hypotheses offers evidence that the measure as operationalized adequately captures the construct. Allen and Yen (1979) and Crocker and Algina (1986) discuss a number of different strategies to assess construct validity, one of which involves examining correlations between the measure of interest and others. Significant correlations in the hypothesized

direction between theoretically related measures provide initial evidence that the constructs behave according to predictions.

I started by analyzing the behaviors of each measure through qualitative examples and descriptive results of the scoring. At a basic level I asked, do the measures capture varying levels of expertise between principals? In the first substudy I provided qualitative examples from principal responses that showed varying demonstrations of expertise between individuals, and I discussed how the scoring rubrics captured these differing levels of expertise. I also used descriptive summaries of the scores to demonstrate how different scenarios prompted different demonstrations of expertise across principals.

Using Crocker and Algina's (1986) recommendation for construct validation, I then examined whether or not the measures within each of the three larger domains related to each other as predicted by theory. While researchers have identified the key components or *subdomains* that comprise the three larger areas of expertise, each of these subdomains is still quite distinct conceptually, and high expertise in one part of a domain does not guarantee high expertise in a second part. For example, just because a principal shows high expertise in data-based decision making does not mean that she/he will show high expertise in effective teaching and learning. Accordingly, in the second substudy I analyzed the correlations and alpha reliabilities between the subdomain scores to determine if they were theoretically distinct and yet internally consistent as scale measures for the three larger domains.

In the third and final substudy I asked, how do overall scores for the three main domains relate to each other? I used aggregated scores of the subdomain results to explore the relationships between the three larger domains of

leadership expertise. In light of the three separate bodies of literature that have developed the primary domains of expertise, I hypothesized that the relationships between these three would be small: researchers have defined distinct areas of expertise, and school leaders who are high in one are not guaranteed to be higher in another. For example, just because a principal demonstrates high expertise in the "content knowledge" necessary to guide teaching and learning in her school does not mean that she will possess the "problem-solving expertise" necessary to oversee the planning and delegating of responsibilities for such improvement. Based on this hypothesis I predicted finding small correlations between the three main domains of leadership expertise. On the other hand, larger correlations between the domains would suggest a) that researchers have identified conceptually similar areas of expertise that are strongly related, or b) that the structures for these areas of expertise may be different than conceptualized. For example, while the literature for "problem-solving expertise" has implied that leaders with expertise in "gathering information" will also be high in "planning" expertise, it may be the case that principals high in "planning" are more likely to possess expertise from one of the other primary domains such as "data-based decision making" or "monitoring instructional improvement." Higher correlations between the domains may offer evidence that researchers in each domain have confined themselves too narrowly to particular theoretical foundations as they have examined or advocated for particular expertise that principals need to do their jobs. Such limitations could restrict their consideration of the alternative areas of expertise that principals require in their work. With the results of this third study I examine the nature of the primary domains' relationships to one another.

Study 3. Examination of Measures' Criterion Validity

Researchers use criterion validity when test scores can be related to other events or behaviors that occur before, during, or after a test is applied (Nunnally & Bernstein, 1994, 94). Criterion validation addresses how well a set of test variables predicts an outcome based on a second set of criterion variables (Pennington, 2003), and it is typically expressed as a correlation between the test and criterion scores (Allen & Yen, 1979, p. 97). Cronbach and Meehl (1955) also argue that it can be used when "one test is proposed as a substitute for another (for example, when a multiple-choice form of a spelling test is substituted for taking dictation)" (p. 282). For example, correlations between the scenario scores and teachers' survey reports of their principals' expertise in monitoring instructional leadership could provide evidence that the two measures both tap a similar construct.

Beyond work by Goldring et al. (2008 and 2009, see below) there are few if any examples of efforts to explore the criterion validity of leadership expertise measures. Leithwod and Stager (1989) and Brenninkmeyer and Spillane (2008) both demonstrated how experts and non-experts principals differed in their problem-solving strategies (and thus their status as experts generally correlates with their different responses to problems). Leithwood and Steinbach (1993) found limited connections between principals' problem-solving expertise as measured in interviews and their staffs' survey reports of their transformational leadership practices (they ultimately argued that both constructs must be tapped to understand principals' broader "quality leadership"). Beyond these efforts researchers have not examined how well other methods of measurement capture

differences in expertise or conditions theoretically related to principals' expertise (such as survey measures of their behavior).

Work from research to measure teacher knowledge provides some guidance in this area. Previous studies examining the criterion validity of vignettes (both written and video) to measure teacher knowledge have evaluated the relationships of scores from the vignettes to other measures of teacher knowledge and/or practice. For example, Stecher, Le, Hamilton, Ryan, Robyn, and Lockwood (2006) examined the correlations between mathematics instruction vignette scores and teachers' survey self-report scores on different aspects of mathematics instruction (p. 116) as well as log and observation results of teacher practices. Kersting (2008) used teacher scores on a math content knowledge test and expert ratings of the teachers as criterion variables; she investigated correlations between these and teacher responses to video clips measuring math content knowledge. All these researchers argued that an integral strategy to exploring the validity of a new measure is to look at its relationship to other theoretically relevant variables. Both papers found high and significant correlations between their vignette scores and some of the other measures; they used these findings to bolster their cases for the validity of their vignette measures.

I first discuss previous criterion validation efforts that used portions of the data from this dissertation before explaining the analyses I conducted for this last study. Earlier papers from these same data have reported mixed findings of relationships between principals' scenario scores and the principal and teacher reports (see Goldring, et al., 2008 and 2009). These papers first hypothesized that principals who demonstrate greater expertise in learning-centered leadership

through the scenarios would a) self-report greater expertise in related areas, or b) self-report more frequent practices in related areas. Because expertise theoretically guides individuals' actions, those principals with more expertise in certain areas of leadership would engage in related activities more than those with less expertise. The papers then hypothesized that principals with greater expertise would have teachers who a) reported their higher expertise or more frequent practices in those areas, or b) more frequently reported the presence of related conditions in the schools. The authors theorized that principals with greater leadership expertise would not only engage in these actions more frequently (and that teachers would report these more frequent practices), but they would also promote related conditions in their schools (for example, principals with greater expertise in standards-based reform would have teachers who more frequently engage in the alignment of standards and school programs). Thus correlations between these three sets of scores would reflect hetero-method measures of the same traits (see Goldring, et al. 2008 for a longer discussion). The authors therefore predicted significant correlations between the scenarios and principal and/or teacher surveys.

Their results did not fully support their predictions. The scenario responses did not correlate highly with principals' self-reported expertise on the principal survey in the same domains of expertise (e.g. standards-based reform, data-based decision making), and scenario correlations with teacher survey reports were greater than the principal surveys. Goldring, et al. (2008) offered three interpretations that inform this current work. First, the evidence suggested that the scenarios and principal surveys comprise measures of different constructs such as principals' tacit knowledge versus their declarative

knowledge, respectively. In other words, the scenarios may do a better job of capturing principals' tacit knowledge (the expertise they employ in their actions), while the surveys may comprise better measures of what principals can declare or say that they know. Second, high correlations between the principal survey measures (from .83 to .85) suggested that these survey scales may be tapping domains of expertise that are indistinguishable from each other, or measuring one overall level of expertise, rather than separate domains. The measures may also be subject to a common source of influence such as self-report bias. On the other hand, the range of correlations between scenario measures (-.02 to .47) suggested there is more of a difference between the domains of expertise they capture. Third, correlations between the scenario results and the teacher survey data returned higher (and more variable) correlations on average than between the scenarios and principal surveys. The authors argued that the wider ranges in correlations offered more evidence that the scenarios tapped more distinguishable domains of expertise. They also contended that the stronger correlations demonstrated that in some cases the principal scenario responses are better able to measure expertise--for example, principals scoring higher on principles of effective teaching and learning are in schools where their teachers report principals have more knowledge of principals of effective teaching and learning (r=.43). Teacher reports of their principals may also comprise better measures of their leaders' expertise in that they a) are not subject to a self-report bias, and 2) the aggregation of multiple teacher scores to the school level (as used in the analyses) limits the possible influence of individuals or groups on the reports of expertise.

In both 2008 and 2009 Goldring, et al. conceded that much works remain to be done to examine just what these different measures capture and what their relationships are. Building on this work, I employed a wider array of variables from the principal and teacher surveys as criterion variables for the different measures of expertise in each of the three domains. From the principal surveys I included school leaders' self-reports of their expertise and practice in related areas, and from the teacher surveys I used teachers' reports of their principals' expertise and practices in related areas. (I have summarized these variables in the principal and teacher survey summaries above). As already discussed, Goldring, et al. (2008 and 2009) found no relationships between the principal survey and scenario scores of principal expertise. The researchers offered a number of possible explanations for the lack of correlations, ranging from differences in the methods used to varying perceptions by principals about their respective levels of expertise (see previous summary). With the variables that I have employed in this study I revisited their hypothesis that principals with high scenario scores would self-report higher expertise in similar areas.

I predicted finding stronger correlations between the scenario scores and principals' reports of their practices. As discussed above, Stecher, et al. (2006) found significant relationships between teachers' vignette scores and their self-reported use of mathematical processes in their teaching. I hypothesized that because expertise guides individuals' practices, those principals with greater levels of expertise would therefore engage more frequently in the related practices.

Goldring, et al. (2008) also commented that teachers' reports of their principals' expertise may be less susceptible to self-report bias and that aggregate

scores of teachers' responses help to limit individuals' or groups' influences on the reports; for these reasons I predicted that teacher reports of their principals' expertise would show higher correlations with the related scenario scores of leadership expertise. For the same reasons I hypothesized that teacher reports of principal practice would also correlate more highly with their respective scenario measures of expertise. In this third study I have presented the correlations between these different measures and discussed whether or not they supported these predictions.

CHAPTER V

STUDY 1 RESULTS: CONTENT VALIDATION THROUGH EXPERT PANEL FEEDBACK ON MEASURES OF LEADERSHIP EXPERTISE

This chapter presents the findings from the first study that used expert panel feedback to evaluate the content validity of the scoring rubrics. I break this chapter into the three main areas of expertise. For each domain of expertise I first explain how I used the literature for these areas to develop the rubrics that I proposed, and I then present experts' specific feedback and detail how I used the feedback to modify the rubrics.

I begin the summary for each measure with a copy of the proposed rubric. Because these experts both answered the survey questions described above and provided written comments, I first offer descriptive results from the survey items to examine larger trends in experts' concerns. Second, I analyze experts' more detailed written comments and discuss significant recommendations about changes to make to the scoring rubrics. Third, I review how experts used the rubrics to score expert principals' example responses, and I discuss whether or not agreement existed between the experts' scores to provide evidence that the rubrics promoted common understanding of each rubric. Finally, I discuss how I used their comments to revise each of the rubrics, and I offer examples of the changes I made. The full range of feedback provided a rich review of the content validity of each of the measures. Appendices F, G, and H include the final definitions and rubrics that I used to score the scenarios for Study 2.

Summary of Proposed Rubric Scoring Guides

Each of the scoring rubrics provided a definition of the key components for a particular domain, and it discussed how to score principals' written responses using the definitions. As discussed earlier, each domain (leadership content knowledge, learning-centered leadership, and problem-solving expertise) contained a number of key components or subdomains that researchers defined. The rubrics assigned a numerical value to principals' written answers based on the quality of their response for each subdomain. Thus a score for each answer was given according to the degree that it demonstrated expertise in one of the subdomains. This analytical strategy followed what Tashakorri and Teddlie (1998) have referred to as "quantitizing"—the "[conversion] of qualitative information into numerical codes that can be statistically analyzed" (p. 126).

All of these rubrics assigned scores to the responses that captured not just how frequently a principal mentioned a concept in a subdomain but also the quality of response a principal provided. The rubrics assigned quantitative scores based on two considerations: (1) how many times a principal referred to a component of each subdomain and (2) whether or not the principal's response went beyond merely mentioning a subdomain to offering a deeper discussion and understanding of it. The table below offers a summary of the types of scores and definitions used in each of the rubrics.

Table 7. **Summary of Rubric Scoring Guides**

Score	Summary of Response
0	No mention of the subdomain
1	A Little Discussion: the principal offers only 1 or 2 mere mentions or superficial discussions of the subdomain with no elaboration
2	Some Discussion: the principal provides 3 or more mere mentions of the subdomain with no elaboration
3	Sufficient Discussion: the principal discusses one subdomain in more detail suggesting a deeper understanding of it
4	Quite a Bit of Discussion: the principal discusses two or more subdomains and develops them with more details that suggest a deeper understanding of them
5	A Great Deal of Discussion: the principal discusses two or more subdomains in more detail and then develops a link or connection between the two subdomains

In sum, a principal could score 1 or 2 if he or she simply mentioned a concept one or more times. However, to score a 3 or better, the respondent had to demonstrate more than superficial knowledge of the subdomain by discussing central components or dimensions of the subdomain in greater detail. The scoring strategy rewarded those responses that went beyond superficial mentions of an area of expertise to offer more substantive discussions; these analyses drew a distinction between those principals who simply mentioned a concept numerous times and those who demonstrated deeper levels of expertise in the domains through more detailed discussions.

Leadership Content Knowledge Rubrics

As outlined earlier Stein and Nelson (2003) laid out three specific components for principals' content knowledge: 1) the different subject matters taught in their schools, 2) pedagogical content knowledge that helps explain how students learn different subjects, and 3) an understanding of teachers as learners and effective ways to teach teachers (p. 426). They argued that principals needed

to understand not only how the subject matter itself differs in structure but also how teaching and learning needs differ for students and teachers according to subject.

Based on their 2003 discussion as well as the additional articles I summarize above I lay out three specific subdomains to identify and evaluate the essential dimensions of a principal's leadership content knowledge. Appendix G contains both the final operational definitions and the instructions and examples for scoring a principal's response.

The first subdomain is subject matter. A principal's response must demonstrate some understanding of the unique characteristics of a particular area of content. A low level response must mention some of the basic characteristics of subject areas or the differences between them to receive a score of 1 or 2. In these cases a respondent might offer a brief discussion of such a topic but provide few details (see Appendix B). To score a 3, 4, or 5 a principal must not only mention the different characteristics of one or more subject areas but also provide more details about these subject areas to demonstrate a deeper knowledge of the subjects or differences between them.

The second subdomain is pedagogical content knowledge--knowledge that is unique to the different subjects that teachers teach. This may include not only the specific strategies that a teacher uses to teach a subject but also theories or beliefs about the best way to teach concepts or methods. The scoring in this subdomain focused on three different areas: 1) theories of how students learn in different subjects, 2) effective teaching strategies for different subject areas, and 3) how teacher knowledge or strategies may differ across subject areas.

The third domain consists of principals' knowledge of "teachers as learners," or professional development strategies for teachers. The primary components in this subdomain consisted not only of theories of professional development and how to support teachers' improvement of their instruction but also how different subject areas might influence the professional development needs of teachers.

Leadership Content Knowledge Results.

I divided this section according to the three subdomains. While all three experts provided written comments for the rubrics in this domain, only two of the three content experts provided answers to the survey items. For this domain I provided a descriptive summary of their responses for each of the subdomains, and I relied heavily on experts' extensive written comments about the rubrics.

Subdomain 1: Subject Matter.

The figure below shows the initial scoring rubric that content experts evaluated in their comments; I summarize their comments and then explain how I used their feedback to modify the rubric.

Figure 1. Proposed Rubric for Subject Matter.

1. Subject Matter

Definition: includes any responses in which the principal mentions, expresses or demonstrates some knowledge of ANY of the following aspects of subject matter covered in the classroom:

- the nature of the content or material that is taught in different subject areas (the scenarios and analysis here focus primarily on mathematics and literacy and reading/language arts)
- differences in the nature of the content across subject areas (for example, a principal might discuss how mathematics possesses a definable body of concepts, symbols, vocabulary, and tools whereas as literacy content may stretch across multiple areas such as language, literature, and composition)
- ways in which subject matter content differences influence other aspects of teaching (for example, because of its more diverse materials a literacy program may be focused around certain assessments of skills, whereas a mathematics program may be more "topic driven" in which agreed-upon content drives the structure of the teaching and instructional time)

0. No Mention of the Dimension

1. A Little

Mere mention of one or two aspects of subject with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

> Specific example of a mere mention of the nature of a subject area:
> "Reading is a tool to enter into the larger world of information and life skills."

2. Some.
Mentions at least three or more different aspects of subject matter but does not develop any of the aspects.

3. Sufficient
Mentions at least one aspect of subject matter and develops at least one aspect. This means the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, a more developed discussion of subject matter should include multiple details in the discussion as well as an explanation of why the approach is valuable or important.)

> Specific example of a single aspect of subject matter that is developed: "It is important that students in this (math) program understand the basic rules of addition and subtraction; these are the important skills you build on and use in other solving problems before learning other things like multiplication and subtraction."

4. Quite a Bit
Mentions at least two aspects of leadership content knowledge and develops two or more; that is, the response goes beyond mentioning the aspects to developing them with more discussion that suggests a deeper understanding of the aspects.

5. A Great Deal
Mentions at least two aspects of leadership content knowledge and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of effective leadership content knowledge to making a link or connection between at least two aspects. For example, a principal may discuss 1) how subject matters differ in their content and learning requirements for teachers and therefore 2) how professional development strategies need to differ according to subject areas so that 3) such programs can ultimately help to improve the pedagogical skills that teachers employ in their classrooms (this last phrase ties together the first two)

First, review of the two experts' responses to the survey items showed many differences of opinion about the rubrics. For example, while both "mostly agreed" that the directions provided clear guidance (item 2.a), they disagreed about whether or not the rubric definitions needed additional dimensions (in item 1.b one "mostly agreed" while another "mostly disagreed"). In situations where the experts disagreed I relied heavily on their written comments (as with item 1.b). Items 2.b. and 2.c. also drew careful scrutiny from reviewers; one or both experts disagreed that the rubric provided clear explanation/examples of what qualified for each level of this subdomain of expertise.

Table 8. Subject Matter Feedback Response			
	Expert		
Question Summary	**1**	**2**	**3**
1. a. rubric provides clear definition	?	4	3
1. b. definition needs additional dimensions	?	4	2
1. c. definition needs fewer dimensions	?	3	3
2. a. directions provide clear guidance	?	4	4
2. b. rubric provides clear explanation of what response qualifies for each level	?	2	3
2. c. rubric provides clear examples for each level	?	1	?

1: completely disagree, 2: mostly agree, 3: neutral (neither agree nor disagree), 4: mostly agree, 5: completely agree

Second, as with their responses to the items above, in the written comments for this subdomain content experts were most concerned about how well the scoring rubric provided a clear explanation of what response qualifies for each level of expertise. Expert 2 made the most specific statements about the examples and explanations:

> The third bullet under subject matter isn't clear to me where it states, "ways in which subject matter content differences influence other aspects of teaching…" What is meant by "other aspects of teaching?"

and

> I'm not sure I understand what is meant by "the nature of the content." Is it referring to an understanding of the content? If not, I would suggest that a statement that refers to an understanding of the content should be included. I also think it would be useful to have an indicator that looks at the principal's stance towards the subject matter. In other words, does the principal see the subject of math, for example, as a set of procedures or ideas, etc."

(Below I explain how I modified the rubric in response to these comments.)

Third, experts' scores of the principals' responses provided final evidence of how well the rubrics presented clear and consistent guidance for identifying subject matter expertise. Table 9 presents the experts' subject matter scores for the principals.

Table 9. Experts' Scores for Subject Matter Responses	Expert		
Scenario	1	2	3
Principal 1			
1	?	0	0
2	?	1	1
3	?	0	0
4	?	0	0
5	?	0	1
Principal 2			
1	1	2	1
2	5	2	1
3	3	3	1
4	0	1	0
5	?	0	0
Principal 3			
1	?	0	0
2	1	3	2
3	1	0	0
4	1	0	0
5	0	0	0

A review of the experts' scores for the principal expert responses showed highly consistent scores between Experts 2 and 3 (Expert 1 did not score all the responses). For example, when the first principal addressed the low school reading scores in scenario 2 and discussed teachers' need to understand "balanced literacy, including clear definitions and examples of each component of a balanced literacy model," both experts rated this as a "1" or a superficial discussion of subject matter. Expert 1 scored approximately two-thirds of the responses and showed the greatest disagreement with the other two scores. For example, principal 3 in scenario 3 wrote the following in response to questions about the value of standardized test scores.

> A standardized test does not inform practice to the extent that teachers learn how to intervene with student. It merely tells you the "what" area to intervene but not "how" to do it. Given the limitations and strengths of standardized tests, as a staff, we need to construct ways to utilize the

> information that they do provide. Through disaggregating the data we can make more informed decisions about "what" we want to address… it will be up to us to figure out "how" to do it. Most likely, we will need to seek out research-based effective practices and find ways to provide professional development and resources to implement those effective practices.

Expert 1 scored this content as a 1 (a superficial mention) for subject matter, even though these comments refer much more generally to theories of how best to teach reading and professional development strategies for teachers. Experts 2 and 3 scored this response as a "0" for subject matter. While the scores provided some evidence that the definitions were helpful to two experts (because of agreement between 2 and 3), the first experts' scores raised questions about their clarity.

Changes to Rubric for Subject Matter

Based on the written comments above and these differences in the experts' ratings I revisited the definitions and the rubrics' examples to clarify further the dimensions of "subject matter" and what qualified under each level of scoring. Key changes I made to this rubric were to provide additional examples and explanations for each level in the rubric. I explained in more detail in the definition what was meant by "subject matter," and I provided more examples in the definition to clarify the "nature" of the content as well as how literacy and math by their nature might influence the pedagogical strategies used for each of them. I added the following text to the definition in the rubric.

> ways in which subject matter content differences influence other aspects of teaching (for example, because of its more diverse materials a literacy program may be focused around training and assessments of certain skills as opposed to pre-set content that is to be covered, while a mathematics program may be more "topic driven" in which agreed-upon content drives the structure of the teaching and instructional time)

I also provided in the definition an example of how math and literacy might differ as content:

> the nature of the content or material that is taught in different subject areas (the scenarios and analysis here focus primarily on mathematics and literacy and reading/language arts). (For example, a principal might discuss how mathematics possesses a definable body of concepts, symbols, vocabulary, and tools, or she might describe how literacy content may stretch across multiple areas such as language, literature, and composition.)

Finally, I included examples at each scoring level to demonstrate comments that would qualify for each different score. The figures below illustrate changes made to the definitions and the first three scoring levels for this subdomain; modifications to the rubric are bolded, italicized, and underlined. The figure below shows the modified rubric with all changes highlighted (this format is used in the rest of this study to illustrate changes to the rubrics).

Figure 2. Modified Rubric for Subject Matter

Subject Matter

Definition: includes any responses in which the principal mentions, expresses or demonstrates some knowledge of ANY of the following aspects of subject matter covered in the classroom. ***For a response to qualify under this category the principal must discuss key components or concepts of the subject matter or the nature of the subject matter.***

- ***different constructs, concepts, or ideas that are central to a particular subject matter. (For example, a principal might discuss how specific arithmetic skills are central to students' mathematical learning or particular reading skills are integral to a students' ability to read.)***
- the nature of the content or material that is taught in different subject areas (the scenarios and analysis here focus primarily on mathematics and literacy and reading/language arts***). (For example, a principal might discuss how mathematics possesses a definable body of concepts, symbols, vocabulary, and tools, or she might describe how literacy content may stretch across multiple areas such as language, literature, and composition.)***
- differences in the nature of the content across subject areas ***(for example, a principal may point out the differences between math and reading that are summarized above).***
- ways in which subject matter content differences influence other aspects of teaching ***(for example, because of its more diverse materials a literacy program may be focused around training and assessments of certain skills as opposed to pre-set content that is to be covered, while a mathematics program may be more "topic driven" in which agreed-upon content drives the structure of the teaching and instructional time)***
- ***comments or opinions that indicate the principal's stance toward the subject matter. (For example, does the principal see math as a set of procedures to solve problems or a set of ideas about numbers to explore and evaluate?)***

0. No Mention of the subcategory at all in the response ***(Examples that would score a "0" include responses that discuss curriculum but do not elaborate on a specific subject area.)***

1. A Little Discussion of the subcategory in the response

Mere mention of one or two aspects of subject with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

Specific example of a mere mention of the nature of a subject area:
"Reading is a tool to enter into the larger world of information and life skills."

2. Some Discussion of the subcategory in the response
Mentions at least three or more different aspects of subject matter but does not develop any of the aspects.

3. Sufficient Discussion of the subcategory in the response
Mentions at least one aspect of subject matter and develops at least one aspect. This means the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, a more developed discussion of subject matter should include multiple details about a subject matter in the discussion as well as an explanation of why the approach is valuable or important.)

Specific example of a single aspect of subject matter that is developed: "It is important that students in this (math) program understand the basic rules of addition and subtraction; these are the important skills you build on and use in other solving problems before learning other things like multiplication and subtraction."

4. Quite a Bit of Discussion of the subcategory in the response
Mentions at least two aspects of subject matter (such as the more developed example above) and develops two or more ; that is, the response goes beyond mentioning the aspects to developing them with more discussion that suggests a deeper understanding of the aspects.

5. A Great Deal of Discussion of the subcategory in the response
Mentions at least two aspects of leadership content knowledge and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of effective leadership content knowledge to making a link or connection between at least two aspects.

For example, a principal may discuss 1) how subject matters differ in their content and learning requirements for teachers and therefore 2) how professional development strategies need to differ according to subject areas so that 3) such programs can ultimately help to improve the pedagogical skills that teachers employ in their classrooms (this last phrase ties together the first two).

Subdomain 2: Pedagogical Content Knowledge.

This figure shows the initial scoring rubric that content experts evaluated in their comments.

Figure 3. Proposed Rubric for Pedagogical Content Knowledge.

Pedagogical Content Knowledge
Definition: includes any responses in which the principal mentions, expresses or demonstrates some knowledge of ANY of the following aspects of pedagogical content knowledge covered in the classroom:

- effective teaching strategies for different subject areas
- how students learn differently in various subjects (for example, mathematics involves applying in some form the agreed-upon concepts, symbols, and problem-solving strategies, while literacy can range in content from learning to writing to evaluating others' compositions)
- how teacher knowledge can differ across subject areas because of their difference in content
- how teaching strategies for different subject areas may differ because of their differing content

0. No Mention of the Dimension

1. A Little
Mere mention of one or two aspects of pedagogical content knowledge with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

Specific example of a mere mention of pedagogical content knowledge: "It's not like in math, with set rules and problems. You've got to cover so much more in the reading program."

2. Some

Mentions at least three or more different aspects of pedagogical content knowledge but does not develop any of the aspects.

3. Sufficient
Mentions at least one aspect of pedagogical content knowledge and develops at least one aspect. This means the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, a more developed discussions of pedagogical content knowledge should include multiple details in the discussion as well as an explanation of why the approach is valuable or important.)

> Specific example of a single aspect of pedagogical content knowledge that is developed: Students have to be given time to read to each other and in small groups. Students should be placed in heterogeneous reading groups so they can listen to each other and share and discuss the book with each other. Parent volunteers or co-teachers can help with the reading groups and the teacher needs to work with each group weekly to listen to them and provide commentary. The teacher must read a book to the class (usually a book above their grade level). The teacher will lead discussions and ask students to visualize, predict and share their feelings about these stories.

4. Quite a Bit
Mentions at least two aspects of pedagogical content knowledge and develops two or more; that is, the response goes beyond mentioning the aspects to developing them with more discussion that suggests a deeper understanding of the aspects.

5. A Great Deal
Mentions at least two aspects of pedagogical content knowledge and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of effective pedagogical content knowledge to making a link or connection between at least two aspects. For example, a principal may discuss 1) how subject matters differ in their content and learning requirements for teachers and therefore 2) how professional development strategies need to differ according to subject areas so that 3) such programs can ultimately help to improve the pedagogical skills that teachers employ in their classrooms (this last phrase ties together the first two).

Responses to the survey items for this subdomain were limited, with only Expert 3 responding to all of the questions. Experts 2 and 3 disagreed significantly on the definitions. Expert 2 "mostly disagreed" that the rubric provided a clear definition for this subdomain (1.a) and "mostly agreed" that it needed additional dimensions (1.b). Expert 3 on the other hand "mostly agreed" that the rubric provided clear definitions (1.a) and did not agree that additional dimensions were needed (1.b). Expert 3 agreed that the directions and explanations were clear. In light of the mixed and limited responses to this I relied more heavily on the scoring evidence and their written comments to guide my rubric modifications.

Table 10. Pedagogical Content Knowledge Feedback Responses			
	Expert		
Question Summary	**1**	**2**	**3**
1. a. rubric provides clear definition	?	2	4
1. b. definition needs additional dimensions	?	4	2
1. c. definition needs fewer dimensions	?	?	2
2. a. directions provide clear guidance	?	?	4
2. b. rubric provides clear explanation of what response qualifies for each level	?	?	4
2. c. rubric provides clear examples for each level	?	?	3

For this subdomain Expert 2 provided the most extensive written comments about the definitions in this rubric.

> I had difficulty getting a good feel for this category and again, I think an overview would help. I think of PCK as the knowledge that educators need of the subject matter to teach it/evaluate it at particular grade levels and the knowledge they also need about how kids at particular grade levels typically think about and interact with the content.

Expert 3 also commented that "I would have liked more examples of each rating. Even if they were just a sentence that encapsulated the idea, not needing to be paragraphs." I discuss my responses to these recommendations at the end of the section.

Table 11 shows the experts' actual scores for the principal responses, and I discuss how these demonstrated that experts agreed very little in their scoring.

Table 11. Experts' Scores for Pedagogical Content Knowledge Responses	Expert		
Scenario	1	2	3
Principal 1			
1	?	1	2
2	?	0	3
3	?	0	1
4	?	0	0
5	?	0	1
Principal 2			
1	3	2	1
2	3	2	1
3	0	1	0
4	0	1	0
5	0	0	0
Principal 3			
1	0	0	0
2	0	3	0
3	1	0	0
4	1	0	0
5	0	0	0

Analysis of these experts' scores of the responses showed large differences in their scores; there was very little consistency in how they rated the responses. Scenarios 1 and 2 generated the largest differences in scores; experts differed in how they scored principals' reactions to a scenario where student math scores for poor students have decreased after a school has adopted a new math curriculum (1) and the school's reading test scores are below the district level's (2). For example, principal 2 wrote the following excerpt in response to scenario 1.

> The above situation is premised in the belief that stagnant or decreasing mathematics achievement is based on the math program and not the instructional approach &/or other relevant factors to student performance. That being said, there is surely considerable data that is available to support the decision making process. I would first look at the historical achievement realized by students in the different math

> classrooms. This might reveal significant patterns that point to variances among teachers as opposed to variances found in curricular approaches.

As shown in the table above, all three experts differed in their scores (3, 2, and 1) when the principal only offered a superficial reference to the instructional strategies for teaching math. Trends in scores for different experts were hard to establish; each expert did not consistently score responses higher or lower than the others. These disagreements in scores raised significant questions about how clear the rubrics were in their guidance to the participants. In light of these results and the written comments above, I summarize the extensive changes I made to the rubric definitions and examples in the paragraph below.

Changes to the Rubric for "Pedagogical Content Knowledge"

In response to Expert 2's call for a better "overview" of this subdomain I made the following additions to the definition. In the introduction I drew a distinction between this subdomain and "subject matter," and I clarified what specifically this area referred to beyond subject matter.

> This area of expertise focuses on the teaching and evaluation skills that teachers use to successfully help children learn subject matter. In contrast to 'subject matter,' responses that qualify for this category emphasize the skills or strategies needed to teach content or evaluate how well students are learning the content.

I also elaborated in the bullet points that this area includes knowledge of evaluation strategies for subject matter: "effective assessment strategies for different subject areas." Finally, in response to Expert 3 I included examples (and non-examples as in the level of "no mention") for scoring levels 1, 3, and 5 that explained superficial ("a little"), more developed ("sufficient"), and highly developed ("a great deal") discussions of the subdomains. The figure below

shows the revised rubric that highlights how I have added the changes above to the rubric.

Figure 4. Modified Rubric for Pedagogical Content Knowledge

Pedagogical Content Knowledge

Definition: ***This area of expertise focuses on the teaching and evaluation skills that teachers use to successfully help children learn subject matter. In contrast to "subject matter," responses that qualify for this category emphasize the skills or strategies needed to teach content or evaluate how well students are learning the content.*** This area includes any responses in which the principal mentions, expresses or demonstrates some knowledge of ANY of the following aspects of pedagogical content knowledge covered in the classroom as they relate to specific subject matter.

- effective teaching strategies for different subject areas
- ***effective assessment strategies for different subject areas***
- how students learn differently in various subjects (for example, mathematics involves applying in some form the agreed-upon concepts, symbols, and problem-solving strategies, while literacy can range in content from learning to write to evaluating others' compositions)
- how teacher knowledge can differ across subject areas because of their difference in content
- how teaching strategies for different subject areas may differ because of their differing content

0. No Mention of the subcategory at all in the response ***(Comments that discuss generic teaching skills or evaluation strategies without connecting them to specific subject matter would count as a "no mention.")***

1. A Little Discussion of the subcategory in the response

Mere mention of one or two aspects of pedagogical content knowledge with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

> Specific example of a mere mention of pedagogical content knowledge: "It's not like in math, with set rules and problems. You've got to cover so much more in the reading program."

2. Some Discussion of the subcategory in the response

Mentions at least three or more different aspects of pedagogical content knowledge but does not develop any of the aspects. ***(For example, a principal might briefly list different evaluation strategies for reading, math, and science classes but provide little or no discussion of how these were appropriate to their subjects.)***

3. Sufficient Discussion of the subcategory in the response

Mentions at least one aspect of pedagogical content knowledge and develops at least one aspect. This means the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, a more developed discussion of pedagogical content knowledge may include multiple details about a strategy or evaluation in the discussion as well as an explanation of why the approach is appropriate for a certain subject area.)

> ***In this example of a principal offers a more developed discussion of how small groups can be used effectively in reading:*** "Students have to be given time to read to each other and in small groups. Students should be placed in heterogeneous reading groups so they can listen to each other and share and discuss the book with each other. Parent volunteers or co-teachers can help with the reading groups and the teacher needs to work with each group weekly to listen to them and provide commentary. The teacher must read a book to the class (usually a book above their grade level). The teacher will lead discussions and ask students to visualize, predict and share their feelings about these stories."

4. Quite a Bit of Discussion of the subcategory in the response

Mentions at least two aspects of pedagogical content knowledge and develops two or more; that is, the response goes beyond mentioning the aspects to developing them with more discussion that suggests a deeper understanding of the aspects.

5. A Great Deal of Discussion of the subcategory in the response

Mentions at least two aspects of pedagogical content knowledge and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of effective pedagogical content knowledge to making a link or connection between at least two aspects.

> ***For example, a principal may discuss in detail 1) how subject matters differ in the teaching strategies that are most effective for each and 2) why different evaluation strategies should therefore be used for each and 3) how principals who recognize these differences and discuss such content specific strategies can best help their teachers improve their teaching (this last phrase ties together the first two).***

Subdomain 3: Teachers as Learners.

The figure below shows the proposed scoring rubric that content experts evaluated in their comments; I summarize their comments and detail how I used their feedback to modify the rubric.

Figure 5. Proposed Rubric for Teachers as Learners.

Teachers as Learners

Definition: includes any responses in which the principal mentions, expresses or demonstrates some knowledge of ANY of the following aspects that pertain to viewing teachers as learners and encouraging their continued learning and professional development:

- how differences in subject area might or can influence professional development needs for teachers (for example, mathematics professional development be tightly organized around specific topics in an adopted curriculum while a literacy program may focus on the theories of learning inherent in a program rather than specific reading or writing content)
- subject-specific effective professional development strategies for teachers

Scoring Guidelines: Assign these scores based on how well a principal's answer includes the following components.

0. No Mention of the Dimension

1. A Little

Mere mention of one or two aspects of teachers as learners with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

> Specific example of a mere mention of the nature of teachers as learners: "The principal needs to meet with each individual grade level team and ask: 'How do you think your children learn best?' 'Does it seem to be working when it relates to reading and math grades?' 'Do you need any other instructional device or training to help you help your kids?' 'What can I do to help you help your students succeed?'"

2. Some

Mentions at least three or more different aspects of teachers as learners but does not develop any of the aspects.

3. Sufficient

Mentions at least one aspect of teachers as learners and develops at least one aspect. This means the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, a more developed discussion of teachers as learners should include multiple details in the discussion as well as an explanation of why the approach is valuable or important.)

> Specific example of a single aspect of teachers as learners that is developed: "It appears the math teachers need to understand better how to teach the basic skills where students are failing. Is it subtraction or multiplication or something else? There are professional development programs that target different areas. We'll need to get our teachers into these particular programs based on where they need to improve their skills."

4. Quite a Bit

Mentions at least two aspects of teachers as learners and develops two or more; that is, the response goes beyond mentioning the aspects to developing them with more discussion that suggests a deeper understanding of the aspects.

5. A Great Deal

Mentions at least two aspects of teachers as learners and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of effective teachers as learners to making a link or connection between at least two aspects. For example, a principal may discuss 1) how subject matters differ in their content and learning requirements for teachers and therefore 2) how professional development strategies need to differ according to subject areas so that 3) such programs can ultimately help to improve the pedagogical skills that teachers employ in their classrooms (this last phrase ties together the first two).

In this subdomain, experts' critical comments in the survey items focused on the definitions and the examples provided. While Experts 2 and 3 disagreed with each other about whether or not the definitions were clear (1a), they "mostly agreed" and "strongly agreed" respectively that the definitions needed additional dimensions. Expert 3 was the most critical of the definitions and the examples in the rubric (items 1a and 2c); she strongly disagreed that these were clear in the rubric. Expert 2 also "mostly disagreed" that the rubric examples were clear for each level (2c). I used their written comments below to make additions to the definitions and examples in this rubric.

Table 12. Teachers As Learners Feedback Responses			
	Expert		
Question Summary	**1**	**2**	**3**
1. a. rubric provides clear definition	?	4	1
1. b. definition needs additional dimensions	?	4	5
1. c. definition needs fewer dimensions	?	3	1
2. a. directions provide clear guidance	?	?	4
2. b. rubric provides clear explanation of what response qualifies for each level	?	?	4
2. c. rubric provides clear examples for each level	?	2	1

Expert 1 commented "could math pd focus on theories of learning and literacy pd focus on specific topics?" Expert 1 also wrote

> You may want to expand this beyond PD. You could have an indicator about how the principal conducts their supervision of teachers, i.e. How they give teachers feedback on their teaching.

This comment emphasized how the original rubric definitions focused primarily on principals' theories or comments about professional development. However, teacher learning can take place in a number of additional contexts, such as faculty meetings, department meetings or conversations, principal-led training,

or in one-one conversations with the principal or other teachers. Finally, Expert 3 wrote that the rubric "would have benefited from more examples and/or counterexamples…it would have been helpful to provide counterexamples of the things that don't fit the criteria."

An examination of experts' scores for the principals showed a range in agreement between the raters. Table 13 summarizes their scores, and I discuss the results below.

Table 13. Experts' Scores for Teachers as Learners Responses	Expert		
Scenario	1	2	3
Principal 1			
1	?	4	4
2	?	1	3
3	?	3	3
4	?	0	0
5	?	2	0
Principal 2			
1	0	0	0
2	3	5	4
3	0	3	3
4	4	1	0
5	0	0	0
Principal 3			
1	0	0	4
2	3	3	3
3	1	0	0
4	0	0	0
5	1	0	0

Experts' scores of principal responses showed greater reliability across scenarios 3, 4, and 5, which provided evidence that the rubrics helped them arrive at a substantial level of agreement for these answers. However, Experts 1 and 3 differed greatly in their scores for two scenarios (see Principal 2, scenario 4, and Principal 3, scenario 1, respectively). Closer examination of these scores and the responses illustrated the need to tie teacher learning in this area more closely to

subject matter. For example, Expert 1 rated Principal 2's comments about teachers' general professional development needs and opportunities in scenario 4 (that asked the principal to respond to a teacher who was resistant to classroom observations) as a "4."

> However, the principal of the school is the instructional leader. If he or she is not the instructional leader, then that becomes a professional development opportunity for that leader. This paradigm among teachers must be addressed directly and not skirted. Each year, instructional areas of focus for the school must be developed in conjunction with teacher leadership. These areas of focus will serve as the anchor for all professional development and teacher evaluations. In this instance, using administrative walkthroughs and teacher evaluations would be used with the instructional areas of focus as the driving theme. Moreover, dealing with the most difficult teachers first will send a strong message to the larger instructional culture.

Expert 1's score of a "4" for this principal's descriptions of an overall professional development strategy offered evidence that this expert did not understand this subdomain's connection to subject matter well enough.

Changes to the Rubric for "Teachers as Learners"

In response to Expert 1's first two written comments above I modified the definition with an additional bullet to elaborate on how a subject's nature might influence the type of professional development that is required for the subject.

- subject-specific effective professional development or teaching strategies for teachers (such as the specific concepts that teachers need to learn and understand through professional development in different subject areas)"

I modified the rubric introduction and definitions to emphasize the range of conditions in which teacher learning may occur.

> Relevant discussions regarding teacher learning about subject matter may occur in a variety of contexts, not just traditional professional development conditions. For a principal's discussion of learning to qualify it must discuss teacher learning in relation to a subject area. For

example a principal may discuss how teachers learn through one-one conversations, meetings with fellow teachers, meetings in professional learning communities, or other places.

As a reply to Expert 3's request for more examples I added a counterexample for "no mention" and explained more in detail the examples for scores of 1 (mere mention), 3 (sufficient discussion), and 5 (a great deal of discussion). For example, for "no mention" I added the following counterexample: "a principal's general discussion of professional development without specific reference to a subject area or areas would be too broad to qualify for this subdomain."

These counterexamples demonstrated responses that would *not* qualify for the different levels in the scoring rubric. The figure below highlights the changes made to the rubric based on the comments I have just summarized.

Figure 6. Modified Rubric for Teachers as Learners

Teachers as Learners

Definition: includes any responses in which the principal mentions, expresses or demonstrates some knowledge of ANY of the following aspects that pertain to viewing teachers as learners and encouraging their continued learning and professional development. ***Relevant discussions regarding teacher learning about subject matter may occur in a variety of contexts, not just traditional professional development conditions. For a principal's discussion of learning to qualify it must discuss teacher learning in relation to a subject area. For example a principal may discuss how teachers learn through one-one conversations, meetings with fellow teachers, meetings in professional learning communities, or other places.***

- ***key strategies for encouraging or organizing professional development or training for teachers in different subject areas***
- ***strategies to evaluate teacher learning in different subject areas***
- how differences in subject area might or can influence professional development needs for teachers (for example, a principal might describe how mathematics professional development could be tightly organized around specific topics in an adopted curriculum while a literacy program may focus on the theories of learning inherent in a program rather than specific reading or writing content, or vice versa)
- ***subject-specific effective professional development or teaching strategies for teachers (such as the specific concepts that teachers need to learn and understand through professional development in different subject areas)***

Scoring Guidelines: Assign these scores based on how well a principal's answer includes the following components.

0. No Mention of the subcategory at all in the response

For example, a principal's general discussion of professional development without specific reference to a subject area or areas would be too broad to qualify for this subdomain.

1. A Little Discussion of the subcategory in the response

Mere mention of one or two aspects of teachers as learners with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

Specific example of a mere mention of the nature of teachers as learners. ***Here the principal discusses questions to discuss with the teachers to understand their learning or training needs, but he offers no discussion of how to pursue larger strategies to help them learn:*** "The principal needs to meet with each individual grade level team and ask: 'How do you think your children learn best?' 'Does it seem to be working when it relates to reading and math grades?' 'Do you need any other instructional device or training to help you help your kids?' 'What can I do to help you help your students succeed?'"

2. Some Discussion of the subcategory in the response
Mentions at least three or more different aspects of teachers as learners but does not develop any of the aspects.

3. Sufficient Discussion of the subcategory in the response
Mentions at least one aspect of teachers as learners and develops at least one aspect. This means the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, a more developed discussion of teachers as learners should include multiple details in the discussion as well as an explanation of why the approach is valuable or important.)

> ***Specific example of a single aspect of teachers as learners that is developed—this develops the need for professional development to be tied to teachers' specific needs in math.*** "It appears the math teachers need to understand better how to teach the basic skills where students are failing. Is it subtraction or multiplication or something else? There are professional development programs that target different areas. We'll need to get our teachers into these particular programs based on where they need to improve their skills."

4. Quite a Bit of Discussion of the subcategory in the response
Mentions at least two aspects of teachers as learners and develops two or more; that is, the response goes beyond mentioning the aspects to developing them with more discussion that suggests a deeper understanding of the aspects.

5. A Great Deal of Discussion of the subcategory in the response
Mentions at least two aspects of teachers as learners and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of effective teachers as learners to making a link or connection between at least two aspects.

> ***For example, a principal may discuss 1) how subject matters differ in their content and learning requirements for teachers—what different concepts or strategies they need to understand--and therefore 2) how professional development strategies need to differ according to subject areas to be more successful in training teachers. This will in turn help guarantee that such programs ultimately help to improve the pedagogical skills that teachers employ in their classrooms (this last phrase ties together the first two).***

Learning-centered Leadership Rubrics

Analyses of the scenarios for this area of expertise focused on four different subdomains: effective teaching and learning; data-based decision-making; standards-based thinking; and monitoring teachers for instructional improvement. For each domain in their 2008 and 2009 pieces, Goldring, et. al's research team developed a coding rubric which incorporated content from 1) a theoretical review of the literature to identify areas of organizational expertise that principals employ in their leadership of a school, 2) items on a principal survey they developed to measure principal expertise and 3) an analysis of the content in a particular professional development program they were evaluating at the time. Through these analyses the team identified central areas of learning-centered leadership expertise, and they established scoring rubrics with

examples for each domain (see Appendices F, G, H for the final scoring rubrics used in this dissertation).

The specific components of "effective teaching and learning" used in this study included not only learning theory or knowledge of curriculum and how they work but also empirically based classroom conditions that foster student learning. Unlike Stein and Nelson's (2003) area of subject matter, this subdomain included those conditions that best facilitate learning, such as cooperation between students, ample time being given to them to work, or the importance of students making connections between the different things they learn. This category also applied more broadly than Stein and Nelson's pedagogical content knowledge because it referred to teaching strategies that can apply across subject areas. While subject-specific teaching strategies may be part of this, this subdomain captured pedagogical strategies that were not subject-specific.

With "data-based decision making" Goldring and Berends (2008) have discussed the different facets of school leaders' use of data as they evaluate conditions in their schools and make decisions and plans to address their students' pressing needs. A school leaders' expertise in this domain includes not only knowledge of the different forms of data at their disposable (such as types of student assessments, reports of student achievement, or summaries of teacher or student characteristics) but strategies to analyze these data. Finally, it can include processes by which principals use their findings to make decisions or formulate plans.

"Monitoring instructional improvement" refers to those strategies by which a principal evaluates a teacher's pedagogy and use of curriculum in his or her classroom. This subdomain included techniques for observing teachers or

evaluating the curriculum in their classrooms (formal or enacted). It also included ways to evaluate the progress a teacher is making in trying a new idea or in improving his existing skills in a specific area. Finally, this subdomain referred to any evaluations by which a teacher might check for alignment between a teacher's practices and the school improvement plan.

Finally, the subdomain "standards-based thinking" captured principals' expertise in the use of learning standards to guide curriculum design and implementation. This subdomain included first and foremost a principal's knowledge of the curriculum standards or expectations for different grade levels. It also referred to strategies to align curriculum with broader learning standards as well as establish alignment between instructional strategies with standards.

Learning-centered Leadership Results.

Subdomain 1: Data-based Decision Making.

Figure 7 shows the initial scoring rubric that content experts evaluated in their comments, and I explain below how I used their comments to make changes to this rubric.

Figure 7. Proposed Rubric for Data-based Decision Making

Data-based Decision Making

Dimensions of data based decision-making referred to in the scale below include but are NOT limited to:

- Information sources, data collection, and data analysis strategies
- Different types of student assessment (e.g., using portfolio and other qualitative methods of assessment, using formative/diagnostic as well as evaluative, and so on)
- Data or information of various sorts (e.g., student achievement data, local demographic data, teacher demographic data, classroom observation data, etc.)
- Data-based decision making
- Evaluation and assessment strategies
- Evidence-based procedures for assessing struggling or low achieving students

0. No Mention

1. A Little

Mere mention of one or two aspects of data based decision-making (mentions any one of the dimensions or a RELATED dimension). *NOTE: saying the same thing 10 times is still a mere mention.*

2. Some.
Mentions at least three aspects of data based decision-making (mentions at least three of the dimensions or RELATED data based decision-making dimension).

3. Sufficient
Mentions at least one aspect of data based decision-making and develops at least one aspect; that is, the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, the respondent might mention data based decision making and go on to talk about using multiple measures of student achievement. Or, the respondent might mention that decisions need to be based on data and go on to note that more than student assessment data should be used in this process.)

4. Quite a Bit
Mentions at least two aspects of data based decision-making and develops two or more; that is, the response goes beyond mentioning an aspect to develop it suggesting a deeper understanding.

5. A Great Deal
Mentions at least two aspects of data based decision-making and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of data based decision-making thinking to making a link or connection between at least two aspects. For example, the respondent might mention data-based decision making and using student achievement data and classroom observation data and how by looking at classrooms where students do well one might be able to identify best practices.

As shown in Table 14 below content experts did not consistently raise significant concerns about the structure, definitions, and examples in the rubric for data-based decision making. When asked about these issues, all three experts responded that they "mostly agree" that the "scoring rubric offers a clear definition for the subdomain" (question 1a); two of the three responded that they "mostly agree" that the directions provide "clear guidance," "clear explanation," and "clear examples" about how to use the rubric (questions 2a, 2b, and 2c, respectively). A third expert reported that he "mostly disagrees" that the rubric provides clear directions and explanations (2a and 2b) but provided few if any comments for changing the rubrics.

Table 14. Data-based Decision Making Feedback Responses			
	Expert		
Question Summary	**1**	**2**	**3**
1. a. rubric provides clear definition	4	4	4
1. b. definition needs additional dimensions	4	4	2
1. c. definition needs fewer dimensions	2	4	4
2. a. directions provide clear guidance	2	4	4
2. b. rubric provides clear explanation of what response qualifies for each level	2	4	4
2. c. rubric provides clear examples for each level	3	4	4

Few of the experts offered detailed comments for this subdomain. Expert 1 wrote that "your rubric is built on three important distinctions: 'mention,' 'develop,' and 'connect.' Fully defining these will help." Expert 2 asked for "more specificity for what constitutes 'understanding'" of data-based decision making. She also commented that "having examples and non-examples would add clarity" to the rubrics.

Despite the limited comments for this section, a review of the experts' scores of the responses showed significant differences across experts. Table 15 summarizes these scores.

Table 15. Experts' Scores for Data-based Decision Making Responses	Expert		
Scenario	1	2	3
Principal 1			
1	2	4	5
2	1	4	3
3	1	4	3
4	0	2	1
5	1	3	1
Principal 2			
1	1	1	0
2	2	5	?
3	1	2	?
4	0	1	?
5	1	4	?

Principal 3			
1	1	1	?
2	0	1	?
3	1	1	?
4	0	1	?
5	1	1	?

Expert 1 consistently scored the answers lower than the other two but provided limited written comments to explain his views or codes. For example, principal 1 for scenario 4 (where the principal responded to a teacher resisting observation) wrote the following response.

> I feel it may be time for the principal and assistant principal to take a good hard look in a mirror (just like you want your teachers to do) and formally assess how we can approach our work differently so we are building organizational capacity for student learning rather than anger. It may be necessary for the principal to become a really good listener and slowly move one staff member at a time to thinking about the power of opening their classroom door to colleagues.
>
> --The principal and assistant principal may want to have someone observe their conferences with teachers and give them feedback.
> --Share research on what good schools do to get better.
> --Move to establish a professional learning community through learning team work.
> --Try to stay as positive as possible during this very difficult time.
> --Talk to colleagues from other schools that seem to have a building without this type of closed door attitude.
> --Critically reflect on why your work with staff is not having the results you want.
> --Just as you don't want your teachers to blame the kids or parents, it will be important that the principal not blame the teachers.
> --Be honest with key staff members that you want to improve your monitoring strategies and see what suggestions they can offer.

In this response the principal superficially mentions formal assessment in the first paragraph and different forms of data to collect (feedback about their observations and advice from other colleagues), and Expert 1 did not catch these mentions at all while the other two scored them as mere mentions.

For other scenarios experts 2 and 3 showed some agreement on the scores (though 3 did not complete all the scores), but these different disagreements emphasized the importance of improving the rubrics to help scorers develop a more common understanding of this subdomain and scoring for each of its levels.

Changes to the Rubric for Data-based Decision-making

After a review of all the information above I made the following changes to the rubric. I addressed both of Expert 2's comments by first specifying what a "mere mention" or more superficial answer might entail:

> Mere mention of <u>one or two aspects</u> of data based decision-making (mentions <u>any one</u> of the dimensions or a RELATED dimension). *NOTE: saying the same thing 10 times is still a mere mention.* (For example, a respondent might refer multiple times to the need to "look at the data" before making a decision, but he or she may not provide specific examples of what data to examine or how to analyze it).

I then provided a more detailed discussion of an example response that demonstrated a principal's deeper understanding of this subdomain for scoring level 3:

> This deeper understanding is evidenced by greater details about how to pursue a particular aspect. These details demonstrate that the principal understands how to analyze data to evaluate a situation. (For example, the respondent might mention data-based decision making and go on to discuss *how to analyze* multiple measures of student achievement. Or, the respondent might mention that decisions need to be based on data and then go on to discuss what specific information beyond student assessment data should be used in this process.)

I also provided a "non-example" for the scoring level of "0":

> For example, principal may discuss making a decision about a math program or professional development strategy with no discussion of examining student achievement data to inform the decision.

Finally, I provided a greater explanation of an example that would qualify for a "5" (a great deal of discussion) in this category.

> For example, the respondent might first discuss in detail the process of analyzing specific student achievement data and second how to review corresponding classroom observation data to corroborate the student achievement data results. She might then describe how by looking at classrooms where students do well one might be able to identify best teaching practices.

The figure below shows the changes made to this rubric.

Figure 8. Modified Rubric for Data-based Decision Making

Data-based Decision Making

Aspects of data based decision-making referred to in the scale below include but are NOT limited to:

- Information sources, data collection, and data analysis strategies
- Different types of student assessment (e.g., using portfolio and other qualitative methods of assessment, using formative/diagnostic as well as evaluative, and so on)
- Data or information of various sorts (e.g., student achievement data, local demographic data, teacher demographic data, classroom observation data, etc.)
- ***Use of data or information to make decisions regarding school matters***
- Evaluation and assessment strategies
- Evidence-based procedures for assessing struggling or low achieving students

0. No Mention of the subcategory at all in the response
For example, principal may discuss making a decision about a math program or professional development strategy with no discussion of examining student achievement data to inform the decision.

1. A Little Discussion of the subcategory in the response
Mere mention of one or two aspects of data based decision-making (mentions any one of the dimensions or a RELATED dimension). *NOTE: saying the same thing 10 times is still a mere mention.* ***(For example, a respondent might refer multiple times to the need to "look at the data" before making a decision, but he or she may not provide specific examples of what data to examine or how to analyze it).***

2. Some Discussion of the subcategory in the response
Mentions at least three aspects of data based decision-making (mentions at least three of the dimensions or RELATED data based decision-making dimension).

3. Sufficient Discussion of the subcategory in the response
This deeper understanding is evidenced by greater details about how to pursue a particular aspect. These details demonstrate that the principal understands how to analyze data to evaluate a situation. (For example, the respondent might mention data-based decision making and go on to discuss how to analyze multiple measures of student achievement. Or, the respondent might mention that decisions need to be based on data and then go on to discuss what specific information beyond student assessment data should be used in this process.)

4. Quite a Bit of Discussion of the subcategory in the response
Mentions at least two aspects of data based decision-making and develops two or more; that is, the response goes beyond mentioning an aspect to develop it suggesting a deeper understanding.

5. A Great Deal of Discussion of the subcategory in the response
Mentions at least two aspects of data based decision-making and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of data based decision-making thinking to making a link or connection between at least two aspects. ***For example, the respondent might first discuss in detail the process of analyzing specific student achievement data and second how to review corresponding classroom observation data to corroborate the student achievement data results. She might then describe how by looking at classrooms where students do well one might be able to identify best teaching practices.***

Subdomain 2: Effective Teaching and Learning.

The figure below shows the initial scoring rubric that content experts evaluated in their comments.

Figure 9. Proposed Rubric for Effective Teaching and Learning

Effective Teaching and Learning

Dimensions of teaching and learning referred to in the scale below include but are NOT limited to:

- student and/or teacher effort produces achievement,
- student learning is about making connections,
- students learn with and through others,
- student learning takes time,
- student and teacher motivation is important to effective teaching and student learning,
- focused teaching promotes accelerated learning,
- clear expectations and continuous feedback to students and/or teachers activate student learning (this does not include the process of monitoring instruction in classrooms),
- good teaching builds on students strengths and respects individual differences,
- good teaching involves modeling what students should learn
- general references to teachers' use of effective teaching and learning practices (this includes discussions of teachers' use of best practices)

Other dimensions might include but are not limited to:

- cognitively or developmentally appropriate or challenging curriculum for students
- applied learning theory
- individualized instruction
- reciprocal teaching
- inquiry teaching or direct instruction

** Note: pay careful attention to discussions of more than one teacher; these may relate more to systemic changes in curriculum that relate more directly to the "standards-based reform/systems thinking."*
*** Note: in situations that discuss professional development or teacher cooperation/collaboration there must be strong, explicit, specific references to effective teaching and learning strategies before it fits under effective teaching and learning.*

0. No Mention

1. A Little

Mere mention of <u>one or two aspects</u> of effective teaching and/or learning with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

2. Some.

Mentions at least <u>three or more different aspects</u> of effective teaching and learning but does not develop any of the aspects.

3. Sufficient

Mentions at least <u>one</u> aspect of effective teaching and learning and develops at least <u>one aspect</u>; that is, the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, the respondent might mention effective instructional strategies in reading and say teachers need to use "writing workshop" or "balanced literacy." Or, the respondent might mention evidence based teaching or assessment and go on to note trying to figure out the strategies that teachers use who have high performing students).

****Note: More developed discussions of effective teaching and learning need to include multiple details in the discussion as well as an explanation of why the approach is valuable or important*
Specific example of single aspect (individualized instruction) that is developed: *"Students must have pre assessment in the critical areas of reading such as vocabulary, phonics, fluency, comprehension, etc. Teachers must know the basic reading levels of their students. Instruction must be tailored to meet these specific needs."*

4. Quite a Bit

Mentions at least two aspects of effective teaching and learning and develops two or more; that is, the response goes beyond mentioning the aspects to developing them with more discussion that suggests a deeper understanding of the aspects.

5. A Great Deal
Mentions at least two aspects of effective teaching and learning and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of effective teaching and learning to making a link or connection between at least two aspects. For example, the respondent might mention and develop how student motivation is critical and then link it to how student effort produces achievement rather than IQ alone. A second example could be that a principal develops 1) how to determine if teachers are using best practices in their teaching, and 2) the importance of using individualized instruction, and she/he then connects them by discussing how individualized instruction should be included as a part of best practices.

For this subdomain experts again reported few concerns in their survey responses. As seen in Table 16, all three experts reported that they "mostly agree" that the rubric offers a "clear definition" for the subdomain, and two of the three reported that they "mostly agree" that the rubric offers "clear guidance" in the directions and "clear examples" of responses that qualify for each level (items 1a, 2b, and 2c). Expert 1 offered lower responses for items 1a, 2b, and 2c but provided limited written comments to clarify the scores (see below).

Table 16. Effective Teaching and Learning Feedback Responses			
	Expert		
Question Summary	**1**	**2**	**3**
1. a. rubric provides clear definition	4	4	4
1. b. definition needs additional dimensions	4	4	2
1. c. definition needs fewer dimensions	2	4	4
2. a. directions provide clear guidance	2	4	4
2. b. rubric provides clear explanation of what response qualifies for each level	2	4	4
2. c. rubric provides clear examples for each level	3	4	4

Experts' detailed comments focused on two aspects of the definitions. Expert 1 commented that "you claim that student learning is about making

connections—connections with what? Whom?" Expert 2 questioned the rubric's lack of discussion of assessment:

> I'm wondering if there's any place for assessment here. You write that this does not include monitoring instruction, but that is different than formative assessment in the classroom, which I would argue is a very important part of teaching & learning. Your caution about not including monitoring leaves open the question about including classroom-based assessment.

This comment pointed to an important oversight in the definition—knowledge of effective assessment strategies by a teacher is an integral part of effective teaching and learning.

As seen in the summary below in Table 17, experts' scoring of the example responses for this subdomain showed much greater agreement than for "data-based decision making." In this subdomain experts similarly scored Principal 1 higher in the first two scenarios and lower in the last three. This principal frequently offered more detailed thoughts about the complexity of improving math instruction such as the following (from scenario 1):

> If this math program is attempting to get kids to think about numbers and not just memorize facts and spit back information then it makes sense that it would be a difficult transition for teachers and for students. This type of learning and instruction is complex and places complex demands on teachers and students.

In addition Experts 1 and 2 scored the last two principals consistently lower. For scenario 1 Principal 2 offered only this brief mention of additional curriculum that may be helpful for teachers, and all the experts identified this as a superficial discussion of "effective teaching and learning:"

> I would also encourage all the teachers to look for some supplemental materials that we could use to "fill-in" the skills areas that they felt were lacking from the new series. This may be different materials depending on the grade levels.

Except for one score of "3" that were significantly higher (see Expert 2 and 3's scores) the agreements here provided initial evidence that the rubrics helped experts develop similar conceptions and scores for this subdomain.

Table 17. Experts' Scores for Effective Teaching and Learning Responses	Expert		
Scenario	1	2	3
Principal 1			
1	4	2	4
2	3	3	4
3	0	1	0
4	0	0	0
5	0	1	3
Principal 2			
1	0	1	1
2	0	2	?
3	0	1	?
4	0	0	?
5	0	0	?
Principal 3			
1	0	0	?
2	1	1	?
3	0	1	?
4	0	1	?
5	0	0	?

Changes to the Rubric for Effective Teaching and Learning

In response to Expert 1's comments about "connections" I modified the rubric to specify more closely that an aspect of effective teaching and learning includes "student learning is about making connections *between different concepts and skills that they learn* (italics mine)." After Expert 2 questioned the lack of discussion of assessments I added the following component to the definition: "cognitively or developmentally appropriate assessment strategies." The figure below highlights the changes I made to the rubric.

Figure 10. Modified Rubric for Effective Teaching and Learning

Effective Teaching and Learning Scenario Coding Rubric

Aspects of teaching and learning referred to in the scale below include but are NOT limited to:

- student and/or teacher effort produces achievement,

- ***student learning is about making connections between different concepts and skills that they learn,***
- students learn with and through others,
- student learning takes time,
- student and teacher motivation is important to effective teaching and student learning,
- focused teaching promotes accelerated learning,
- clear expectations and continuous feedback to students and/or teachers activate student learning (this does not include the process of monitoring instruction in classrooms),
- good teaching builds on students strengths and respects individual differences,
- good teaching involves modeling what students should learn
- general references to teachers' use of effective teaching and learning practices (this includes discussions of teachers' use of best practices)

Other dimensions might include but are not limited to:

- cognitively or developmentally appropriate or challenging curriculum for students
- ***cognitively or developmentally appropriate assessment strategies to evaluate student learning***
- applied learning theory
- individualized instruction
- reciprocal teaching
- inquiry teaching or direct instruction

** Note: pay careful attention to discussions of more than one teacher; these may relate more to systemic changes in curriculum that relate more directly to the "standards-based reform/systems thinking."*
*** Note: in situations that discuss professional development or teacher cooperation/collaboration there must be strong, explicit, specific references to effective teaching and learning strategies before it fits under effective teaching and learning.*

0. No Mention of the subcategory at all in the response
For example a respondent might summarize teaching strategies he has observed without offering an opinion of them or discussing why these are effective.

1. A Little Discussion of the subcategory in the response
Mere mention of one or two aspects of effective teaching and/or learning with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention. ***For example a principal may discuss briefly the need for "good teaching" or the importance of setting "clear expectations" but then provide no details about what such actions would entail.***

2. Some Discussion of the subcategory in the response
Mentions at least three or more different aspects of effective teaching and learning but does not develop any of the aspects.

3. Sufficient Discussion of the subcategory in the response
Mentions at least one aspect of effective teaching and learning and develops at least one aspect; that is, the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, the respondent might mention effective instructional strategies in reading and say teachers need to use "writing workshop" or "balanced literacy." Or, the respondent might mention evidence based teaching or assessment and go on to note trying to figure out the strategies that teachers use who have high performing students).
****Note: More developed discussions of effective teaching and learning need to include multiple details in the discussion as well as an explanation of why the approach is valuable or important*

> ***Example of single aspect (individualized instruction) that is developed (in this case the principal discusses specific steps to take in implementing more individualized instruction for students):*** *"Students must have pre assessment in the critical areas of reading such as vocabulary, phonics, fluency, comprehension, etc. Teachers must know the basic reading levels of their students. Instruction must be tailored to meet these specific needs."*

4. Quite a Bit of discussion of the subcategory in the response
Mentions at least two aspects of effective teaching and learning and develops two or more; that is, the response goes beyond mentioning the aspects to developing them with more discussion that suggests a deeper understanding of the aspects.

5. A Great Deal of discussion of the subcategory in the response
Mentions at least two aspects of effective teaching and learning and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of effective teaching and learning to making a link or connection between at least

two aspects. For example, the respondent might mention and develop how student motivation is critical to student learning and then link it to how student effort produces achievement rather than IQ alone. A second example could be that a principal develops 1) how to determine if teachers are using best practices in their teaching, and 2) the importance of using individualized instruction, and she/he then connects them by discussing how individualized instruction should be included as a part of best practices.

Subdomain 3: Monitoring Instructional Improvement.

Figure 11 shows the proposed rubric that content experts evaluated in their comments; I summarize their comments and then explain how I used their feedback to modify the rubric.

Figure 11. Proposed Rubric for Monitoring Instructional Improvement.

Monitoring Instructional Improvement

Dimensions of monitoring instructional improvement referred to in the scale below include but are NOT limited to:

- Benchmarking: setting teacher performance levels and evaluating teacher progress toward those
- Procedures for monitoring teachers
- Observing a teacher who was trying new instructional practices or using new curricular materials
- Monitoring the curriculum used in classrooms to see that it reflects the school's improvement efforts
- Monitoring classroom instructional practices to see if they reflect the school's improvement efforts

These codes do not include descriptions of coaching or mentoring, in which more a more knowledgeable professional observes and models instruction and offers advice or feedback. This also does not include collaboration, in which a principal might help teachers work together or coordinate time to share ideas and information.

Examples:

"*I would make sure teachers were aware of the evaluation process and of our intention to closely monitor the academic progress of students*."

"*I would first determine if the new science program was even being used by teachers. To do this I would drop in on classrooms to observe on a regular basis, and would have my science specialists do the same*."

In Example 1, there is an explicit reference to monitoring – "intention to closely monitor." In Example 2, although the word monitoring is not used, this is clearly what the respondent intends. The respondent proposes to monitor science teaching to see if a new science program is being used in the classroom.

0. No Mention

1. A Little

Mere mention of one or two aspects of monitoring instructional improvement (mentions any one of the dimensions or a RELATED dimension). NOTE, saying the same thing 10 times is still a mere mention.

2. Some.

Mentions at least three aspects of monitoring instructional improvement (mentions at least three of the dimensions or a RELATED monitoring instructional improvement dimensions).

3. Sufficient

Mentions at least one aspect of monitoring instructional improvement and develops at least one aspect; that is, the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, the respondent might mention monitoring instructional improvement and go on to discuss specific conditions or

criteria for which she or he might look in the classroom. Or, the respondent might discuss monitoring conditions in a classroom and then elaborate on how these relate to the school's larger improvement efforts.)

4. Quite a Bit
Mentions at least two aspects of monitoring instructional improvement and develops two or more; that is, the response goes beyond mentioning an aspect to develop it suggesting a deeper understanding.

5. A Great Deal
Mentions at least two aspects of monitoring instructional improvement and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of monitoring instructional improvement thinking to making a link or connection between at least two aspects. For example, the respondent might discuss 1) setting teaching performance levels and 2) specific procedures for monitoring instructional improvement toward those levels, and then she might explain how these steps help to promote the overall school improvement plan.

As shown in Table 18 all three experts again responded that they "mostly agree" that the rubric offers "a clear definition" for the subdomain (1a) along with clear directions and examples (2a and 2c). All three "mostly disagreed" that there needed to be fewer dimensions in the rubric (item 1c). Experts 2 and 3 did, however, say that they "mostly agreed" or "completely agreed" that the definition needed additional dimensions. I examined the written comments below for further directions regarding additional dimensions.

Table 18. Monitoring Instructional Improvement Feedback Responses			
	Expert		
Question Summary	**1**	**2**	**3**
1. a. rubric provides clear definition	4	4	4
1. b. definition needs additional dimensions	2	4	5
1. c. definition needs fewer dimensions	2	2	2
2. a. directions provide clear guidance	4	4	4
2. b. rubric provides clear explanation of what response qualifies for each level	3	4	4
2. c. rubric provides clear examples for each level	4	4	4

Expert 1 expressed concern that the definition was too unclear about a principal stating high expectations:

> If you believe that simply stating expectations is not enough to be included in monitoring expectations then be sure to state so. The samples

included lots of examples of principals stating expectations, but there was no action associated with that, so I didn't code it as such.

I modified the definition to specify that a principal's simple description or statement about the need for high expectations did not qualify as an act of monitoring instructional improvement. Expert 2 questioned whether or not "informal monitoring" qualified here: "I know about both informal and formal walkthroughs used by principals—formal is more for evaluation of staff and informal is more about monitoring instruction."

Experts' different scoring of the principals' responses (as seen in Table 19 below) for this subdomain demonstrated the need for changes to the rubric.

Table 19. Experts' Scores for Monitoring Instructional Improvement Feedback Responses	Expert		
Scenario	1	2	3
Principal 1			
1	3	1	0
2	1	1	0
3	1	3	0
4	1	1	0
5	0	3	0
Principal 2			
1	0	1	0
2	1	5	?
3	0	0	?
4	1	1	?
5	1	4	?
Principal 3			
1	3	1	?
2	1	0	?
3	0	0	?
4	0	1	?
5	0	0	?

While experts tended to agree that the definitions and examples were clear they frequently disagreed in their scores. Expert 3 found no instances of "monitoring instructional improvement" for Principal 1 while Experts 2 and 3 had both high

and low evaluations of this principal. An excerpt of this principal's response to scenario 1 demonstrates how the experts viewed the content differently.

> For example, if there is a high degree of capacity built in the school for student learning then I would encourage math teachers to analyze the math data by teacher. This would be very scary for teachers but it might yield the best information. Essentially, teachers would pick a common task/assessment based on the new math program and administer it to students. They would analyze the student work together based on a common set of standards. These would be standards that the students and teachers know well. After scoring the work they would look at the student results and discuss strengths and gaps. This common analysis of student work has the possibility of yielding the best information. I would ask hard questions about what they are learning about themselves as learners and educators as they implement this new program.

As shown in the table above, the experts gave three different scores of 3, 1, and 0 to this response. Although the principal provided no substantive discussion of how to monitor instruction in the math classes, only two experts scored this low, while one expert viewed this as a developed discussion of monitoring instructional improvement. The experts had greater agreement for most of the last two principals, but they still disagreed strongly as in the case of scenarios 2 and 5 for Principal 2. Expert 2 found much greater levels of "monitoring instructional leadership" expertise for this principal than did Expert 3. Thus, while experts tended to rate the rubrics favorably in their survey responses, their scores indicate they had limited agreement in understanding how to apply this subdomain and its components.

Changes to the Rubric for Monitoring Instructional Improvement

In response to Expert 1's comment about simply stating high expectations I modified the definition to specify that a principal's simple description or statement about the need for high expectations did not qualify as an act of

monitoring instructional improvement. After reading Expert 2's concerns about "informal" teaching I added to the definitions knowledge of "procedures for monitoring teachers *formally and informally"* to insure that coders would look for both types of expertise in principals' responses. I also revisited the examples for the different scoring levels to better illustrate types of responses that would qualify for each. The figure below includes highlighted changes to the rubric.

Figure 12. Modified Rubric for Monitoring Instructional Improvement

Monitoring Instructional Improvement Scenario Coding Rubric

Aspects of monitoring instructional improvement referred to in the scale below include but are NOT limited to:

- Benchmarking: setting teacher performance levels and evaluating teacher progress toward those ***(this may include evaluation from outside the classroom through strategies such as examining students' progress in a particular teacher's classroom)***
- Procedures for monitoring teachers formally and informally
- Observing a teacher who was trying new instructional practices or using new curricular materials
- Monitoring the curriculum used in classrooms to see that it reflects the school's improvement efforts
- Monitoring classroom instructional practices to see if they reflect the school's improvement efforts

These codes do not include descriptions of coaching or mentoring, in which a more knowledgeable professional observes and models instruction and offers advice or feedback. These also do not include collaboration, in which a principal might help teachers work together or coordinate time to share ideas and information. ***Finally, cases in which principals describe how they simply state expectations with no evaluation/monitoring do not qualify in this case.***

Examples:

"*I would make sure teachers were aware of the evaluation process and of our intention to closely monitor the academic progress of students*."

"*I would first determine if the new science program was even being used by teachers. To do this I would drop in on classrooms to observe on a regular basis, and would have my science specialists do the same*."

In Example 1, there is an explicit reference to monitoring – "intention to closely monitor." In Example 2, although the word monitoring is not used, this is clearly what the respondent intends. The respondent proposes to monitor science teaching to see if a new science program is being used in the classroom.

0. No Mention at all of the subcategory in the response
For example, a principal may discuss the need to understand what is going on in classrooms but provide no discussion of how to monitor or observe teachers as they work.

1. A Little Discussion of the subcategory in the response
Mere mention of one or two aspects of monitoring instructional improvement (mentions any one of the dimensions or a RELATED dimension). NOTE, saying the same thing 10 times is still a mere mention. ***Here, a principal could state that she observes in a classroom without giving details about how she does this, or she might refer to evaluating curriculum without discussing the criteria she would use.***

2. Some Discussion of the subcategory in the response
Mentions at least three aspects of monitoring instructional improvement (mentions at least three of the dimensions or a RELATED monitoring instructional improvement dimensions).

3. Sufficient Discussion of the subcategory in the response

Mentions at least one aspect of monitoring instructional improvement and develops at least one aspect; that is, the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, the respondent might mention monitoring instructional improvement and go on to discuss specific conditions or criteria for which she or he might look in the classroom. ***Or, the respondent might discuss monitoring conditions in a classroom and then elaborate on how these conditions relate directly to the school's larger improvement efforts.)***

4. Quite a Bit of Discussion of the subcategory in the response
Mentions at least two aspects of monitoring instructional improvement and develops two or more; that is, the response goes beyond mentioning an aspect to develop it suggesting a deeper understanding.

5. A Great Deal of Discussion of the subcategory in the response
Mentions at least two aspects of monitoring instructional improvement and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of monitoring instructional improvement thinking to making a link or connection between at least two aspects. For example, the respondent might discuss in detail 1) setting teaching performance levels and 2) specific procedures for monitoring instructional improvement toward those levels, and then she might explain how these steps help to promote the overall school improvement plan.

Subdomain 4: Standards-based and Systems Thinking.

The figure below shows the initial scoring rubric for "standards-based and systems thinking;" I summarize content experts' comments below and explain how I modified the rubric.

Figure 13. Proposed Rubric for Standards-based and Systems Thinking.

Standards-based and Systems Thinking Scenario Coding Rubric

Dimensions of standards and system thinking referred to in the scale below include but are NOT limited to:

- Standards-based reform
- Standards (e.g., curriculum standards, content standards, learning standards, performance standards, etc.)
- Curriculum design, implementation, evaluation, and refinement
- What students should know and be able to do at any grade level or in any school subject
- Alignment or coherence in general,
- Alignment or coherence in reference to student assessment, curriculum standards, professional development, curricular materials, etc.
- Alignment or coherence of instruction, assessments, and materials.
- Accountability (e.g., holding staff accountable for learning, holding students accountable)
- Systemic reform as it relates to standards or curricula
- Systems theory as it relates to standards or curricula
- The political, social, cultural, and economic systems and processes that impact schools

There are possible overlaps between this code and data-based decision making. We use this general rule: if a principal discusses beginning with data and then moving to curricular decisions, we first consider this as a discussion of data-based decision-making and then look to determine if standards/alignment/systems thinking are also mentioned. If so, this may be a double code.

Discussions of state and national assessments are not in and of themselves standards unless standards are explicitly mentioned. We treat most of these discussions as data-based decision making because they refer to the understanding and use of assessments that include data.

For professional development to be included in this code there must be explicit discussions of standards, alignments, or accountability.

Be careful to look closely at any system, school-wide or community-wide references as a part of the systemic theory or larger systems that influence the school.

Scoring
0. No Mention

1. A Little
Mere mention of one or two aspects of standards and system thinking (mentions any one of the dimensions or a RELATED dimension). NOTE, saying the same thing 10 times is still a mere mention.

2. Some.
Mentions at least three aspects of standards and system thinking (mentions at least three of the dimensions or RELATED standards and system thinking dimension).

3. Sufficient
Mentions at least one aspect of standards and system thinking and develops at least one aspect; that is, the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, the respondent might mention standards based reform and go on to talk about performance standards or content standards. Or, the respondent might mention that alignment is important and go on to note that assessment must be aligned with content standards.)

4. Quite a Bit
Mentions at least two aspects of standards and system thinking and develops two or more; that is, the response goes beyond mentioning an aspect to develop it suggesting a deeper understanding.

5. A Great Deal
Mentions at least two aspects of standards and system thinking and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of standards and system thinking to making a link or connection between at least two aspects. For example, the respondent might mention aligning curriculum and assessment and then note that when you do this you can hold teachers accountable for student achievement.

Content experts reported the greatest concerns for the examples in this rubric (see Table 20). While all three were neutral or "mostly agreed" that the scoring rubric offered a clear definition for this subdomain (1a), two of the three were neutral or "mostly disagreed" that the rubric included clear examples of responses that qualified for this area (2c).

Table 20. Standards-based and Systems Thinking Feedback Responses			
	Expert		
Question Summary	**1**	**2**	**3**
1. a. rubric provides clear definition	3	4	4
1. b. definition needs additional dimensions	4	2	2
1. c. definition needs fewer dimensions	3	4	1
2. a. directions provide clear guidance	4	4	4
2. b. rubric provides clear explanation of what response qualifies for each level	2	4	4
2. c. rubric provides clear examples for each level	3	2	4

Experts offered few written comments except for Expert 2 who commented more in detail about the need for better examples and explained how examples and non-examples would help scorers:

> …again, examples and non-examples would be helpful. I found myself wondering about the level of inference that was appropriate, especially with coherence. Many of the responses seemed to give a nod to the idea, but there was a range of detail and depth of responses.

I discuss below how I modified the examples based on this comment.

Experts' scoring of the example responses provided mixed evidence for the reliability of their coding. Table 21 reports these scores.

Table 21. Experts' Scores for Standards-based and Systems Thinking Responses	Expert		
Scenario	1	2	3
Principal 1			
1	3	3	1
2	2	5	4
3	1	1	4
4	1	0	1
5	0	4	1
Principal 2			
1	1	0	3
2	0	3	?
3	0	0	?
4	0	0	?
5	1	0	?
Principal 3			
1	1	1	?
2	1	1	?
3	1	0	?
4	0	0	?
5	0	1	?

Coding of Principal 1's answers generated the greatest differences in the scores, with Expert 1 often assigning the lowest scores (as with scenarios 3-5) and Expert 2 assigning higher scores (see her scores for 2 and 5). The response to scenario 3 shows how experts differed greatly in their evaluations.

> I would celebrate the good questions and concerns of these professionals and begin the conversation about ongoing formative assessment. I may want ask questions of staff like: What assessments do you currently use in your classroom that would provide us with evidence of student learning? Do you use common assessments among content areas? If not, would you consider developing and administering common assessments and then discussing the results?...What would be a useful assessment? How can we hold ourselves accountable for student learning? If not, the state test, what other measure?

For this excerpt in which the principal offers detailed questions about the use of assessments and their alignment with content, two experts rated this as only a mere mention (1), while the third rated this as a 4 (quite a bit of discussion). On the other hand, scoring of the next two principals by Experts 1 and 2 was much more consistent as they both scored the principals low on this subdomain of expertise. Despite the latter scores, disagreements over Principal 1's scores emphasized the need for me to offer additional examples as the experts recommended in their comments.

Changes to the Rubric for Standards-based and Systems Thinking

My changes to this rubric focused on the examples I included to illustrate the different scoring levels. In response to Expert 2's statement I added more explicit examples for the rubric scores of "no mention (0)," "a little/a mere mention (1)," "sufficient (3)," and "a great deal (5)." These examples provided more tangible references by which to evaluate the "range of detail and depth of responses" that Expert 2 highlighted. The figure below illustrates these changes made to the rubric.

Figure 14. Modified Rubric for Standards-based and Systems Thinking

Standards-based and Systems Thinking Scenario Coding Rubric

Aspects of standards and system thinking referred to in the scale below include but are NOT limited to:

- Standards-based reform
- Standards (e.g., curriculum standards, content standards, learning standards, performance standards, etc.)

- Curriculum design, implementation, evaluation, and refinement ***(this may include specific steps or challenges to accomplishing these conditions in school, strategies to pursue these goals, or comments about the importance or role of these in successful schools)***
- What students should know and be able to do at any grade level or in any school subject
- Alignment or coherence in general,
- Alignment or coherence in reference to student assessment, curriculum standards, professional development, curricular materials, etc.
- Alignment or coherence of instruction, assessments, and materials.
- Accountability (e.g., holding staff accountable for learning, holding students accountable)
- Systemic reform as it relates to standards or curricula
- Systems theory as it relates to standards or curricula
- The political, social, cultural, and economic systems and processes that impact schools

There are possible overlaps between this code and data-based decision making. We use this general rule: if a principal discusses beginning with data and then moving to curricular decisions, we first consider this as a discussion of data-based decision-making and then look to determine if standards/alignment/systems thinking are also mentioned. If so, this may be a double code.

Discussions of state and national assessments are not in and of themselves standards unless standards are explicitly mentioned. We treat most of these discussions as data-based decision making because they refer to the understanding and use of assessments that include data.

For professional development to be included in this code there must be explicit discussions of standards, alignments, or accountability.

Be careful to look closely at any system, school-wide or community-wide references as a part of the systemic theory or larger systems that influence the school.

Scoring

0. No Mention of the subcategory in the response

For example, a principal might discuss specific skills to teach in a course without explaining how the skills are part of the larger curriculum or how they relate to what students should know at a particular time.

1. A Little Discussion of the subcategory in the response

Mere mention of one or two aspects of standards and system thinking (mentions any one of the dimensions or a RELATED dimension). NOTE, saying the same thing 10 times is still a mere mention. ***For example, a principal might comment that "setting high standards" is important but not elaborate on how to do this. He might also call for "aligning the curriculum with the standards" without explaining what this entails.***

2. Some Discussion of the subcategory in the response

Mentions at least three aspects of standards and system thinking (mentions at least three of the dimensions or RELATED standards and system thinking dimension).

3. Sufficient Discussion of the subcategory in the response

Mentions at least one aspect of standards and system thinking and develops at least one aspect; that is, the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. ***(For example, the respondent might mention standards based reform and go on to talk about performance standards or content standards with specific details about the standards. Or, the respondent might mention that alignment is important and go on to note that assessment must be aligned with content standards. He or she might also discuss how one goes about achieving such alignment.)***

4. Quite a Bit of Discussion of the subcategory in the response

Mentions at least two aspects of standards and system thinking and develops two or more; that is, the response goes beyond mentioning an aspect to develop it suggesting a deeper understanding.

5. A Great Deal of Discussion of the subcategory in the response

Mentions at least two aspects of standards and system thinking and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of standards and system thinking to making a link or connection between at least two aspects. ***For example, the respondent might discuss specifics of different standards and how to use these to align curriculum and assessment; she might then note that when you do this you can hold teachers accountable for student achievement.***

Problem-solving Expertise Rubrics

Despite the fact that Leithwood and Stager (1989) examined principal responses in administrative conditions while Brenninkmeyer and Spillane (2008) and Brenninkmeyer and Weitz White (2005) examined responses in more instructional conditions, their studies nonetheless found a number of similar differences in expert versus nonexpert school leaders' problem-solving strategies (among them, use of prior planning in addressing an issue, focusing on benefits to students versus benefits to staff, and collecting information about an issue versus making assumptions about it). Leithwood and Stager (1989) derived their categories from qualitative analyses of principal responses to scenarios and reported those categories in which they found differences between expert and nonexpert principals, while Brenninkmeyer, et al. (2005 and 2008) included categories from Leithwood and Stager's (1989) work and Bullock, James, and Jamieson's (1995) research.

I focused on the key areas where one or both previous lines of research examined differences and for which the scenarios prompted the greatest discussions that might differ between expert and nonexpert principals. In total I included four subdomains for problem-solving expertise: a) the degree to which principals collect information before addressing a situation, b) the extent to which school leaders delegate tasks, c) their discussion of facing and addressing conflict with other(s), and d) their use of planning and establishing strategies to address both large and small school conditions. Appendix H elaborates on the resulting scoring rubrics that I used for problem-solving expertise; I discuss briefly the theoretical bases for the different rubrics below.

I used three criteria to prioritize the key subdomains of problem-solving expertise for this project. First, I considered if the scenarios used in this study were more or less likely to prompt a respondent to speak about a particular subdomain, and I asked content experts to comment on these likelihoods. With a number of the subdomains in which these researchers found differences the scenarios in this project provides few if any prompts (for example, informing parents or concern with feelings). Second, I examined whether or not both lines of research (Leithwood & Stager (1989) and Brenninkmeyer, et al., 2005 & 2008)) included the subdomain as a significant dimension of problem-solving expertise. I prioritized those subdomains for which both sets of work discussed differences between experts and nonexperts. Third, I selected those categories for which principals might provide varying degrees of discussion instead of the simple presence or lack of a subdomain. With both Leithwood and Stager's (1989) and Brenninkmeyer, et al's (2004 and 2008) work the researchers focused more on the binary differences between responses (i.e. they focused more on the mere presence or absence of certain concepts in principals' responses). For example they looked for use of "relevant anecdotes" versus "poor anecdotes," or focusing on student versus staff perceptions in situations. However, the analyses in this dissertation targeted the quality of response that principals provide as evidence of their expertise; this scoring evaluated their depth of discussion more than the simple presence of one trait versus another. For example, I chose such subdomains as planning and collecting data because participants might offer responses that demonstrated varying degrees of their expertise in these areas.

First, in regard to respondents' efforts to understand the problems they faced, Leithwood and Stager found that principals relied on three different

sources of information: past experiences, assumptions they made, or new information they collected (p. 142). Experts were explicit about the assumptions they made, and they often discussed strategies to collect new information regarding the scenario either through additional data or discussions with others. Experts "stressed the value of careful information collection" (p. 148). Nonexpert principals on the other hand often jumped to conclusions about the situation and/or did not discuss the need for additional information to understand the situation.

Second, Brenninkmeyer and Spillane found significant differences between expert and nonexpert principals' discussion of "task delegation" (2008, p. 464), with experts being more likely to discuss strategies for delegating authority, the specific tasks that they would delegate to others, or what they would do to transfer appropriate authority or responsibility to others. Nonexperts often paid less attention to these details, either not specifying them or implying that they would address the conditions before them (in the scenarios) on their own.

Third, Brenninkmeyer and Spillane (2008) and Bullock, James, and Jamieson (1995) each examined differences in individuals' readiness and discussion of "facing conflict" within their staffs, either with themselves personally or between other staff members. They focused on differences in the two groups' willingness to face or address conflict that might arise in a particular situation versus avoiding it. Brenninkmeyer and Spillane reported nonsignificant differences between experts and nonexpert principals and their willingness to face or avoid conflict.

Finally, Leithwood and Stager reported that experts and nonexperts differed markedly in their responses to "planning;" in this area "experts spent more effort planning for the solution process and identified more detailed steps to be included in the process than did nonexperts" (1995, p. 148). Brenninkmeyer and Spillane found significant differences between expert and nonexpert principals in the degree to which they discussed planning in their responses (2008, p. 464). These differences included how much experts used "detailed prior planning," identified "detailed steps in solution process," or stressed the "importance of information collection plans for follow-up" (Leithwood & Stager, 1995, p. 140). While these two bodies of research focused primarily on short-term planning, more recent literature illustrates how such planning can also relate to longer-term, broader visions and goals for a school. First and foremost, this research points to the crucial role that school leaders can play in planning and organizing "shared visions and goals for their schools" (Leithwood and Jantzi, 1999). Hallinger and Heck's 1996 literature review of connections between principals and school effectiveness found that the only consistent interactive mediating variable between the two was the presence of a shared vision and goals within a school. Leithwood and Jantzi (1999) defined planning more broadly than just short-term strategies; they argued it included the means to develop larger visions and goals: "Planning includes the explicit means used for deciding on purposes and goals, determining the specific nature of the goals that are set, and beginning to understand what might be entailed in their accomplishment" (p. 683). Finally, Leithwood and Riehl (2003) further argued that "leaders influence student learning by helping to promote vision and goals…" and that successful school leadership practice "includes actions aimed

at developing goals for schooling and inspiring others with a vision of the future" (p. 3).

Problem-solving Expertise Results

Subdomain 1: Gather Information to Understand the Situation.

Figure 15 shows the initial rubric that content experts reviewed for the subdomain of "gathering information."

Figure 15. Proposed Rubric for Gathering Information

Gathering Information

Definition: includes any responses in which the principal mentions, expresses or demonstrates some knowledge of collecting new information before addressing an issue:

- different sources of information a principal would reference to find out more about a problem (such as people or data)
- strategies to find out such information, such as how to collect it or analyze it
- discussion of the importance or role of additional information to understand a situation

0. No Mention of the Dimension OR a respondent makes assumptions about a situation without providing supporting information (jumping to a conclusion about what is happening is also evidence of little or no expertise in collecting new information).

1. A Little

Mere mention of one or two aspects of subject with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

For example, a respondent might mention the need to look at standardized test scores before understanding what is happening with student achievement in reading.

2. Some.

Mentions at least three or more different aspects of gathering information but does not develop any of the aspects.

3. Sufficient

Mentions at least one aspect of gathering information and develops at least one aspect. This means the response goes beyond mention of the aspect to develop it suggesting a deeper understanding. (For example, a more developed discussion of gathering information should include multiple details in the discussion as well as an explanation of why the approach is valuable or important.)

For example, a principal could discuss the importance of asking additional personnel about the condition and then go on to detail specific individuals and why their perspectives are important.

4. Quite a Bit

Mentions at least two aspects or strategies for gathering information and develops two or more; that is, the response goes beyond mentioning the aspects or strategies to developing them with more discussion that suggests a deeper understanding of those aspects.

5. A Great Deal

Mentions at least two aspects of gathering additional information and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of gathering additional information to making a link or connection between at least two aspects. For example, a principal may discuss 1) how specific data such as standardized test scores would provide insights into what is happening with the math curriculum and 2) how conversations with specific teachers would also provide information regarding the situation. She might then describe how she would use the two sources of information together to reach a deeper understanding of the conditions.

All three experts for this subdomain provided extensive written comments, but only Experts 2 and 3 completed the Likert scale items. No experts recommended changes to the directions, and the first two experts marked that they "mostly agree" or "completely agree" that the directions provided clear guidance about using the rubric (see Table 22, 2a). Experts also marked for question 2b that they "completely agree" that the scoring guide provides a "clear explanation" and about what responses qualify for each level of expertise.

Table 22. Gathering Information Feedback Responses			
	Expert		
Question Summary	**1**	**2**	**3**
1. a. rubric provides clear definition	?	4	4
1. b. definition needs additional dimensions	?	1	1
1. c. definition needs fewer dimensions	?	1	1
2. a. directions provide clear guidance	?	4	5
2. b. rubric provides clear explanation of what response qualifies for each level	?	5	5
2. c. rubric provides clear examples for each level	?	3	4

As seen in Table 23 below, experts' scores of the example responses revealed a strong agreement between them. For example, all three agreed on three of Principal 1's answers (2-4) , one for Principal 2 (1), and two for Principal 3 (1 and 3). On three others they were within a point of each other. Finally, Experts 2 and 3 agreed entirely on a total of thirteen responses. However, Expert 1 frequently rated the responses lower than the other two, providing evidence of the need to further explain the rubrics. This excerpt from the first principal's response to scenario 5 (in which teachers and administrators struggle to have meaningful conversations about improved teaching and learning) demonstrates Expert 1's lower scores:

> I would continue to bring in outside people to do inservice at every staff meeting to talk about teaching and learning. I would try to get some book studies going. I would go to grade level meetings and ask them what I could provide that would help them. I would put articles in their mailbox. Again, teachers want to do what is best for kids.

Here the principal briefly discusses collecting feedback from the teachers, but only the Experts 2 and 3 correctly identified this as a mere mention of "gathering information."

Table 23. Experts' Scores for Gathering Information Responses	Expert		
Scenario	1	2	3
Principal 1			
1	2	4	4
2	0	0	0
3	0	0	0
4	0	0	0
5	0	1	1
Principal 2			
1	4	4	4
2	2	4	4
3	4	3	4
4	1	4	4
5	2	3	4
Principal 3			
1	4	4	4
2	2	3	3
3	4	4	4
4	1	3	3
5	2	0	0

Expert 1 provided the most detailed commentary on this subdomain by arguing that it "should also refer to values and assumptions. My understanding of this literature suggests that explicit statement of one's values and assumptions aids problem-solving especially when information is missing, unobtainable, or ambiguous." Further review of Leithwood and Stager (1989) confirmed this recommendation, and I changed the definition to include a principal's "explanation of specific assumptions she or he is making about the situation and

the potential strengths or limitations of those assumptions." In response to the lower scores that Expert 1 assigned to the answers (and disagreed with the other two experts) I provided more extensive examples for the different scoring levels. The next figure illustrates changes I made to the rubric.

Figure 16. Modified Rubric for Gathering Information

Gather Information to Understand the Situation

Definition: includes any responses in which the principal mentions, expresses or demonstrates some knowledge of collecting new information before addressing an issue. ***This also includes any responses where the principal discusses his or her assumptions about a situation.***

- ***explanation of specific assumptions she or he is making about the situation and the potential strengths or limitations of those assumptions***
- different sources of information a principal would reference to find out more about a problem (such as different people or types of data)
- strategies to find out such information, such as how to collect it or analyze it
- discussion of the importance or role of additional information to understand a situation
- ***discussion of the role that additional information can play in informing the principal's assumptions***

0. No Mention of any of the aspects OR a respondent makes assumptions about a situation without providing supporting information (jumping to a conclusion about what is happening is also evidence of little or no expertise in collecting new information). ***In making these assumptions the respondent may also fail to clarify that the statements are indeed assumptions.***

For example, a principal might discuss how a difficult teacher has "no interest in working with other teachers here" or "is not interested in being here" without qualifying the statements as assumptions based on limited observations.

1. A Little Discussion of the subcategory in the response

Mere mention of one or two aspects of subject with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

For example, a respondent might mention the need to look at standardized test scores before understanding what is happening with student achievement in reading. ***However, she or he provides no additional evidence of how to do this or why additional evidence would help to inform the situation.***

2. Some Discussion of the subcategory in the response

Mentions at least three or more different aspects of gathering information but does not develop any of the aspects.

3. Sufficient of Discussion of the subcategory in the response

Mentions at least one aspect of gathering information and develops at least one aspect. This means the response goes beyond mention of the aspect to develop it suggesting a deeper understanding. (For example, a more developed discussion of gathering information should include multiple details in the discussion as well as an explanation of why the approach is valuable or important.)

For example, a principal could discuss the importance of asking additional personnel about the condition and then go on to detail specific individuals and why their perspectives are important. ***Or a principal might qualify why she or he has limited knowledge of the situation and discuss in detail how other perspectives would help to inform him or her of the conditions.***

4. Quite a Bit of Discussion of the subcategory in the response

Mentions at least two aspects or strategies for gathering information and develops two or more; that is, the response goes beyond mentioning the aspects or strategies to developing them with more discussion that suggests a deeper understanding of those aspects.

5. A Great Deal of Discussion of the subcategory in the response

Mentions at least two aspects of gathering additional information and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of gathering additional information to making a link or connection between at least two aspects. For example, a principal may discuss 1) how specific data such as standardized test scores would provide insights into what is happening with the math curriculum and 2) how conversations with specific teachers would also provide information regarding the situation. She might then describe how she would use the two sources of information together to reach a deeper understanding of the conditions.

Subdomain 2: Addressing Conflict with Others.

Figure 17 below the proposed scoring rubric that content experts evaluated in their comments.

Figure 17. Proposed Rubric for Addressing Conflict with Others

Addressing Conflict

Definition: includes any responses in which the principal mentions, expresses or demonstrates some knowledge of ANY of the following aspects regarding addressing conflict with or between faculty members:

- the importance of facing conflict with others so as to address disagreements or misunderstandings
- strategies to address conflict with others
- the benefits that come from addressing conflict

0. No mention of the dimension OR the respondent discusses avoiding conflict if possible (this implies the individual will not address a disagreement with another person).

1. A Little

Mere mention of one or two aspects of subject with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

With a mere mention here a respondent might discuss briefly her plan to speak with another person with whom she has a disagreement.

2. Some.

Mentions at least three or more different aspects of delegating tasks but does not develop any of the aspects.

3. Sufficient

Mentions at least one aspect of delegating tasks and develops at least one aspect. This means the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, a more developed discussion of addressing conflict should include multiple details in the discussion as well as an explanation of why the approach is valuable or important.)

A developed description of addressing conflict might include a principals' elaboration on specific strategies she would use to discuss a disagreement with a teacher so that the two reach a common understanding and resolve the conflict.

4. Quite a Bit

Mentions at least two aspects of addressing conflict and develops two or more; that is, the response goes beyond mentioning the aspects to developing them with more discussion that suggests a deeper understanding of the aspects.

5. A Great Deal

Mentions at least two aspects of addressing conflict and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of addressing conflict to making a link or connection between at least two aspects. For example, a respondent may describe 1) why it is important to address a particular conflict with a staff member 2) what particular strategy he can use to resolve the disagreement and 3) how the resolution can help promote better communication and cooperation between the two (this last phrase ties together the first two).

For their survey item responses Experts 2 and 3 "mostly agreed" that the rubric provided clear definitions, directions, and examples for this subdomain. However, they also "mostly agreed" that the rubric needs additional dimensions. I examined their more extensive recommendations for these dimensions in the written comments below.

Table 24. Addressing Conflict Feedback Responses			
	Expert		
Question Summary	**1**	**2**	**3**
1. a. rubric provides clear definition	?	4	4
1. b. definition needs additional dimensions	?	4	4
1. c. definition needs fewer dimensions	?	1	1
2. a. directions provide clear guidance	?	4	5
2. b. rubric provides clear explanation of what response qualifies for each level	?	4	5
2. c. rubric provides clear examples for each level	?	4	4

Expert 2 raised the importance of a principal evaluating when to engage a situation or conflict, and argued that the existing definition assumed that all engagement conflict is helpful (which may not always be the case):

> I think you might wish to include a dimension about assessing the relative importance of a conflict—what a principal friend always referred to as "deciding whether this is a hill you are prepared to die on." Engaging in a conflict for the sake of a conflict is not helpful. Administrators need sufficient wisdom to assess whether what presents itself as a conflict really is a conflict, or whether a simple compromise, or even capitulation, is more appropriate. At the other end, a shrewd administrator should be able to identify those conflicts which are absolutely critical and which cannot be allowed to slide into compromise because of the future repercussions of not reaching a clear and unequivocal resolution.

Expert 3 agreed:

> I am not sure whether there is an embedded assumption that all situations include conflicts. It may often be desirable to find ways to avoid conflict, but that is not always possible or, for that matter, desirable. It's a naïve (and potentially poor) leader who believes this.

Though the literature does not specify this aspect of addressing a conflict, a principal's ability to evaluate a situation or conflict and determine whether or not it merits engagement certainly offers evidence of a greater expertise in this subdomain. As I explain in the last paragraph I added to the definitions to address this oversight in the original rubric.

As with their scores for "gathering of information," experts' scores of the example scenarios agreed highly for the first two principals (see Table 25 below). However, Principal 3's answers again generated differing scores, with Expert 1 often scoring the answers much lower than Experts 2 and 3. Their scores for principal 3's response to scenario 2 demonstrated these differences particularly well.

> First, for the teacher who remarked, "It must be the kids." A private conversation would take place to clarify that my values do not agree, that all students can learn, I am personally responsible for their success and I will be writing the evaluation. It is a something that needs to be clearly articulated with that particular staff member as a non-negotiable.

In this excerpt the principal offers specific details for the discussion as well as a clear rationale for having the conversation. Expert 1 scored this as a mere mention while Experts 2 and 3 more correctly scored the deeper discussion the principal provided. Such differences as this one demonstrated the need to modify the definitions and examples for more complex discussions of this subdomain.

Table 25. Experts' Scores for Addressing Conflict Responses	Expert		
Scenario	1	2	3
Principal 1			
1	1	0	0
2	0	0	0
3	0	0	0
4	1	0	0
5	1	0	1
Principal 2			
1	1	1	1
2	1	0	1
3	0	0	0
4	0	0	0
5	0	0	0
Principal 3			
1	2	0	0
2	1	3	3

3	0	4	4
4	0	0	0
5	0	3	5

Based on Expert 2 and 3's comments about when to address a conflict I added to the definition these two bullet points:

- strategies to evaluate the importance of addressing a conflict (i.e. whether or not a conflict is important enough to engage)
- strategies to determine how far to push in engaging a conflict (e.g. is it important to "win" a conflict, or is a compromise preferred?)

I also added to the examples for the scoring levels of 1 and 3 to clarify content that would qualify for these levels. The following figure illustrates the changes I made to the rubric.

Figure 18. Modified Rubric for Addressing Conflict with Others

Addressing Conflict with Others

Definition: includes any responses in which the principal mentions, expresses or demonstrates some knowledge of ANY of the following aspects regarding addressing conflict with or between faculty members:

- the importance of facing conflict with others so as to address disagreements or misunderstandings
- ***strategies to evaluate the importance of addressing a conflict (i.e. whether or not a conflict is important enough to engage)***
- ***strategies to determine how far to push in engaging a conflict (e.g. is it important to "win" a conflict, or is a compromise preferred?)***
- strategies to address conflict with others
- the benefits that come from addressing conflict
- ***what one can learn from addressing a conflict***

0. No mention of the dimension OR the respondent discusses avoiding conflict if possible (this implies the individual will not address a disagreement with another person).

1. A Little Discussion of the subcategory in the response
Mere mention of one or two aspects of subject with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

With a mere mention here a respondent might discuss briefly her plan to speak with another person with whom she has a disagreement, ***but she might offer few details about how to do this in a productive way.***

2. Some Discussion of the subcategory in the response
Mentions at least three or more different aspects of ***addressing conflict*** but does not develop any of the aspects.

3. Sufficient Discussion of the subcategory in the response
Mentions at least one aspect of ***addressing conflict*** and develops at least one aspect. This means the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, a more developed discussion of addressing conflict should include multiple details in the discussion as well as an explanation of why the approach is valuable or important.)

A developed description of addressing conflict might include a principals' elaboration on specific strategies she would use to discuss a disagreement with a teacher so that the two reach a common understanding and resolve the conflict.

4. Quite a Bit of Discussion of the subcategory in the response

Mentions at least two aspects of addressing conflict and develops two or more; that is, the response goes beyond mentioning the aspects to developing them with more discussion that suggests a deeper understanding of the aspects.

5. A Great Deal of Discussion of the subcategory in the response
Mentions at least two aspects of addressing conflict and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of addressing conflict to making a link or connection between at least two aspects. For example, a respondent may describe 1) why it is important to address a particular conflict with a staff member 2) what particular strategy he can use to resolve the disagreement and 3) how the resolution can help promote better communication and cooperation between the two (this last phrase ties together the first two).

Subdomain 3: Delegation of Tasks.

This figure below shows the initial scoring rubric that content experts evaluated in their comments.

Figure 19. Proposed Rubric for Delegation of Tasks.

Delegation of Tasks
Definition: includes any responses in which the principal mentions, expresses or demonstrates some knowledge of ANY of the following aspects regarding delegating responsibilities:

- Specific reasons for assigning particular responsibilities to other staff members (for example, it may be more efficient, or those individuals might possess more information about particular aspects of a project or issue)
- Strategies for delegating tasks to other staff members (such as reasons for whom to select or what information and responsibilities to assign to them)
- Specific mention of individuals or people to whom to assign tasks

0. No Mention of the Dimension OR the respondent discusses or implies that he will take on the project entirely by himself (this implies that he will not delegate any responsibilities to others).

1. A Little
Mere mention of one or two aspects of subject with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

> In a "mere mention" a principal might discuss briefly the need to ask a reading specialist to follow up with a teacher about specific students' low reading scores.

2. Some.
Mentions at least three or more different aspects of delegating tasks but does not develop any of the aspects.

3. Sufficient
Mentions at least one aspect of delegating tasks and develops at least one aspect. This means the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, a more developed discussion of task delegation should include multiple details in the discussion as well as an explanation of why the approach is valuable or important.)

> A more developed discussion could include a principal's discussion of the need to ask specific math teachers to collect test score and homework data about their low-scoring students before they as a team consider what new math program to use.

4. Quite a Bit
Mentions at least two aspects of delegating tasks and develops two or more; that is, the response goes beyond mentioning the aspects to developing them with more discussion that suggests a deeper understanding of the aspects.

5. A Great Deal
Mentions at least two aspects of gathering tasks and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of task delegation to making a link or connection between at least two aspects. For example, a respondent may describe 1) why it is important to include teachers in the evaluation of a math program 2) what particular roles they can play in the

evaluation process and 3) how their participation helps to build support for the plan that results from the evaluation process (this last phrase ties together the first two).

Experts 2 and 3 agreed strongly in their responses to the survey items, and both called for few if any changes to the rubrics in their answers. However written comments and a closer review of their scores of example answers demonstrated the need for changes to the rubrics.

Table 26. Delegation of Tasks Feedback Responses			
	Expert		
Question Summary	**1**	**2**	**3**
1. a. rubric provides clear definition	?	4	4
1. b. definition needs additional dimensions	?	3	2
1. c. definition needs fewer dimensions	?	1	1
2. a. directions provide clear guidance	?	4	5
2. b. rubric provides clear explanation of what response qualifies for each level	?	4	5
2. c. rubric provides clear examples for each level	?	4	5

Expert 3 wrote that the definition as originally written implies that delegation is always better:

> I am a little concerned that you assume that delegation will be necessary, and the more delegation the better. The scenarios do not all suggest that multiple delegations would be necessary or appropriate. Furthermore, it is possible that some respondents, when they talk about collecting information, do not intend to do so themselves, but did not elaborate on this in their responses.

This comment highlighted an oversight in the definition: in some cases a principal may decide (correctly) that he or she needs to address a situation without others' input or assistance. The original rubric made no allowance for such a situation; it only rewarded greater delegation of a task or tasks as evidence of a principal having more expertise in a certain area. In the closing paragraph I discuss how I responded to Expert 3's comment.

Table 27 shows the scores that experts assigned to principal responses.

Table 27. Experts' Scores for Delegation of Tasks Responses	Expert		
Scenario	1	2	3
Principal 1			
1	1	0	0
2	0	0	0
3	0	0	0
4	1	1	1
5	1	1	1
Principal 2			
1	1	0	0
2	0	0	0
3	0	0	0
4	0	1	1
5	1	0	1
Principal 3			
1	1	1	1
2	0	2	3
3	0	2	3
4	0	3	4
5	1	0	0

Experts' scores of the example responses showed striking consistency across the different scenarios for Principals 1 and 2. Give the low scores for these responses, however, some of this agreement may be due to the limited discussion of delegation that principals provided in their answers (so there may have been higher agreement simply because there was nothing to score). Scores for Principal 3, which have a wider range in the scores, illustrated how there was greater disagreement as experts used higher scores. Expert 1 consistently marked the answers lower for Principal 3. A review of the response to the second scenario illustrated how more complex discussions of assigning responsibility generated different scores. Principal 3's answer focused on starting a new group in the school to respond to lower reading scores.

> …It would also be time to initiate a literacy commission in school comprised of teachers from all department to look at the data regarding literacy, establish schoolwide goals for literacy and develop a staff develop plan to address reading across the curriculum.

As shown in Table 27, the experts generated three different scores for these comments. While Expert 1 did not believe this was an example of any delegation, Expert 3 scored this as a more developed discussion of delegation. Thus while experts for the first two principals showed high agreement, their scores for the third principal demonstrated the need to increase agreement for more complex discussions of delegating tasks.

First, based on Expert 3's comments I modified the definition to include the possibility that a principal still shows expertise by explaining specifically the value of *not* delegating authority. I also included in the definition a principal's discussion of which specific tasks or plans to delegate in addressing a situation. In light of the of disagreements between experts' scores which I discussed in the previous paragraph, I added examples to the scoring levels in an effort to demonstrate more clearly conditions that would qualify for the different scoring levels.

Figure 20. Modified Rubric for Delegation of Tasks

Delegation of Tasks

Definition: includes any responses in which the principal mentions, expresses or demonstrates some knowledge of ANY of the following aspects regarding delegating responsibilities:

- Specific reasons for assigning (or not assigning) particular responsibilities to other staff members (for example, it may be more efficient, or those individuals might possess more information about particular aspects of a project or issue). ***(Note: here a principal with expertise may also explain why he or she made the decision not to delegate responsibility.)***
- Strategies for delegating tasks to other staff members (such as reasons for whom to select or what information and responsibilities to assign to them)
- ***Specific tasks to delegate to others***
- Specific mention of individuals or people to whom to assign tasks
- ***Plans to transfer authority for something***

0. No Mention of the Dimension OR the respondent discusses or implies that he will take on a complex task or project entirely by himself or herself (this implies that he/she will not delegate any responsibilities to others).

1. A Little Discussion of the subcategory in the response

Mere mention of one or two aspects of subject with no development of the aspect(s). NOTE: mentioning the same thing 10 times with no development is still a mere mention.

In a "mere mention" a principal might discuss briefly the need to ask a reading specialist to follow up with a certain teacher. However, he provides few details few details about what the specialist should do or discuss with the teacher. Also, the principal might not discuss the value or benefit of delegating this responsibility to a specialist.

2. Some Discussion of the subcategory in the response

Mentions at least three or more different aspects of delegating tasks but does not develop any of the aspects.

3. Sufficient Discussion of the subcategory in the response

Mentions at least one aspect of delegating tasks and develops at least one aspect. This means the response goes beyond mention of an aspect to develop it suggesting a deeper understanding. (For example, a more developed discussion of task delegation should include multiple details in the discussion as well as an explanation of why the approach is valuable or important.)

A more developed discussion could include a principal's discussion of the need to ask specific math teachers to collect test score and homework data about their low-scoring students before they as a team consider what new math program to use. ***He or she might also discuss specific tasks for them to undertake or explain the advantage of having the team address this need.***

Alternatively, a principal might explain specifically why he or she will take on an issue with such details as why she can do the best job or why others do not have the capacity to address the issue.

4. Quite a Bit of Discussion of the subcategory in the response

Mentions at least two aspects of delegating tasks and develops two or more; that is, the response goes beyond mentioning the aspects to developing them with more discussion that suggests a deeper understanding of the aspects.

5. A Great Deal of Discussion of the subcategory in the response

Mentions at least two aspects of gathering tasks and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of task delegation to making a link or connection between at least two aspects. For example, a respondent may describe 1) why it is important to include teachers in the evaluation of a math program 2) what particular roles they can play in the evaluation process and 3) how their participation helps to build support for the plan that results from the evaluation process (this last phrase ties together the first two).

Subdomain 4: Planning and Goal Setting.

Figure 21 below shows the initial scoring rubric that content experts evaluated in their comments.

Figure 21. Proposed Rubric for Planning and Goal Setting.

Planning and Goal Setting

Definition: includes any responses in which the principal mentions, expresses or demonstrates some knowledge of ANY of the following:

- the concepts school mission, vision, or strategy and the contents of or differences between each of these
- the process of creating a vision to improve student achievement
- developing a strategy to implement the vision
- building action plans that align with and execute the school design or the school redesign process
- building action plans that create and align all the elements that contribute to student learning within the school
- the principal's role in creating and leading a learning culture and organization within the school. To qualify for this code the discussion must focus on actions that explicitly foster learning between teachers and/or staff members. For example, a principal's simple mention of promoting "collaboration" would not qualify unless she/he discussed how such a strategy would improve the school learning culture or the exchange of pedagogical ideas between staff members.
- general reference to planning, vision, mission, or strategy. This may include references to teacher planning only if the principal mentions such planning as a component of an overall vision or plan for the school. A simple mention of promoting teacher planning with little other context would not fit under this.

The code refers more broadly to organizing all the school resources and activities around establishing a school's vision or strategies; it can include but does not focus exclusively on curriculum or teaching. In some cases a principal may overlap with or discuss another code (such as standards-based reform or systems thinking) while elaborating on the planning process. In these cases code those sections as part of the planning discussion (see examples below).

This code also focuses on knowledge of the planning process and its different components. In summary the code refers to the activity or concepts involved in developing a plan or vision. It does not include principals' descriptions of actively evaluating or assessing progress in achieving a vision, strategy, or action plan. For such actions to be coded as "planning and goal setting" the individual must discuss these actions in the context of larger plans or strategies for the school.

0. No Mention of Dimension

1. A Little
Mere mention of one or two aspects of planning and goal setting (mentions any one of the dimensions or a RELATED dimension). NOTE, saying the same thing 10 times is still a mere mention.

2. Some.
Mentions at least three aspects of planning and goal setting (mentions at least three of the dimensions or a RELATED planning and goal setting dimension).

3. Sufficient
Mentions at least one aspect of planning and goal setting and develops at least one aspect; that is, the response goes beyond mention of planning or goal setting to develop it suggesting a deeper understanding. (For example, the respondent might mention planning and goal setting and go on to discuss specific steps to develop the vision for the school. Or, the respondent might discuss setting the school's vision and then list specific strategies he or she would use to implement the vision).

4. Quite a Bit
Mentions at least two aspects of planning and goal setting and develops two or more; that is, the response goes beyond mentioning an aspect to develop it suggesting a deeper understanding.

5. A Great Deal
Mentions at least two aspects of planning and goal setting and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of planning and goal setting to making a link or connection between at least two aspects. For example, the respondent might discuss 1) the process for establishing specific strategies that align with the school vision and 2) how to build action plans that execute the strategy, and then she might explain how these steps all connect to the overall goal of improving student achievement in the school.

Experts' responses to the survey items for this final subdomain showed that two experts "mostly agreed" that the definitions, guidance, and examples in the rubric were clear. They further commented that they "strongly" or "mostly disagreed" that the rubrics needed more or fewer dimensions. These responses provide positive initial support for the rubrics in their original forms, but the written comments and expert scores raised critical issues about necessary changes to them.

Table 28. Planning and Goals Setting Feedback Responses			
	Expert		
Question Summary	**1**	**2**	**3**
1. a. rubric provides clear definition	?	5	4
1. b. definition needs additional dimensions	?	1	2
1. c. definition needs fewer dimensions	?	1	1
2. a. directions provide clear guidance	?	4	5
2. b. rubric provides clear explanation of what response qualifies for each level	?	4	5
2. c. rubric provides clear examples for each level	?	4	5

While he did not respond to the survey items, Expert 1 provided a strong challenge to the definition in this subdomain. He commented that the scoring rubric extended beyond the research cited for this subdomain of expertise:

> What you put into the rubric seemed inconsistent with the research reported in the first part. Neither Leithwood's research cited in the first section nor other research on planning and goal-setting in problem-solving focuses explicitly on vision and mission, but rather on setting goals for solution and planning strategy (as you indicate).

While the originally cited literature fits this description, more recent work has raised the importance of a leader's ability to plan not just to address short-term issues but to examine and organize the larger vision and mission of the school (need citations here). Leithwood and Stager (1989) and others do not include such long-range planning in their definitions, but this has emerged as a crucial component for a leader's expertise in planning and goal setting. I discuss my response to Expert 1 at the end of this section.

Table 29 reports the experts' scores for the principals below.

Table 29. Experts' Scores for Planning and Goal Setting Responses	Expert		
Scenario	1	2	3
Principal 1			
1	0	1	0
2	1	0	1
3	2	0	0
4	1	0	0
5	0	2	1
Principal 2			
1	0	3	4
2	0	0	0
3	1	0	1
4	1	2	3
5	0	1	1
Principal 3			
1	0	2	3
2	0	3	4
3	1	4	4
4	1	3	3
5	0	3	4

In their scoring of the example answers all three experts consistently scored Principal 1 low in this subdomain of expertise. Experts 2 and 3 agreed substantially in their rating of the last two principals, with all of their scores being within one point of each other or identical. Expert 1 consistently scored the last two principals lower in this subdomain, and a review of the responses illustrated how the experts disagreed when a principal laid out particular steps in a program. For example, Principal 3 wrote the following response to address a new math program that has not seen improvements in test scores:

> It sounds as if I have a department that is more concerned about a "program" than instructional practice. After targeting and identifying "with data" the exact problems we are facing. I would hold a facilitated day with the department to focus the conversation about instructional practice and where the program is aligning or not to meet the needs of students. It may be time to go in a different direction, but we must do that intelligently.

As shown in the table below for scenario 1 scores, Expert 1 indicated there was "no mention" of planning, while Experts 2 and 3 scored this answer that listed different steps as including "some" discussion (Expert 2) of the planning and a "sufficient" discussion of planning (Expert 3). Disagreements such as this illustrated the need to explain better the scoring levels for this rubric so raters understand that more detailed discussions of steps qualify for demonstrations of expertise.

Expert 1's comment forced me to examine additional research regarding a leaders' expertise in planning and goal setting. As I have already stated, original work such as Leithwood and Stager (1989) and Brenninkmeyer and Spillane (2008) do not explicitly include broader, longer term thinking on such issues as school mission, vision, strategies. These texts referred to shorter term, more immediate approaches such as "demonstrates some order and structure to solution" (Brenninkmeyer & Spillane, 2008, p. 463) or "uses detailed prior planning, and identifies detailed steps in solution process" (Leithwood & Stager, 1989). After more careful review of the literature I have not removed these bullets from the definitions. While neither of the two pieces above explicitly cites longer term planning, they also do not exclude it. Furthermore, recent literature certainly points to the important role that school leaders can play in planning and organizing their schools around clear, "shared visions and goals" (Leithwood and Jantzi, 1999). In their 1996 review of literature examining connections between the principal's role and school effectiveness, Hallinger and Heck found that the presence of these goals in schools was the only mediating variable that was consistently interactive with principal leadership. Leithwood

and Jantzi (1999) defined planning more broadly than just short-term strategies; it included the means to develop larger visions and goals: "Planning includes the explicit means used for deciding on purposes and goals, determining the specific nature of the goals that are set, and beginning to understand what might be entailed in their accomplishment" (p. 683). Leithwood and Riehl (2003) further argued that "leaders influence student learning by helping to promote vision and goals…" and that successful school leadership practice "includes actions aimed at developing goals for schooling and inspiring others with a vision of the future" (p. 3). In light of this evidence I drew the definitions for this rubric more broadly to include both short-term and long-term planning and goal setting. I also modified the research summary for this subdomain to include additional literature that speaks to this longer-term dimension of problem-solving expertise. The figure below shows the changes I made to this last rubric.

Figure 22. Modified Rubric for Planning and Goal Setting

Planning and Goal Setting

Definition: includes any responses in which the principal mentions, expresses or demonstrates some knowledge of ANY of the following:

- ***the use of detailed prior planning to address a situation or challenge***
- ***how to identify specific steps required to address a situation***
- ***the importance of following a plan to successfully address a situation or solve a problem***
- the concepts school mission, vision, or strategy and the contents of or differences between each of these
- the process of creating a vision to improve student achievement
- developing a strategy to implement the vision
- building action plans that align with and execute the school design or the school redesign process
- building action plans that create and align all the elements that contribute to student learning within the school
- the principal's role in creating and leading a learning culture and organization within the school. To qualify for this code the discussion must focus on actions that explicitly foster learning between teachers and/or staff members. For example, a principal's simple mention of promoting "collaboration" would not qualify unless she/he discussed how such a strategy would improve the school learning culture or the exchange of pedagogical ideas between staff members.
- general reference to planning, vision, mission, or strategy. This may include references to teacher planning only if the principal mentions such planning as a component of an overall vision or plan for the school. A simple mention of promoting teacher planning with little other context would not fit under this.

The code refers both to short term plans or strategies to address a situation as well as the broader plans for organizing all the school resources and activities around a school's vision or strategies; it can include but does not focus exclusively on curriculum or teaching. In some cases a principal may overlap with or discuss another code (such as standards-based reform or systems thinking) while elaborating on the planning process. In these cases code those sections as part of the planning discussion (see examples below).

This code focuses on knowledge of the planning process and its different components for both short and long term issues. In summary the code refers to the activity or concepts involved in developing a plan or vision. It does not

include principals' descriptions of actively evaluating or assessing progress in achieving a vision, strategy, or action plan. For such actions to be coded as "planning and goal setting" the individual must discuss these actions in the context of larger plans or strategies for the school.

0. No Mention of Dimension at all in the response
For example, a principal may discuss a response to a situation without laying out a sequence of steps to address it.

1. A Little Discussion of the subcategory in the response
Mere mention of one or two aspects of planning and goal setting (mentions any one of the aspects or a RELATED aspect). NOTE, saying the same thing 10 times is still a mere mention. For example, a principal might mention the importance of setting and agreeing on a clear school vision but provide no specific details about how to do that. ***Alternatively, he or she might discuss the importance of developing a plan to address a situation but not provide any details about it.***

2. Some Discussion of the subcategory in the response
Mentions at least three aspects (or RELATED aspects) of planning and goal setting.

3. Sufficient Discussion of the subcategory in the response
Mentions at least one aspect of planning and goal setting and develops at least one aspect; that is, the response goes beyond mention of planning or goal setting to develop it suggesting a deeper understanding. (For example, the respondent might mention planning and goal setting and go on to discuss specific steps to develop the vision for the school. Or, the respondent might discuss setting the school's vision and then list specific strategies he or she would use to implement the vision).

4. Quite a Bit of Discussion of the subcategory in the response
Mentions at least two aspects of planning and goal setting and develops two or more; that is, the response goes beyond mentioning an aspect to develop it suggesting a deeper understanding.

5. A Great Deal of Discussion of the subcategory in the response
Mentions at least two aspects of planning and goal setting and develops two or more AND makes connections between at least two of the aspects mentioned; that is, the response goes beyond mentioning and developing two or more aspects of planning and goal setting to making a link or connection between at least two aspects. For example, the respondent might discuss 1) the process for establishing specific strategies that align with the school vision and 2) how to build action plans that execute the strategy, and then she might explain how these steps all connect to the overall goal of improving student achievement in the school.

Summary.

The purpose of this first content validation study was to solicit experts' evaluations of how well the rubrics captured the content of leadership expertise subdomains developed in the literature. Experts' responses in this first study presented a complex picture of the needed changes to the rubrics. Both experts' rich discussions of the rubrics and the results of their scoring of principal responses provided insights into the necessary revisions to better define and illustrate the subdomains. While the expert responses and their scores sometimes disagreed (for example experts' survey answers for "gathering data"

called for few changes to the rubrics, but their scoring illustrated the need for improved explanations in them), the data most frequently demonstrated that changes to each of the rubrics were essential. This evidence derived from experts' written comments and responses to the survey questions along with their varying scores for the principal expert responses.

Experts' survey responses focused most frequently on the rubrics' definitions and examples; they rarely called for changes in the directions. Across the rubrics experts frequently "mostly agreed" that the rubrics provided clear definitions of the subdomains, but they also specified a number of additional dimensions for each. For example, while all three experts "mostly agreed" that the rubric provided a clear definition of "effective teaching and learning," all three still "mostly agreed" that additional dimensions were needed. Their written statements in this subdomain recommended such changes as including knowledge of "cognitively or developmentally appropriate assessment strategies." These recommendations helped me to add and modify integral dimensions in the rubrics to insure that they covered the full range of content discussed in each of the subdomains.

Experts' scores of principals' responses offered evidence of how experts often did not arrive at common understandings of the subdomains or scoring levels after reading the rubrics. Across the subdomains low agreement between their scores most frequently demonstrated that experts still needed additional clarification about what different responses qualified for each subdomain. While reviewers agreed significantly in the subdomains for "problem-solving expertise" (as with "delegation of tasks" and "addressing conflict"), their scores more frequently did not match in other areas. As I have detailed above, I

responded to these differences in scores by providing additional examples to clarify what texts would qualify for the different scoring levels on the rubric.

Written comments provided the most specific guidance for modifying the rubrics to clarify them further; experts made pointed recommendations for changes. As with the survey responses, they focused primarily on the definitions and the examples, and I made significant changes to both across the subdomains. These responses helped to improve the content coverage of these rubrics and insure that the rubrics capture the subdomains that they define.

Above all, the content experts' responses and comments across all the domains illustrated the complex nature of evaluating the quality of response that principals demonstrate in written scenarios. These content experts' differing scores demonstrated the difficulty of attaining high agreement without an arbitration process or further discussions between raters to reconcile differences in understanding. In response to these results I scored and arbitrated example principal responses with another graduate student before scoring the full set of principal responses to the scenarios. This allowed me to achieve satisfactory reliability in my scoring before I worked with the actual data from the principals. I discuss this strategy in more detail in the next chapter.

CHAPTER VI

STUDY 2 RESULTS: EXAMINATION OF THE LEADERSHIP EXPERTISE MEASURES' CONSTRUCT VALIDITY

This chapter presents the findings from the second study that analyzed the principals' scores for the revised rubrics to evaluate the construct validity of the measures of leadership expertise. As stated previously in the methods section, construct validity involves examining how a measure relates to other measures. A researcher hypothesizes about its relationship to other measures, and confirmation of these hypotheses offers evidence that the measure as operationalized adequately captures the construct.

I have divided this chapter into three sections that present findings for the three main questions of whether or not these measures behave according to theory and how the measures relate to each other. These main questions are the following:

1. Do the measures capture different levels of principal expertise?
2. Do the measures within each of the three larger domains relate to each other as predicted by theory?
3. Do the overall scores for the three main domains relate to each other to provide evidence of a larger single construct of expertise?

In each section I summarize the central question briefly before presenting findings.

Substudy 2. A. Capturing Different Levels of Expertise With the Rubrics

For the first part of this second study I asked, do the measures capture differing levels of principal expertise? Two primary sets of findings illustrate how the scenarios measured different levels of expertise: a) qualitative examples

of specific answers that demonstrate the content of "high expertise" and "low expertise" scores and b) descriptive summaries of the scores for each of the subdomains that show that different scenarios generated variation across scores.

1. Qualitative Examples of Variation in Scoring

In this section I have provided contrasting cases of responses for each of the subdomains. Each set of examples that I have included here demonstrates differences in the quality of response and how rubrics scored those responses differently. For example I compared text from a principal with high levels of expertise in pedagogical content knowledge to that of a lower expertise response and explained how and why their scores reflected different levels of expertise. These examples illustrate that the scoring rubrics did in fact distinguish between higher and lower levels of expertise based on the responses that principals provided.

Leadership Content Knowledge

Subject Matter

With this subdomain principals demonstrated their expertise in the specific content that students learn in their classes. Thus the rubric for this subdomain assigned higher scores only to those responses in which principals demonstrated an understanding of the unique characteristics of a particular area of content such as math or reading. With this first subdomain no principal received a higher score than a 1 for a "mere mention"—none of the participants discussed subject matter content specific to one area. The two examples below

show how principals often discussed subject matter in general or provided only a brief discussion of subject matter.

In responding to scenario 3 (in which there are slumping reading scores and teachers differ in their responses the lower scores) one principal wrote the following:

> These kids would be writing during every subject area. They would also be involved with higher order thinking skills like that are provided by the Jr. Great Book program. (Quality of Response Code of 1)

Here the principal briefly recommended that students engage in writing across subject areas and mentions one specific program, but he or she provided no specifics about the writing strategies or the "Jr. Great Book" program that demonstrate a deeper understanding of subject matter for reading.

A second principal offered the following comments in response to scenario 3:

> In addition, all teachers are teachers of reading regardless of the subject area. Reading skills encompasses all content areas. (Quality of Response Code of 1)

Again this principal made general comments about teaching reading across the curriculum, but she or he provided no additional details about what specific strategies teachers would use for teaching reading.

For this subdomain the principal responses did not show great differences in their expertise in subject matter. The greatest differences consisted of those who did not mention this subdomain and those who offered only a mere mention of it.

Pedagogical Content Knowledge

This subdomain focused on leaders' understanding of the subject-specific teaching strategies that teachers employ in presenting content to their students. A principal with higher expertise in this subdomain offered the following response to the video segment showing a teacher presenting a reading lesson to her students. This response consisted of advice that she or he offered to the teacher to improve the reading lesson:

> I would recommend that the teacher demonstrate revising the sentence by writing in what students suggested then divide the students into small groups to revise one. The groups could then write their sentences on chart paper, post them and read them to the class. In this way many more students participate in the revision and each student sees many more examples. The next step would be authoring sentences that incorporate expressive language. (Quality of Response Code of 3)

In this excerpt the principal provided detailed recommendations about how to improve the reading lesson through more structured group activities. After providing these additional details for the activity the principal explained specific benefits for the changes: more students would participate, and they would see more examples of sentence revisions. This person also discussed a logical next lesson to build on what they just learned. In this case the principal provided not only detailed advice but also showed a deeper understanding of why these changes in the pedagogy would be beneficial to students' understanding of how to revise sentences. Finally, this response demonstrated that the principal understood how to connect this with a larger sequence of lessons for the students to build on their learning. This response received a "3" that reflected the principal's more developed understanding of the pedagogical strategy changes she or he recommended.

This example contrasted with the following response from another principal giving advice for the same scenario:

> She started the lesson by reading a segment of the book and asking the students what they noticed. I don't think the students understood "what they noticed" as this is vague. She needed to give an introduction as to what the lesson would involve and what the students were looking for prior to reading the passage. The lesson moved too slowly and was not engaging. Having the students come to the chart took too much time and allowed others to become bored. (Quality of Response Code of 1)

In this case the respondent simply summarized what she saw in the video and provided only a brief recommendation for the teacher with no specific details or reasons for the changes to the lesson ("she needed to give an introduction as to what the lesson would involve and what the students were looking for prior to reading the passage"). Even though this response is similar in length to the first example, the principal provided no deeper discussion of the pedagogical value for his or her advised changes. There is no evidence of the principal's deeper expertise in the area of pedagogical content knowledge for reading in this response, and this response scored a "1" as a result.

Here we see that the rubrics scored different qualities of response from principals. A comparison of these two cases illustrates how the scoring rubrics captured a higher quality of response when the first principal not only provided more detailed advice but also discussed why these changes would improve the students' opportunity to learn from the reading lesson. It also revealed how the lengths of responses were not solely responsible for higher scores in these rubrics—just because a principal wrote more about something did not guarantee he or she would receive a higher score.

Teachers as Learners

This subdomain focused on principals' strategies to help teachers improve in their instruction. Again, the scoring rubric assigned higher scores to responses in which principals discussed strategies specific to content areas. In the first response to scenario 3 below (which asks the principal to address slumping reading scores) a principal recommended a specific program for inclusion in the professional development for teachers.

> Appropriate professional learning in the teaching of reading, such as the Guided Reading Lesson, would be arranged for all teachers. A presentation for all teachers that demonstrates the need for all teachers to teach reading in their content area would be needed. I would provide teachers with the action research that supports this. The Design Team with input from all teachers would investigate to determine if addition resources, professional learning, and interdisciplinary planning might be needed. (Quality of Response Code of 3)

This response contrasted with a second principal who wrote only the following for scenario 3:

> Professional learning for teachers will be implemented based on the National Reading Panel research. (Quality of Response Code of 1)

While this individual cited specific research in this response, she or he offered no further discussion of what this research entails or what approaches would best help teachers learn new instructional strategies or improve their teaching. This comparison demonstrated how principals had to discuss specific ways to support teachers' professional development to score higher on this rubric.

Learning-centered Leadership

Data-based Decision Making

This subdomain focused on a principal's understanding of how to use data to inform his or her decisions for school improvement. To score higher

according to this subdomain a principal had to provide a deep discussion of the data he or she would collect or the strategies he or she would use to analyze the information regarding in-school conditions. I first present a higher expertise example for this subdomain and contrast it with a low expertise response to show how the rubrics captured different levels of expertise.

One principal provided the following response to scenario 2, which asked how the principal would respond to a situation in which student math scores have dropped in recent years but teachers differ in their support for continuing to use the math program.

> I would analyze the data with members of the staff:
> --Individual students progress
> --Classroom teacher results
> --Subgroup results
> --Examine teacher lesson plans and compare with results
> --Interview teachers and students for feedback
> --I would then share this information with those in a position to determine changes that might be possible to be made
> ALSO:
> --Provide professional development for full implementation with the program as written and recommended
> --Request that those teachers who were having success mentor those having difficulty with the program
> --Utilize grade level and cross grade level collaboration and planning to assure that the program was consistently being implemented and best practices with regard to it were being used by all
> --Check program's alignment with "tests"
> --Check for incremental improvement over time
> --Discourage "whatever works" philosophy!
> (Quality of Response Code of 3)

In the first part of this response the principal discussed how to disaggregate student test data according to teacher and other subgroups, and he or she detailed the need to collect additional data such as teacher lessons and student and teacher perspectives on the curriculum. Toward the end of the answer the principal discussed the need to evaluate curricular alignment with tests and to

evaluate student progress over a period of time. This individual demonstrated higher expertise in data-based decision making by not only listing strategies to disaggregate the data but also providing additional data sources to use in evaluating the impact of the math curriculum.

Another principal provided the following response for the second scenario.

> The data should be studied to determine the problems. A committee would look at the results and a possible solution. Efforts need to center on how carefully the teachers used the new program and then study the effect on the students.
>
> (Quality of Response Code of 1)

This response provided only a superficial demonstration of how to collect or analyze data. Here the principal mentioned studying data and assembling a team to do so, but he or she provided no specific details about how to do this or what data to use in evaluating the program or its effects on students. A comparison of the two responses above demonstrates that individuals exhibited higher or lower expertise in data-based decision making when they provided differing levels of detail about data or the processes they would use to analyze this information. The scoring rubrics captured these different qualities of response by assigning different quality of response scores to principals' answers.

Effective Teaching and Learning

This subdomain included principals' understanding of key theories about succsssful instruction and student learning. Unlike with leadership content knowledge, principals did not have to include subject-specific content to score higher in this subdomain. The examples below demonstrate how principals

scored higher (or lower) based on how extensively they explained their theories about teaching and learning.

The first principal offered a highly detailed strategy to address the school-wide reading problems discussed in scenario 3.

> You cannot blame the kids for what they don't know. The school must adopt a philosophy that all teachers are responsible for the success of the students. If you are using the same techniques that you have always done and are expecting different results, you are misguided and will be disappointed.
>
> The principal must develop with the teachers a comprehensive reading program. The program must include: Take home backpacks of books for students in grades k and 1 (5 per classroom) to be shared by the grade level. This will encourage students to read more at home especially if the home is not a literature rich environment (usually not when the free and reduced lunch rate is high). There needs to be diverse classroom libraries that students may use and read at all times. Students have to be given time to read to each other and in small groups. Students should be placed in heterogeneous reading groups so they can listen to each other and share and discuss the book with each other. Parent volunteers or co-teachers can help with the reading groups and the teacher needs to work with each group weekly to listen to them and provide commentary. The teacher must read a book to the class (usually a book above their grade level). The teacher will lead discussions and ask students to visualize, predict and share their feelings about these stories. The books should reflect the diversity of the classroom and present to students literature from different ethnic backgrounds and genres. The school should have a library with an ever growing circulation of books that encompasses all levels of reading expertise and include non-fiction, poetry and biographical selections. Finally, the school should develop and present Family Reading Nights to help parents, help their kids succeed in reading. (Quality of Response Score of 5)

Here the principal discussed two different strategies to help students in reading: a book program to encourage them to read, and the use of heterogeneous reading groups. In both cases the principal provided extensive reasons for the benefits of these approaches, and he or she explained a larger purpose in the first paragraph (establishing an idea of collective responsibility amongst teachers).

This response contrasted with a principal who provided the following response to scenario 3.

> It is imperative that as educators we must address the needs of all students through differentiated instruction. It is incumbent upon us all to identity strategies, best practices, and interventions to improve learning outcomes for all students. It addition, all teachers are teachers or reading regardless of the subject area. Reading skills encompasses all content areas. (Quality of Response Score of 1)

While this individual supported the use of differentiated instruction he or she offered little additional explanation of what this entails or why this would be useful for raising reading scores. These two examples demonstrate how the rubrics captured different levels of expertise for "teaching and learning."

Monitoring Instructional Improvement

Principals who scored higher in this subdomain demonstrated greater understanding of how to observe and evaluate teachers' instructional strategies and curriculum. This subdomain examined principals' discussions of benchmarking instruction or curriculum to identify improvement or their alignment with the school improvement plan.

For this first example one principal (in addressing scenario 3) not only discussed the strategies for monitoring instructional improvement (informal observations and review of lesson plans) but listed specific questions that would guide the observations.

> An inventory of teaching strategies being practiced in the classrooms needs to be examined. Are the teaching practices "best practices" or researched practices? Informal observations and review of lessons plans would support this inventory. The question also needs to be answered, "Do the teaching practices of the past align with the standards of today?" If the teachers are not teaching the correct information then improvement will not be seen. (Quality of Response Code of 3)

This higher quality response contrasted with the next principal's answer that only briefly explained the importance of examining content delivery and offered no strategies to do this.

> I believe that the entire school is responsible for the achievement of our students and their success. We as a school need to revisit our delivery of the curriculum and begin to assess early and often to monitor student progress. Through this monitoring we will be able to identify areas that we need to address through better instruction and resources. (Quality of Response Code of 1)

These two examples show how the rubrics captured different qualities of response, even when the answers were similar in length. Those answers that better explained the purposes or strategies for "monitoring instructional improvement" received higher scores.

Standards-based Reform and Systems Thinking

This subdomain focuses on principals' understanding of how standards help to guide school-wide reforms or align curriculum, learning benchmarks, or evaluation strategies. Principals with greater expertise in this area discussed in more detail the importance of aligning instruction, assessments, and other materials according to standards or broader school goals.

In this first example response to scenario 2 (in which math scores have dropped in the school) one principal offered a more developed discussion of this subdomain when he/she discussed a number of different strategies to pursue and emphasized that they be anchored in the learning standards:

> I believe that teachers should use every resource they need to teach! One math program is not the answer to instruction or improving student achievement. I would encourage teachers to use whatever resources they need to teach the math standards. I would work hard to provide them with funding for the resources and continue to support all teachers. As long as the standards are being taught to students and high expectations

> are in place, I would support teacher's methodology to deliver the curriculum to the students. (Quality of Response Code of 3)

This response contrasted with the next answer in which a principal discussed briefly the importance of alignment but did not offer any strategies for how to achieve this alignment.

> The math program should be based on the standards the students need to master, not based on a specific program. There needs to be alignment with the standards, the instruction, and the assessments. Therefore, to address this situation, I would make sure the above is indeed put in place. (Quality of Response Code of 1)

These examples demonstrate how the rubrics captured principals' greater understandings of the importance of standards and broader organizational alignment in school reforms.

Problem-solving Expertise

Gathering Information

This subdomain encompassed principals' discussions of collecting additional information before addressing a situation or deciding on a course of action. The scoring rubric focused on principals' discussion of strategies to collect such information or what types of information they would gather.

In this first example a principal offered two more developed discussions of gathering information in responding to scenario 2 (in which the school faces declining math scores with a recently adopted math program). First, he or she described analyzing the alignment between the math curriculum and the Georgia state standards. Second, this individual detailed the additional questions to address about the program.

> My first step would be to inquire as to whether the math program is aligned with the Georgia Performance Standards and the assessment

> instrument that is used. Are teachers well versed in the standards and the assessment instrument?
>
> If there is alignment, I'd proceed to inquire as to whether or not the program was implemented as intended (integrity).
>
> Again, if the answer if "yes", further inquiry would be needed as to why students are not mastering the content/skills.
>
> --What are the areas/sub-skills on the assessment that are causing low test scores? Are these addressed sufficiently in the instructional materials?
>
> --Is there a sub-group of students who are not performing well on the assessment (achievement gap)?
>
> --Are pre-assessments used to identify level of mastery and/or areas of concern prior to beginning instruction?
>
> --Are there other issues involved? (I.e. veteran teachers who have expertise in teaching math; focus on covering the text versus content mastery;)
>
> --What supplemental materials are used to reinforce learning?
>
> --Were there teachers whose students did well on the test? Were their students representative of the school population? If the answers are "Yes", then these teachers may need to share best practices and/or provided time for other teachers to observe their instruction using the adopted program.
>
> All of these things would need to be considered prior to throwing out the math program. (Quality of Response Code of 5)

This response contrasted with the following response to scenario 2, in which another principal only briefly detailed how she or he would collect information before responding to the situation.

> Look at the student test data. See what needs improving. Get the teachers together and see how that skill is being presented. Note the differences. Decide on the BEST way to approach the issue. Set plan in place. Review and evaluate with formative data. Re-tool the plan if necessary. (Quality of Response Code of 1)

With this second example the principal summarized looking at the test data with teachers but did not elaborate on how to do this or what specific data to use. The rubric scored this superficial discussion with only a "1."

Planning and Goal Setting

A key component of this subdomain targeted a principal's understanding of how to establish a detailed strategy to respond to an issue. Many of the analyses for this subdomain focused on how well the principal laid out a series of steps to address a scenario as evidence that he or she presented a plan. The following examples show differences in the quality of response for two principals and explain how their respective scores captured differences in planning and goal setting expertise.

First, one principal provided the answer below to scenario 2, which (as summarized previously) asked how the principal would respond to a situation in which student math scores have decreased recently, but teachers differ in their support for continuing to use the math curriculum.

> First of all I would ask for a data team representative of both sides of the issues do an in-depth analysis of the scores and look at all sub groups within the building. I would also invite a district resource person, preferably the district's math director, to serve on the team. As data is being reviewed and discussed the team would also need to look at teacher expectations and the time frame as to when teachers starting to use whatever works. A survey would need to be developed and then completed by all teachers using the math program. The team would also need to review the independent research on this math program to assess the possibility of cultural bias. The results/findings of the team's work, data analysis, and survey input would be presented to the entire faculty so each grade level could then discuss the findings on their grade level and reach a consensus regarding their recommendation. Then, each grade level team leader would bring their decision to the School Improvement Team for a discussion and vote on what to do in regards to continuing this particular math program.
>
> (Quality of Response Code 3)

The planning rubric included an individual's expertise in "the use of detailed prior planning to address a situation or challenge." In this response the principal presented a detailed, more developed summary of the strategy he or she would pursue to address the situation. This discussion established a clear chronological

order for what he or she would do to gather information about the program and organize the faculty to make an informed decision:

1. assign relevant members to a team to begin analyzing the math achievement data, collecting teacher feedback about the program, and reviewing relevant research and evaluations of the math program.
2. this team would present its findings to the faculty to promote grade level discussions about how to respond to the decreasing test scores.
3. team leaders would present their decisions to the larger School Improvement Team for a decision about the program

The structure embedded in this response demonstrated this principal's higher level of expertise in planning a response to the decreasing math scores, and corresponding score of "3" captured this more developed discussion. This response contrasted strongly with that of the next principal, who provided a much more simplistic discussion of the steps needed to address the situation.

> Look at the student test data. See what needs improving. Get the teachers together and see how that skill is being presented. Note the differences. Decide on the BEST way to approach the issue. Set plan in place. Review and evaluate with formative data. Re-tool the plan if necessary.
>
> (Quality of Response Code 1)

Here the individual provided only brief, superficial details about the actions she or he would take, and there is little rationale for the order or chronology of the response. While the first principal often discussed how information from each step (such as teacher surveys or test score data) would be used in a following stage, the second principal offered little explanation of this type. A comparison of these two responses illustrates how the scoring rubric captured different levels of expertise in planning.

Delegation of Tasks

This subdomain focused on a principal's understanding of how to assign responsibilities to others. The rubric included both respondents' reasons and strategies for delegating (or not delegating if the conditions warranted it) tasks to others in their schools.

The two examples below show how the rubrics captured varying levels of expertise in the responses. In responding to scenario 4 (in which teachers questioned the value of standardized test scores to improve student learning) the first principal specified who would participate in a data analysis team and what exactly this team would do to share information with the rest of the school.

> There should be a Data Team at the school made up of a representative from each grade level. Their job would be to develop a comprehensive thumbprint of how the students in the school are doing in all subjects. This would include test data but would also add, report information, non-fiction writing, teacher commentary on how the students are doing completing performance standards, and alternative assessments ideas. The Data Team would share the information about progress and develop instructional changes to be made in the classrooms. These suggestions would be shared with each grade level and the response from the students would be shared at the next meeting. Further, the entire staff would need to meet several times during the year to share ideas, successes and concerns noted from the testing, daily instruction and formative assessments. The test results are only a snap shot of everything going on. That information needs to be used but the other information needs to be included at the school level. (Quality of Response Code 3)

The second principal's response offered much less explanation of how or why to delegate responsibility to others in the school.

> Test data should broken down (disaggregated) so that it is more teacher friendly. Graphic organizers can be used to illustrate the data in different formats. Ultimately, pre assessments will need to be conducted by each teacher at the beginning of the school year to get a more accurate view of the students performance capabilities. Data teams will be developed by grade to assist with ongoing assessment of student data based on semester benchmarks. (Quality of Response Code 1)

Here the respondent simply described what individual teachers will do and references "data teams," but there is no detail regarding who will participate on these teams or what they will do to assess data. These contrasting cases demonstrate that the rubrics assigned varying scores based on the expertise that principals showed in their responses.

Addressing Conflict with Others

Principals scored higher for this subdomain when they discussed in more detail their strategies for addressing conflict or the value of doing so. The two examples below show that the rubrics captured varying levels of discussion for how to address conflict.

In this response to scenario 5 (in which a teacher opposes the principal's entering the classroom to observe) one principal not only discussed a strategy for addressing the teacher but also explained the reasons and value for providing positive feedback to the faculty. In his/her view, building connections with teachers was integral to their effectiveness with students.

> As an instruction leader the culture of the school would have to change to one of support and praise. I would build the trust by giving positive feedback to teachers throughout the year to recognize their abilities and hard work. I believe teachers would see feedback not as a "gotcha" mentality but one of support and how I could help them to get resources or obtain training in areas that they would like to explore. Building a school that sees the administrative as supportive and fair is the key to increase teacher effectiveness and in the end student achievement. (Quality Response Code of 3)

A second principal offered this response to scenario 5.

> I would reiterate to teachers that they are no longer working in "silos." With the advent of the Professional Learning Communities, I would guide teachers to see the benefit of working as a team and the positive benefits to reviewing student work together. It is a mindset change and if I was unable to change that mindset through persuasion (initially), I would

> strongly encourage the teachers that refuse to get on board, to seek employment elsewhere. (Quality Response Code of 1)

While this individual planned to use Professional Learning Communities as a way to encourage collaboration, he or she offered no details of how to do this, nor did the principal discuss much of the value of such a strategy. Unlike the first participant, this person offers little explanation of why or how to address the resistance the teacher has shown.

Summary

The examples above have shown that principals varied in their responses to the scenarios and how the scoring rubrics rewarded "more developed" discussions of each subdomain. Principals who provided more detailed, specific discussions of the different subdomains scored higher according to the rubrics. The cases above illustrate how those individuals who demonstrated deeper understandings of a subdomain received higher ratings of expertise. Examples such as in the subdomains "pedagogical content knowledge" and "monitoring instructional improvement" showed that these scores depended more closely on the quality of respondents wrote rather than the length of their responses (in these two subdomains similar length answers received quite different scores). In the next section I examined whether or not different scenarios prompted principals to demonstrate levels of expertise.

2. Capturing Different Levels of Expertise Across the Scenarios

Here I have provided mean scores and standard deviations for each subdomain that show that principals' responses to the different scenarios

generated different scores. If the qualitative examples above used contrasting individual cases to demonstrate variation in levels of expertise, the descriptive standard deviations and means in this section summarize the range of responses from sample principals to provide further evidence that the measures generated scores that differed according to principals' expertise.

In this section I also examine possible measurement bias caused by the scenario prompts for different areas of expertise. I ask, did each scenario appear to prompt principals adequately to provide evidence of their expertise in each subdomain? Previous work from Goldring, et al. (2008) found that different scenarios prompted principals to varying degrees to demonstrate their expertise. I reviewed three sources of information to identify more closely those scenarios that best prompted for each subdomain of expertise:

a. the descriptive statistics for each subdomain across the scenarios,
b. content experts' ratings of how well each scenario prompted a principal to demonstrate his or her expertise (experts marked a 4-item likert scale that included "none at all," "a little bit," "somewhat," and "a great deal"), and
c. the context of the scenario texts and how well they prompted respondents to demonstrate their expertise in different areas.

In light of the differences between the scenarios I calculated "selected average" aggregate scores for each principal to give him or her a single score for each subdomain. These scores used only those scenarios that generated the highest average responses with the greatest variations and where the scenario texts prompted principals more directly to discuss the domain. I used the following criteria from the information above to generate selected average scores:

a. those scenarios which generated means and standard deviations of at least .5 or higher, or provided the highest variation for a subdomain,
b. those scenarios that experts rated as prompting a higher demonstration of expertise from principals, and
c. those scenarios whose text more directly prompted princpals to demonstrate expertise in a particular subdomain.

In a final check on the selected averages, I scored the principal experts' responses from Study 1 to examine if their average responses on the selected scenarios were also higher than on the other scenarios. With this last examination for each domain I asked, did expert principals also demonstrate greater expertise in responding to the selected averages? I have included these results as evidence of how well the different scenarios elicited demonstrations of expertise.

I have divided this section according to the three main areas of expertise. Tables 29, 30, and 31 present the averages and standard deviations for each domain, and I discuss evidence that the scenarios generated differential responses from principals. For each table I include in bold text those scores that I used to generate the selected averages. In each section I provide examples of my decisions to include different scenarios for one of the "selected average" scores to illustrate how I generated these final selected averages for the subdomains. I conclude each section with a final table that shows whether or not expert principals' average scores were also high on the selected scenarios.

Leadership Content Knowledge

Table 30 presents descriptive statistics for the Leadership Content Knowledge subdomain scores, and I use these data to discuss the selected averages for each subdomain.

	Table 30. Subdomains for Leadership Content Knowledge Expertise		
Scenario	Subject Matter	Pedagogical Content Knowledge	Teachers as Learners
1	***Mean: .23*** ***SD: .43***	**Mean: .91** **SD: .97**	Mean: 0 SD: 0
2	*Mean: .02* *SD: .15*	*Mean: .12* *SD: .32*	***Mean: .35*** ***SD: .57***
3	***Mean: .14*** ***SD: .351***	***Mean: .4*** ***SD: .73***	**Mean: .67** **SD: .81**
4	*Mean: 0* *SD: 0*	*Mean: .12* *SD: .50*	Mean: 0 SD: 0
5	*Mean: 0* *SD: 0*	Mean: 0 SD: 0	Mean: 0 SD: 0
6	*Mean: 0* *SD: 0*	Mean: 0 SD: 0	Mean: 0 SD: 0
Selected Averages	Mean: .19 SD: .29	Mean: .65 SD: .65	Mean: .51 SD: .52

N=43

These results summarize principals' scores across the scenarios. For "subject matter" only two scenarios (1, the video of a teacher presenting a lesson, and 3, in which student scores in math were falling) generated many responses with relevant content. 10 principals (23%) provided a discussion of this subdomain in scenario 1, while 6 (14%) discussed it in subdomain 3. Few if any principals discussed this subdomain in the other scenarios (only 1 mentioned it in scenario 2). With "pedagogical content knowledge" I also used scores from scenarios 1 and 3. In the first scenario 63% (N=26) of principals discussed this

subdomain; their responses ranged from scores of 1 (superficial discussion) to 3 (more developed discussion). In the third scenario 32.5% (N=14) of principals discussed the subdomain, with their scores ranging from 1 to 4. On average, principals discussed dimensions of both "subject matter" and "pedagogical content knowledge" more frequently and with greater variation in these two scenarios than in any of the others.

As evidenced by their low means and standard deviations, Scenarios 2, 4, 5 and 6 did not elicit many if any responses regarding these two subdomains. Closer scrutiny of the first and third scenario texts helped to explain these differences: these both prompted principals more explicitly to discuss their understanding of the subject matter and teaching skills necessary for a particular subject area. For example, scenario 3 asked principals to address a situation in which reading scores are dropping in the school and both reading and non-reading teachers need assistance in helping to raise scores. In answering this scenario principals demonstrated more explicitly and more frequently their knowledge of reading content and teaching strategies as they discussed different strategies to improve instruction for reading. Finally, a review of content experts' evaluation of the scenarios supported these decisions: the two responding content experts marked that scenarios 1 and 3 would "somewhat" prompt for pedagogical content knowledge (these were the highest scores they offered for any scenario in this subdomain). Based on these data I used only scores from scenarios 1 and 3 to generate principals' average scores for "subject matter" and "pedagogical content knowledge."

For "teachers as learners," respondents discussed this subdomain most in the second and third scenarios (which focused on reduced math scores and

teachers' mixed reactions to lower reading scores, respectively). No principals discussed "teachers as learners" in any of the other scenarios. Content experts' ratings for scenarios 2 and 3 mirrored these descriptive results: they marked that these two scenarios "somewhat" prompted for this subdomain. These were the highest scores that experts assigned to any of the scenarios for this area.

Table 31 summarizes my ratings for the principal experts' responses. In this table all the selected scenarios' values are bolded. This allows one to compare which scenarios were selected in the process above with the means and standard deviations for the expert principals' scores. Therefore, if a bolded value in this table is also high, this means that expert principals on average scored higher on a scenario that was selected through the analysis above. As I mentioned in the introduction for this substudy, I present these findings as additional evidence for whether or not the scenarios prompted for demonstrations of expertise.

In this table we see that expert principals' scores matched the sample principals' responses across all three subdomains. On all of the bolded selected scenario scores expert principals also demonstrated higher levels of expertise—these means and standard deviations were equal to or higher than the other scores in each subdomain. The non-bolded values (for those scenarios not selected) were less than the results for the selected average scores. For "leadership content knowledge," this review of expert principals' responses provided further evidence that the selected scenarios prompted for greater demonstrations of expertise.

	Table 31. Expert Subdomain Scores for Leadership Content Knowledge Expertise		
Scenario*	Subject Matter	Pedagogical Content Knowledge	Teachers as Learners
2	*Mean: .33* *SD: .58*	*Mean: 1* *SD: 1*	***Mean: 2*** ***SD: 2.65***
3	***Mean: 1.67*** ***SD: 1.53***	***Mean: 1*** ***SD: 1***	**Mean: 3** **SD: 0**
4	*Mean: 0* *SD: 0*	*Mean: .12* *SD: .50*	Mean: 0 SD: 0
5	*Mean: 0* *SD: 0*	Mean: 0 SD: 0	Mean: 0 SD: 0
6	*Mean: 0* *SD: 0*	Mean: 0 SD: 0	Mean: 1 SD: 0
Selected Averages	Mean: 1.67	Mean: 1	Mean: 2.5

* *Because scenario 1 was a video, expert principals could not respond to the first scenario; scores from that scenario are not included here.*

One final point for these subdomains merits discussion. On average, principals' scores in all of these subdomains across the selected averages were very low. Even when using just the scores from the selected scenarios, principals wrote limited comments about these areas of expertise, and they most frequently provided only mere mentions or superficial discussions of them.

Learning-centered Leadership

Table 32 summarizes the descriptive data for the Learning-centered Leadership subdomains.

	Table 32. Subdomains for Learning-centered Leadership Expertise			
Scenario	Effective TL	Data-based DM	Monitoring Instructional Improvement	Standards-based Reform
1	***Mean: 1.58*** ***SD: .85***	Mean: .02 SD: 1.52	Mean: 0 SD: 0	Mean: .07 SD: .26
2	***Mean: .72*** ***SD: .67***	***Mean: 1.47*** ***SD: 1.3***	***Mean: .19*** ***SD: .39***	***Mean: 1.44*** ***SD: 1.05***
3	***Mean: .95*** ***SD: 1.02***	***Mean: .67*** ***SD: .837***	**Mean: .14** **SD: .52**	***Mean: 1.02*** ***SD: 1.06***
4	***Mean: .33*** ***SD: .47***	***Mean: 1.93*** ***SD: 1.22***	Mean: .05 SD: .21	***Mean: .19*** ***SD: .55***
5	Mean: .37 SD: .49	Mean: .12 SD: .32	*Mean: .84** *SD: 1.00*	Mean: .33 SD: .72
6	Mean: .37 SD: .49	Mean: .19 SD: .55	Mean: .02 SD: .15	Mean: .12 SD: .32
Selected Average	Mean: .90 SD: .46	Mean: 1.36 SD: .78	Mean: .16 SD: .32	Mean: .88 SD: .59

**Despite this scenario's relatively high mean and standard deviation I did not include it because of its negative correlation with the other two selected average scenario scores.*

The descriptive data in Table 31 present principal scores across scenarios. Scores for the subdomain "effective teaching and learning" were highest on average and showed the greatest variation for scenarios 1 through 4. Here criteria beyond the descriptive statistics guided creation of the selected averages. Despite the scores for scenarios 5 and 6, they did not focus directly on this subdomain (scenario 5 asked a principal to respond to a teacher resisting classroom observations, and scenario asked how a principal would promote better discussions in faculty meetings). Furthermore, content experts gave these

two scenarios the lowest scores for their prompting of this subdomain: they marked that these scenarios would prompt for "teaching and learning" only "a little bit." In light of these additional factors I used scenarios 1 through 4 for this selected average.

For the subdomain "data-based decision making," I compared scores from scenarios 2 and 5 to demonstrate how much variation lies between the principals' responses for the different scenarios. For the second scenario 26% of respondents (N=43) offered no mention of data-based decision making, 40% provided only 1-2 superficial mentions of the concept, 2% offered 3 or more superficial mentions, 30% provided one in-depth discussion of standards-based thinking, and 2% offered more than one developed discussion of standards-based thinking. This contrasted with scenario 5, in which 88% of respondents included no mention of data-based decision making, and 12% provided only one or two mere mentions of this subdomain. Principals on average discussed aspects of data-based decision making more frequently and with greater variation in the second scenario (which asked them to address a situation in which slumping math scores have left teachers adhering closely to a curriculum or discarding it for "whatever works") than in scenario 5 (this prompted them to respond to a group of teachers increasingly opposed to having administrators monitor instruction regularly in their classrooms).

As evidenced by their low means and standard deviations, Scenarios 1, 5 and 6 did not elicit many if any responses regarding using data-based decision making. In this case scenarios 2-4 generated both higher mean scores along with larger standard deviations in the scores. Examination of the scenario texts further helped to explain these differences: all three prompted principals more

explicitly to discuss their understanding or use of data in addressing conditions within their schools. For example, scenario 4 asked principals to respond to a situation in which a group of teachers questioned the value of standardized math and reading test scores for their understanding of students and their teaching strategies. In answering this scenario principals demonstrated more explicitly and more frequently their expertise in using data. As summarized above, Scenario 5 on the other hand asked principals to discuss their views of monitoring the instruction of a resistant teacher, and it clearly prompted fewer discussions of data-based decision making (see Appendix A for exact scenario texts). Finally, the three content experts' ratings of the scenarios matched these results for data-based decision making: after scoring the example responses they marked that these three scenarios would prompt for data-based decision making "somewhat" or "a great deal." Based on these data and observations, I used only scores from scenarios 2-4 to generate principals' average scores for data-based decision-making.

For "monitoring instructional improvement" content experts' comments were key to choosing the scenarios for the selected average. While scenarios 2, 3, and 5 had the highest means and standard deviations, scenario 5 (which addressed a teacher's resistance to observation) correlated negatively with the scores from 2 and 3 (which dealt with decreasing math and reading scores, respectively). Content experts rated the second and third scenarios most likely to prompt for discussions of this subdomain of expertise, and I therefore used only scenarios 2 and 3 for this selected average. In these two scenarios only a limited number of respondents referred to this subdomain (8 in scenario 2 and 4 in

scenario 3). Their scores ranged from 1 at the lowest (a mere mention) to a 3 (a more developed discussion).

Finally, scenarios 2, 3, and 4 generated the highest means and standard deviations for "standards-based reform and systems thinking." While scenario 5 also generated relatively high descriptive statistics (a mean of .33 and a standard deviation of .72), content experts commented that this scenario prompted for the subdomain only "a little bit." The content in this scenario (in which a teacher resisted administrator observation) had little relation to the subdomain. However, content experts predicted that scenarios 2, 3, and 4 were more likely to prompt for this subdomain. The descriptive results matched these predictions, with 58%, 65%, and 14% of respondents, respectively, discussing "standards-based reform and systems thinking" in their answers. Their scores ranged from a low of 1 (a mere mention) to a high of 3 (a more developed discussion).

A review of expert principals' responses to the scenarios (see Table 33 below) shows that their scores matched those of the sample principals: across almost all the subdomains their average selected scenario responses were equal to or higher than means and standard deviations for the scenarios not used. The two notable exceptions came in scenario 6 for both "data-based decision making" and "standards-based reform," in which the means and standard deviations were equal to or higher than the selected scenario results. In both of these cases, however, the higher scores were due to single expert principals who offered more developed discussions of the subdomains—these results did not offer strong evidence that scenario 6 would elicit higher scores across numerous principals in these two subdomains. Overall these results supported the scenarios selected above: expert principals offered greater demonstrations of

expertise on the same scenarios selected from analyses of the sample principals' responses.

	Table 33. Expert Subdomain Scores for Learning-centered Leadership Expertise			
Scenario*	Effective TL	Data-based DM	Monitoring Instructional Improvement	Standards-based Reform
2	***Mean: 3*** ***SD: 0***	***Mean: 2.33*** ***SD: 2.31***	***Mean: .33*** ***SD: .58***	***Mean: 2*** ***SD: 2.65***
3	***Mean: 1.67*** ***SD: 1.53***	***Mean: 2.33*** ***SD: 1.15***	**Mean: 1.33** **SD: 1.53**	***Mean: 3*** ***SD: 0***
4	***Mean: 0*** ***SD: 0***	***Mean: 3*** ***SD: 0***	Mean: 0 SD: 0	***Mean: 1*** ***SD: 1.73***
5	Mean: 0 SD: 0	Mean: 2 SD: 1	*Mean: .33* *SD: .58*	Mean: 0 SD: 0
6	Mean: 0 SD: 0	Mean: 2.33 SD: 1.15	Mean: .33 SD: .58	Mean: 1.67 SD: 2.89
Selected Average	Mean: 1.56	Mean: 2.56	Mean: .83	Mean: 2

* *Because scenario 1 was a video, expert principals could not respond to the first scenario; scores from that scenario are not included here.*

To summarize for this domain, as with the previous domain "leadership content knowledge," on average, sample principals' scores of expertise across the selected averages (as shown in Table 31) was quite low.

Problem-solving Expertise

Table 34 presents descriptive results from the Problem-solving Expertise subdomains.

	Table 34. Subdomains for Problem-Solving Expertise			
Scenario	Gather Information	Planning and Goal Setting	Delegate Authority	Address Conflict
1	Mean: 0 SD: 0	Mean: 0 SD: 0	Mean: 0 SD: 0	Mean: 0 SD: 0
2	***Mean: 1.26*** ***SD: 1.16***	***Mean: 1.23*** ***SD: 1.361***	***Mean: .58*** ***SD: .794***	*Mean: .07* *SD: .457*
3	***Mean: .49*** ***SD: .736***	***Mean: .63*** ***SD: 1.047***	**Mean: .47** **SD: .667**	Mean: 0 SD: 0
4	***Mean: .51*** ***SD: .827***	***Mean: .6*** ***SD: 1.116***	**Mean: .7** **SD: .964**	Mean: 0 SD: 0
5	Mean: .16 SD: .531	Mean: .26 SD: .581	*Mean: .21* *SD: .559*	***Mean: .84**** ***SD: 1.067***
6	Mean: .14 SD: .351	**Mean: .58** **SD: .906**	**Mean: .81** **SD: .906**	Mean: 0 SD: 0
Selected Average	Mean: .75 SD: .68	Mean: .76 SD: .67	Mean: .64 SD: .49	Mean: .84 SD: 1.07

**Only one scenario response was used for this subdomain because of low means and standard deviations for the other scenarios.*

As with the previous two tables, the results above show the differences in principal scores across the scenarios. I first discuss scores from the subdomain "gather information" for scenarios 2 and 6 to demonstrate how scores varied between the principals. For the second scenario, 26% of respondents (N=43) made no mention of gathering information to address the situation, 47% (20 principals) provided only 1-2 superficial mentions of this subdomain, 9% (4 principals) offered 3 or more superficial mentions, 16% (7 principals) provided one in-depth discussion of gathering information to address the scenario, and 2% (1 principal) offered two more developed discussions of gathering information. Responses to scenario 6 differed greatly: 86% of respondents (37 principals)

included no mention of gathering information, and just 14% (or 6 principals) provided only one or two mere mentions of this subdomain. As demonstrated in these results, respondents on average discussed aspects of gathering information more frequently and with greater variation in scenario 2 (its prompt is summarized above) than in scenario 6 (which asked principals how they would improve staff discussions about better teacher instruction and student achievement).

The descriptive results for scenarios 1, 5, and 6 above demonstrate how these scenarios did not elicit many if any responses regarding "gathering information." Scenarios 2-4 generated higher average scores as well as larger standard deviations for this subdomain. Closer reviews of these scenario texts helped explain the differences: all of them prompted respondents more frequently to discuss their expertise in gathering information before addressing the situation. When principals answered scenario 3 to discuss how they would address dropping reading scores in the school, they more frequently and explicitly described their strategies for gathering specific types of information before developing a response or solution. Scenario 5 on the other hand asked principals to discuss their strategies for monitoring a resistant teacher's instruction, and most of the principals (except for 4) offered no discussion of gathering additional information before responding to the teacher. Third and finally, content experts' ratings of the scenarios marked that scenarios 2, 3, and 4 would prompt for gathering information "somewhat or a great detail." Using these data, I selected only those scores from scenarios 2-4 to generate principals' average scores gathering information.

For the subdomain "planning and goal-setting" scenarios 2-4 and 6 provided the greatest variations in scores. The means for these scenarios ranged from .58 (scenario 6) to 1.23 (scenario 2), and their standard deviations ranged from .906 (scenario 6) to 1.361 (scenario 2). These scenarios addressed topics such as dropping test scores (scenarios 2 and 3) and promoting faculty discussions around curriculum and instruction (scenario 6). Scenarios 1 and 5 generated few if any comments about this subdomain. Content experts reported that scenarios 2-4 and 6 would prompt for this suddomain "a great deal" or "somewhat"—the two highest marks that could be given to the scenarios.

The descriptive statistics for the subdomain "delegate authority" were similar to those of "planning and goal setting": scenarios 2-4 and 6 generated the highest variations in scores. Means ranged from .47 (scenario 3) to .81 (scenario 6). No respondents discussed this subdomain in scenario 1, and just 7 principals discussed "delegating authoring" in scenario 5. Again, content experts predicted that scenarios 2-4 and 6 would elicit demonstrations of this expertise "a great deal" and at least "somewhat." In light of this evidence, I used these for the selected average for "delegate authority."

Finally, only scenario 5 was used to generate the average score for "resolving conflict." All the other scenarios generated no discussions of this subdomain (1, 3, 4, and 6) or very few (only 1 principal discussed this area of expertise in scenario 2). A review of the scenarios' content helped explain these scores: only in scenario 5 did a principal have to explain how he or she would respond to a conflict in which teachers opposed their entering the classroom to observe instruction. For the fifth scenario 49% of principals offered no discussion

of this subdomain, while 35% (N=15) provided a "mere mention" of it, and 16% (N=7) offered more developed discussions of it.

Table 35 shows expert principals' scores for this domain; I compare these results to those of the sample principals in the previous table.

	Table 35. Expert Subdomain Scores for Problem-Solving Expertise			
Scenario	Gather Information	Planning and Goal Setting	Delegate Authority	Address Conflict
2*	***Mean: 1.67*** ***SD: 1.15***	***Mean: 1.67*** ***SD: 1.15***	***Mean: 1*** ***SD: 0***	*Mean: .33* *SD: .58*
3	***Mean: 1.67*** ***SD: 1.53***	***Mean: .33*** ***SD: .58***	**Mean: 2.33** **SD: 1.15**	Mean: 1 SD: 1.73
4	***Mean: 1.33*** ***SD: 1.53***	***Mean: 1*** ***SD: 1.73***	**Mean: 1.33** **SD: 1.53**	Mean: 0 SD: 0
5	Mean: 1.33 SD: 1.53	Mean: 0 SD: 0	*Mean: 0* *SD: 0*	***Mean: 1.67*** ***SD: 2.89***
6	Mean: 1 SD: 1.73	**Mean: 0** **SD: 0**	**Mean: 1.67** **SD: 1.53**	Mean: .33 SD: .58
Selected Average	Mean: 1.67	Mean: 1	Mean: 1.58	Mean: 1.67

* *Because scenario 1 was a video, expert principals could not respond to the first scenario; scores from that scenario are not included here.*

Expert principals' means were highest in the selected scenarios for all of the subdomains. These results offered further evidence that the selected scenarios in each subdomain elicited the greatest demonstrations of expertise. There was one exception: scenario 5 for "gather information." For "gather information" one principal offered a more developed discussion (for a score of "3") to the response that raised the average for scenario 5. This offered only limited evidence that scenario five would generate greater demonstrations of "gathering information" expertise if administered to more principals. As with the two previous two domains, this comparison supported the scenarios selected in Table 34.

Summary

To summarize the findings for this substudy, both the specific text examples and the descriptive results illustrated that these measures captured different levels of expertise in principals' comments. These results demonstrated that principals' answers differed across the scenarios, and they showed that the rubrics assigned numeric values according to their different quality responses. With the creation of "selected average" scores these measures incorporated only those scenarios that best prompted for the subdomains, and these final scores captured more varied evidence of principals' expertise. For each domain a review of expert principals' responses to the scenarios offered strong additional evidence that the selected scenarios elicited the greatest demonstrations of expertise in the different areas.

However, a final discussion must address the low scores for the selected averages across the subdomains. As summarized above, principals' selected average scores for the subdomains ranged from 0 to 3, but the mean scores for the selected averages for all the subdomains (a range of .16 o 1.36, as shown at the bottom of the previous three tables) offered evidence that on average even the selected scenarios generated minimal discussions of the subdomains for their responses. While these low scores raise questions of the principals' expertise (principals overall may have had low expertise in these areas), a more important question focuses on the scenarios themselves: did the measures adequately elicit demonstrations of expertise for these subdomains?

Despite most content experts' positive ratings for the selected average scenarios (which helped to guide selection of the scenarios for the averages), some experts did express concern about the scenarios' ability to elicit

demonstrations of expertise. For example, one content expert for learning-centered leadership wrote the following about how well the scenarios prompted for a full range of "effective teaching and learning" expertise: "I'm not sure your prompts lead participants to the level of detail" described in the rubric. A second expert raised the following concerns after reflecting on the definitions and sample responses for "leadership content knowledge":

> I thought the definition was OK, but I didn't see specificity about the nature of subject matter, differences between subjects or connections in the responses in 14 of the 15 responses. This leads me to wonder if the scenarios spurred respondents to think in these terms, or if they were given suggestions to attend to these three subcatergories…What was difficult was not perceiving enough content in the scenarios that would lead people to be content-specific in their responses. If the scenarios had a greater degree of content, perhaps the responses would have elicited more content.

If one of the important contributions of these measures is to capture more graduated levels of expertise for school leaders (as opposed to Leithwood, et al.'s binary coding for the mere presence of problem-solving expertise), then the results from the selected averages along with the comments above question just how well the scenarios accomplished that. While there is variation in these scores, it is quite limited, and it is a factor to which I return in the final discussion of this dissertation. As structured, they may not have prompted participants explicitly enough for them to demonstrate their full expertise in these different areas. If this is the case, then the low scores may represent shortcomings with the measures more than principals' actual expertise in a particular subdomain.

Substudy 2. B. Relationships Within the Main Domains of Expertise

In the literature I have cited for educational leadership expertise, researchers have argued that particular subdomains comprise the three larger

domains. This substudy examined those theoretical arguments that the subdomains would help to capture the three broader constructs. Using the selected average scores from the previous substudy I examined the relationships among the subdomains within each of the three larger areas of expertise. Here I asked, how do the subdomain measures in each domain relate to one another? Allen & Yen (1979) and Crocker and Algina (1986) have discussed a number of different strategies to assess construct validity, one of which involves examining correlations between the measures of interest. Significant correlations between the measures of interest and theoretically relevant measures would provide initial evidence that the larger constructs behave according to theoretical predictions.

Because of the theoretical arguments that each of the subdomains falls under one of three primary domains of expertise, one might assume that the subdomain scores within each domain will correlate because they comprise a larger construct of expertise. However, this may not be the case; the subdomains may not overlap much conceptually. Researchers have argued which constructs comprise the larger domains, but they acknowledge that these subdomains are theoretically distinct constructs. Given the conceptual differences between the subdomains I actually predicted that there would *not* be strong correlations between the subdomains in each main area of expertise. For example, in the area of "leadership content knowledge" a principal may have a higher expertise in subject matter but not know much about how to guide and encourage a teacher in pursuing professional development in that subject matter. Nonetheless, higher scores in these two areas combined would help identify a principal who is higher overall in "leadership content knowledge" expertise—each of the scores

plays a complementary role in measuring this expertise. Low correlations between the subdomains in each area would therefore offer evidence that these are conceptually distinct but still might help capture a broader domain of expertise. Clark and Watson (1995) write that moderate correlations between items of .15 to .50 offer preliminary evidence that they are internally consistent and may yet comprise a larger construct. I also reported a Cronbach's alpha coefficient of reliability to provide an additional measure of the subdomains' internal consistency in measuring the broader domains. While I predicted finding low correlations, I predicted that there would be higher alpha coefficients—such results would offer evidence that while the subdomains were conceptually distinct they nonetheless comprised internally consistent scale measures of the larger domains.

In this substudy I again divided the findings according to the three primary domains of expertise. I first analyzed correlations between the selected averages for the subdomains within each larger domain of expertise, and I then discussed the alpha reliability values for each to examine further the overall internal consistency of these scores in measuring the larger domains.[4] While the correlations helped to explain the internal relationships between the subdomain scores, the alphas provided additional insight into how well the scores might create scales to measure the larger domains of expertise. Given the theoretical arguments for these areas, I have predicted that these analyses would produce lower correlations but higher alpha reliability coefficients.

[4] In a later section I also examined if there were greater correlations between subdomains across *all* three primary domains.

For each of the domains I have closed with a discussion of the strategy for generating overall scores for each of the three primary domains of leadership expertise. For example, low correlations between the subdomains would suggest that they are conceptually distinct, but a higher alpha reliability score would provide evidence that the subdomains together provide an internally consistent scale to tap a larger construct such as learning-centered leadership. Such results would support the use of an average of the subdomain scores to generate a single overall value for each domain of expertise. I have discussed the statistical results for each area alongside the larger theoretical arguments to justify my development of aggregate scores for each domain. For example, even if the alpha reliabilities were low for some of the subdomains, I nonetheless generated aggregate scores for the domain to explore their larger relationships to each other. These aggregate scores for each main domain allowed me explore in the final section of this substudy the larger theoretical question of how the broader domains relate to each other.

Leadership Content Knowledge

Table 36. Correlations for Leadership Content Knowledge

	Subject Matter	Pedagogical Content Knowledge	Teachers as Learners
Subject Matter	1		
Pedagogical Content Knowledge	.45**	1	
Teachers as Learners	.03	.30*	1

N=43

*. Correlation is significant at the 0.1 level (2-tailed).

**. Correlation is significant at the 0.05 level (2-tailed).

Cronbach's Alpha: .49

As shown in Table 36, the correlations between the different subdomains for leadership content knowledge ranged from .03 to .45, with only the correlation between "subject matter" and "teachers as learners" not being significant. A review of the scenarios selected for each of these subdomains provided mixed evidence for a possible "scenario bias" in which scenarios similarly prompted respondents for two different subdomains. For example while "subject matter" and "pedagogical content knowledge" had a significant correlation of .45 and used the same scenarios, "pedagogical content knowledge" and "teachers as learners" had a significant correlation of .3 and used different scenarios.

Overall, these mixed correlations provided initial evidence for their being conceptually distinct: principals who showed higher expertise in one subdomain did not always show higher in expertise in another (the lack of correlation between "subject matter" and "teachers as learners" offers the strongest support for this). At the same time, these values offered mixed evidence of how well the subdomains represented a larger construct of expertise. Nunnally and Bernstein (1994) commented on such correlations: "a domain of items is of interest only if the average correlation among items is positive" (p. 228), and the average correlation (at .26) fits this criterion. In Table 32 we see significant correlations between "subject matter" and "teachers as learners" each to "pedagogical content knowledge," and these two correlations fell within the range of .15 to .5 that Clark and Watson (1995) argue are satisfactory for inter-item correlations. The correlations warranted an examination of their subdomains' overall reliability, and the alpha reliability score of .49 showed that the subdomains had a limited internal consistency in tapping a larger domain. Following researchers'

theoretical arguments that these three subdomains comprise a larger domain of expertise, I nonetheless generated an aggregate score by taking an average of the subdomains, and I used this overall score for the domain in the final substudy in this chapter.

Learning-centered Leadership

Table 37. Correlations for Learning-centered Leadership

	Data-based DM	Effective Teaching & Learning	Standards-based Thinking	Monitor Instructional Improvement
Data-based DM	1			
Effective Teaching & Learning	.48**	1		
Standards-based Thinking	.20	.56**	1	
Monitor Instructional Improvement	.28*	.06	0.21	1

N=43

*. Correlation is significant at the 0.1 level (2-tailed).

**. Correlation is significant at the 0.05 level (2-tailed).

Cronbach's Alpha: .61

As in the previous section, Table 37 shows the correlations for the learning-centered leadership subdomains. All the correlations except two fell in the range that Clark and Watson (1995) recommended, and the correlations between "effective teaching and learning" and "data-based decision making," "standards-based thinking," and "monitoring instructional improvement" were statistically significant. The average correlation between the subdomains was positive at .3, which fit Nunnally and Bernstein's (1994) simple threshold for being a "domain of interest." While the mixed correlations offered evidence that the subdomains were conceptually distinct areas of expertise, their shared

variance (as high as .31 between "effective teaching and learning" and "standards-based thinking") suggested conceptual connections between them that warranted a review of their internal consistency. The Cronbach's alpha of .61 showed a marginal reliability: taken as a whole these subdomains may comprise a scale measure of the larger domain of learning-centered leadership. Because this value more closely approached the commonly accepted threshold of .7 (Peterson, 1994) it offered some statistical support for computing a larger average score for this domain.

Problems-solving Expertise

Table 38. Correlations for Problem Solving Expertise

	Planning	Gather Info	Delegate Authority	Address Conflict
Planning	1			
Gather Info	.36**	1		
Delegate Authority	.14	-0.04	1	
Address Conflict	.39**	0.22	0.04	1

N=43

*. Correlation is significant at the 0.1 level (2-tailed).

**. Correlation is significant at the 0.05 level (2-tailed).

Cronbach's Alpha: .49

Table 38 offers evidence that the subdomains for problem-solving expertise were distinct; four of the six correlations fell within or approach the recommended range of .15 and .50. The subdomain "gathering information" correlated significantly with both "planning" and "addressing conflict," but two of the correlations were close to zero or negative (between "delegate authority" and both "gather info" and "address conflict"). These correlations provided little

evidence for a scenario bias in which scenarios prompted heavily for more than one subdomain of expertise: while "planning and goal setting" and "delegate authority" used the same selected scenarios, there was no significant correlation between them. Nonetheless, the average correlation between these items was positive at .19. As with leadership content knowledge, the alpha reliability score of .49 provided little evidence for the internal consistency of these subdomains to measure the larger construct of problem-solving expertise. Relying primarily on researchers' theoretical arguments that these comprise the larger domain of "problem-solving expertise," I generated an overall score for the domain by taking an average of the subdomain scores, and I used these results in the next substudy.

Summary

To summarize the findings from this substudy, limited correlations between the subdomains within all three primary domains provided evidence that these were conceptually distinct from one another, but they raised questions about whether or not all of these helped to capture the larger domains of expertise. Nunnally and Bernstein (1994) comment on correlations between items in a scale: "if the average correlation is zero or near zero, the items as a group have no common core" to measure a larger construct (p. 228). Low correlations within each of the three groups offered evidence that particular subdomains may not comprise the larger common cores for the domains. Along with these mixed correlations, the low alpha reliabilities called into question whether or not each of these groupings was internally consistent in measuring the larger expertise areas. For all three areas of expertise, the low alpha

reliabilities offered statistical evidence that the average scores were not unidimensionsal measures for their larger domains, but they included subdomains in each that did not relate adequately to the larger group to generate a meaningful score for the overall domain. While researchers have argued that these particular subdomains comprise the three larger domains of expertise, the statistical analyses here question just how well some of them helped to measure these main categories. Two questions arise from these analyses. First, are there particular subdomains in each area that might be dropped because they do not contribute adequately to measuring the larger domains? Second, do the relationships between the subdomains raise questions about the larger relationships between the domains—although researchers have proposed these domains as distinct areas of expertise are there indeed significant relationships between them? I turn to these two questions in the last substudy for this chapter.

As stated above, I relied primarily on researchers' theoretical arguments that these were larger unitary constructs of expertise as the main justification to generate larger average scores for the domains, and I used those scores for the next substudy to examine the behaviors of the domains and to address the two questions I've raised above.

Substudy 2. C. Relationships Among the Main Domains of Expertise

In this last construct validation substudy I used the three overall average scores from above to examine the how the larger domains relate to each other. I asked specifically, what are the relationships among the three larger domains of expertise? As I summarized in the methods chapter, I predicted finding small or no correlations between the domains because of their conceptual distinctions in

the literature. Along with the correlations I have provided a final alpha reliability for the domain scores to examine their internal consistency as measures of a broader construct of expertise that encompasses all three domains. I first presented these results to examine the broader relationships between the domains, and I then looked more closely at the correlations between subdomains *across* the domains of expertise to explain these relationships.

Table 39 below shows the correlations and alpha reliability for the aggregated domain scores.

Table 39. Correlations Between Primary Domains

	Correlations Between Primary Domains		
	Leadership Content Knowledge	Learning-centered Leadership	Problem-solving Expertise
Leadership Content Knowledge	1		
Learning-centered Leadership	.44**	1	
Problem-solving Expertise	.34*	.70**	1

N=43

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Cronbach's Alpha: .74

These results show significant positive correlations between all three large domains of expertise. Two of the correlations fell within the range of .15 and .50 that Clark and Watson discussed as showing some internal consistency while still being distinct. While the third correlation between "problem-solving expertise" and "learning-centered leadership" of .70 was higher, this means that the two domains shared only 49% of their variance—roughly half of their variances were due to other factors. The alpha reliability score met a generally

accepted threshold of .7 (Clark & Watson, 1995; Peterson, 1994) for internal consistency, but one qualification is important for this score. A measure's reliability increases as it contains additional items (Nunnally & Bernstein, 1994), and this higher alpha reliability could certainly be due in part to the fact that it used the aggregations of the eleven subdomains discussed in the previous substudy. Therefore not all of the increase in this alpha coefficient was due to an increased internal consistency of the three subdomain measures. Nonetheless, the findings offered initial evidence that relationships between the different areas of expertise and their broader internal consistency are stronger than I hypothesized in the methodology chapter.

In light of the results in the previous substudy, however, a closer examination of the subdomain relationships *within* and *across* domains was necessary to explain in more detail how the mixed correlations and alpha coefficients in the previous substudy generated the higher correlations and alpha reliability coefficient for the aggregated scores in Table 39. Table 40 below shows the correlations between all the subdomains from the three primary areas of expertise.

		Leadership Content Knowledge			Learning-centered Leadership				Problem-solving Expertise			
	Subdomains	Subject Matter	Ped Content Knowledge	Teachers as Learners	Data-based DM	Effective Teach & Learn	Standards-based Think	Monitor Instruction	Gather Information	Planning	Delegate Authority	Resolve Conflict
Leadership Content Knowledge	Subject Matter	1.00										
	Ped Content Knowledge	.45**	1.00									
	Teachers as Learners	0.03	0.30*	1.00								
Learning-centered Leadership	Data-based DM	**0.14**	**0.25**	**0.16**	1.00							
	Effective Teach & Learn	**0.08**	**.53****	**0.27**	.48**	1.00						
	Standards-based Think	**0.01**	**.41****	**0.16**	0.20	.56**	1.00					
	Monitor Instruction	**0.24**	**0.05**	**0.17**	0.28*	0.06	0.21	1.00				
Problem-solving Expertise	Gather Information	**0.22**	**0.10**	**0.03**	**.67****	**.37***	**0.10**	**0.30***	1.00			
	Planning	**0.19**	**0.46***	**0.19**	**0.58****	**0.54****	**0.40****	**.32***	.36**	1.00		
	Delegate Authority	**-0.21**	**0.21**	**.37***	**0.07**	**.36***	**.51****	**0.12**	-0.04	0.14	1.00	
	Resolve Conflict	**0.06**	**0.02**	**0.22**	**0.27**	**0.01**	**0.17**	**.39****	0.22	0.39**	0.04	0.06
	N=43											
	**. Correlation is significant at the 0.01 level (2-tailed).											
	*. Correlation is significant at the 0.05 level (2-tailed).											
	Triangles: subdomain correlations *within* domains											
	Bolded Numbers: subdomain correlations *across* domains											

A comparison of the correlations both within and across the domains in this table helped to identify those stronger subdomain relationships across the domains that would explain the higher correlations between the aggregated domain scores in the previous Table 39. These relationships raised two additional questions: a) do some of the correlations between subdomains suggest that researchers have identified conceptually similar areas of expertise in their respective literature, and b) do the correlations between these subdomains question the conceptual distinctions that researchers have drawn between the primary domains of expertise? For example, if correlations between subdomains were stronger across the domains than within, one must ask if they might be might be better grouped into alternative constructs of expertise. I address these two questions in the remainder of this substudy.

First, given researchers' theoretical arguments about these three broader constructs, one would predict that subdomain correlations *within* the domains would be greater than subdomain correlations *across* the domains. With the data in Table 36, this would mean that correlations within the triangles would be greater than the bold-faced correlations. However, Table 36 shows numerous

cases in which correlations *across* the domains were higher. For example, correlations between a) "gathering information" and "data-based decision making" and b) "planning" and "data-based decision making" were both higher than any of the correlations within the domains "learning-centered leadership" or "problem-solving expertise." The correlation between the subdomains "effective teaching and learning" and "pedagogical content knowledge" was higher than all the subdomain correlations within "leadership content knowledge" and all but one of the correlations in "learning-centered leadership." Relationships such as these raise the question of how conceptually distinct some of these subdomains were. I present three examples below.

First, the correlation between the subdomains "gather information" and "data-based decision-making" was statistically significant at .67—the two shared just under half of their variance (.45). A review their definitions demonstrated the conceptual overlap for these two subdomains. Both referred to principals' understandings of a) the different types of information a principal would collect or use to make a decision, b) the role that such information would play in understanding a situation, and c) the value or importance of such information to a principals' decision making. Text from one principal's response to scenario 2 (in which students' math scores are dropping and teachers differ in their opinions of whether or not to continue with the math curriculum) demonstrated how he or she scored highly in both of these subdomains.

> If the new program is research-proven to be effective in developing student achievement in math, the first question to ask is was the program implemented completely and properly? What areas of student achievement in math are not improving or declining? What is the comparison of student achievement/performance in those areas with the new program versus the old program? What assessment is being used --- the same or a new assessment? Does the current assessment measure

> what is being taught? Is the lack of achievement a function of the absence of content knowledge or lack of practice with the assessment format? What were the research conditions for the new program?
> Armed with this information, a schoolwide effort, using grade groups which would come together for a faculty discussion, would plan for analyzing the problem and developing a plan of action.

In this response the principal offered detailed examples of information that would be crucial to addressing the situation. Reviewers scored this response with a "3" for both "data-base decision making" and "gathering information" because it included more developed discussions of both the data and information she or he would gather and use. In this case the similarities in definitions and the resulting correlations between the subdomains offered evidence that the two subdomains may be tapping a single construct that focuses on principals' understanding and use of data in their work.

Second, the correlation of .53 between "pedagogical content knowledge" and "effective teaching and learning" revealed similar overlaps for these subdomains. While their shared variance was only .29, the conceptual similarities between these two subdomains raised questions about whether or not they comprised separate constructs. "Pedagogical content knowledge" focused on teaching strategies and student learning theories for specific-subject areas (such as math and reading), while "effective teaching and learning" included broader theories of pedagogy and student learning. "Effective teaching and learning" would therefore encompass the subject-specific definitions in "pedagogical content knowledge" along with broader discussions, and this explains some of the correlation we see between these two subdomains. Again, text from a principal response demonstrated this overlap:

> I would recommend that the teacher demonstrate revising the sentence by writing in what students suggested then divide the students into small

> groups to revise one. The groups could then write their sentences on chart paper, post them and read them to the class. In this way many more students participate in the revision and each student sees many more examples. The next step would be authoring sentences that incorporate expressive language.

This principal's more developed recommendation for teaching a reading lesson included specific details about the lesson, a justification for its use, and a discussion of how it would relate to the next reading lesson. This response was scored as a "3" for both "pedagogical content knowledge" and "effective teaching and learning."

In brief, while most of the subdomains showed limited correlations across the domains which offered evidence that they were conceptually distinct, particular correlations and conceptual similarities between subdomains such as the ones above raised questions about whether some of these were conceptually distinct enough to warrant keeping them separate--or whether they should be collapsed them into more generally defined subdomains. Such stronger correlations across the domains help to explain in part why the larger domain correlations in Table 39 above showed such high correlations—but in these cases the scenario measures as defined may have tapped a similar construct in different domains.

A review of the scenarios used in the selected averages offered little evidence of "scenario bias" within or across subdomains that might have explained these correlations. While some selected averages shared the same scenarios within subdomains (such as "effective teaching and learning" and "standards-based thinking") or across domains (such as "monitoring instructional improvement" and "teachers as learners"), they frequently did not correlate significantly. Thus while some subdomains did use the same scenarios

in their selected averages, these scenarios did not prompt principals in a uniform pattern across those areas of expertise to bias their responses.

Finally, additional correlations across the domains raised the question of how these larger constructs were structured overall. I predicted that there would be no or low correlations between the subdomains across domains based on the argument that researchers have thus far developed these domains as conceptually distinct. The results in Table 40 did not match this prediction--many correlations in Table 40 indicated that while subdomains were conceptually distinct they nonetheless offered evidence of additional relationships across the domains that have not yet been examined. In the final discussion chapter I consider possible explanations and implications for these additional relationships between the domains.

Summary

Qualitative and descriptive summaries from the scoring illustrated that the scenarios varied in eliciting participants to demonstrate expertise in different areas. Analyses of correlations within the domains showed the subdomains were conceptually distinct but offered little internal consistency as scale measures of the larger domains. Finally, in substudy 2.c. higher and significant correlations between the three domains did not support my hypothesis that these primary areas of expertise were entirely distinct. A closer examination of all the subdomains' correlations in Table 40 provided evidence that a) some of them may be conceptually similar such that they should be collapsed and used as a single subdomain, and 2) there may be relationships between conceptually

distinct subdomains across domains that researchers have not yet examined. I return to these results in my final discussion.

CHAPTER VII

STUDY 3 RESULTS: EXAMINATIONS OF THE MEASURES' CRITERION VALIDITY

For this chapter I have presented the findings from the criterion validation study in which I examined relationships between the scenario measures and related principal and teacher survey reports of principals' expertise and practices. As in previous chapters I have structured the three substudies around the three larger domains of expertise, and in each I discuss relationships between the subdomains and their respective criterion measures. As summarized in the "methodology" section, I used scales from principal self-report and teacher surveys that asked principals and their teachers to report on principals' expertise in particular areas and how frequently they engaged in particular related practices. I hypothesized that principals who showed higher expertise in the scenarios would self-report having greater expertise in these areas and engaging in related practices more frequently. I also hypothesized that principals with higher expertise would have teachers who scored them higher in such areas on a survey and report that they engaged more often in related activities.

The closing discussion for this study examines its limitations. For example, significant questions remain about what exactly the principal survey questions of expertise capture (Goldring et al., 2008, questioned whether these surveys may capture a more declarative form of knowledge, as opposed to the expertise they demonstrate in their scenario responses). There are also questions about what other factors may influence the results of these different measures

(such as self-report bias from principal surveys, or method bias for both surveys). The final chapter of this study reflects on its limitations.

Substudy 3. A. Leadership Content Knowledge

Table 41 presents the correlations between the subdomain measures and the principal survey measures for this domain. In this first paragraph, I specify the correlations I expected to find. First, I hypothesized that principals who demonstrated greater expertise on the scenarios would self-report their expertise higher in related areas on the surveys. I therefore predicted finding higher correlations between principals' "subject matter" expertise scores on the scenarios and their self-reports of such expertise on the surveys. I also predicted higher correlations between principals' scenario scores for "teachers as learners" and their scores for the survey scale of "methods for creating learning cultures." As they possess greater expertise in helping teachers learn, I hypothesized that they would report knowing more about how to create cultures in which students and teachers learn.

Next I hypothesized that principals who demonstrated higher expertise on the scenarios would engage in these practices more frequently. I predicted that principals with greater expertise in supporting "teachers as learners" would therefore self-report engaging more frequently in steps to "encourage staff development."

Table 33 shows that there were no significant correlations with the principal survey criterion measures: principals who showed higher expertise on the survey responses did not provide higher self-reports for either their expertise or their engagement in related practices. Given the sample size (only 36

principals were included after missing cases were deleted), the correlations with self-reports of expertise (.22 and .21) showed a possible relationship, but even these offered evidence of shared variances of approximately only .04. There was no correlation between principals' scenario scores for "teachers as learners" and their self-reports of encouraging staff development. These results did not provide much support for the two hypotheses above.

	Table 41. Correlations Between Leadership Content Knowledge Scores and Principal Self-reports			
		Leadership Content Knowledge Subdomains		
		Subject Matter	Pedagogical Content Knowledge	Teachers as Learners
Principal Self-report of Expertise	Subject Matter	**0.22**		
	Methods for Creating Learning Cultures			**0.21**
Principal Self-report of Practice	Encourage Staff Development			**-0.04**

N=36

**. Correlation is significant at the 0.05 level (2-tailed).

*. Correlation is significant at the 0.10 level (2-tailed).

From the teacher surveys I first predicted that principals with greater expertise in supporting "teachers as learners" would have teachers who reported their having more expertise in "supporting teachers' professional development." I also hypothesized that teachers would report their principals engaging more frequently in those related areas where they have more expertise. For example, principals with greater expertise in supporting "teachers as learners" would have teachers who reported them engaging more frequently in such areas as

"developing teachers' capacity," "encouraging improvement in teaching," "interacting with teachers regarding instruction," and "showing interest in professional development" for teachers.

Table 42 shows the correlations between the scenario scores and teacher reports. While there are no significant correlations between these variables, particular correlations suggest that these relationships may be stronger than the principal reports. For example, principal expertise scores for "teachers as learners" had a correlation of .29 with teachers' reports of principal practices to develop teachers' capacity. Principal expertise scores for "teachers as learners" also had a correlation of .24 with teachers' reports of their principal interacting with their teachers to improve instruction. Given the sample size of 38, these results offer limited evidence of stronger relationships between the scenario scores and teachers' reports about their principals.

	Table 42. Correlations Between Leadership Content Knowledge Scores and Teacher Reports			
		Leadership Content Knowledge Subdomains		
		Subject Matter	Pedagogical Content Knowledge	Teachers as Learners
Teachers' Reports of Principal Expertise	Support Teachers' Professional Development			**0.15**
Teachers' Reports of Principal Practice	Evaluate Instruction		**0.16**	
	Develop Teachers' Capacity			**0.29**
	Encourage Improvement of Teaching			**-0.02**
	Interact with Teachers Regarding Instruction			**0.24**
	Show Interest Teacher Pro Dev			**0.11**

N=38

**. Correlation is significant at the 0.05 level (2-tailed).

*. Correlation is significant at the 0.10 level (2-tailed).

To summarize for this first domain, the mixed correlations provided limited evidence that teacher reports of their principals more closely related to the scenario measures than the principals' self-report. The higher correlations for the teacher reports of principal practice were only slightly higher than principals' own reports in the surveys.

Substudy 3. B. Learning-centered Leadership

Table 43 below summarizes the principal survey variables I used to examine the criterion validity for the learning-centered leadership scores. I hypothesized that principals with higher scenario scores would self-report

higher levels of expertise in the same areas. For example, principals who showed greater expertise in data-based decision making on the scenarios would self-report having higher expertise on the survey. I also predicted finding higher correlations between the scenarios and principals' reports of their practices: those with greater expertise in the subdomains would engage more frequently in related activities.

Table 43. Correlations Between Learning-centered Leadership Scores and Principal Self-reports

		Learning-centered Leadership Subdomains			
		Data-based Decision Making	Effective Teaching & Learning	Monitoring Instruction	Standards- Based Thinking
Principal Self-report of Expertise	Data-based Decision-making	***0.12***			
	Principles of Effective Teaching & Learning		***0.27***		
	Monitoring Instructional Improvement			***0.10***	
	Standards-based Thinking				***0.05***
Principal Self-report of Practice	Data-based Decision-making	***-0.04***			
	Examine Student Work		***-0.13***		
	Monitor Instructional Improvement			***0.08***	

N=36

**. Correlation is significant at the 0.05 level (2-tailed).

*. Correlation is significant at the 0.10 level (2-tailed).

As shown by the results in Table 43, there were few relationships between the scenarios and principals' self-reports of their expertise and practice. The correlation between principals' scenario scores for "effective teaching and learning" and their self-reports for this domain was highest at .27, but even this was not significant. Negative correlations between the scenario scores and principals' reports of practice (as with data-based decision making) offered

further evidence of little or no relationship between the scenario and principal practice report score. These results offered no evidence to support my hypotheses that principals who showed greater expertise would self-report higher expertise or engagement in these types of activities.

Next I hypothesized that if principals demonstrated higher expertise through the scenarios they would have teachers who reported their possessing more expertise in those subdomains or their more frequent engagement in related activities. For example, principals with greater expertise in "monitoring instructional improvement" would have teachers who reported that they more frequently participated in actually "monitoring instructional improvement" in classrooms.

Table 44. Correlations Between Learning-centered Leadership Scores and Teacher Reports

		Learning-centered Leadership Subdomains			
		Data-based Decision Making	Effective Teaching & Learning	Monitoring Instruction	Standards-Based Thinking
Teachers' Reports of Principal Expertise	Effective Teaching and Learning		**.56****		
Teachers' Reports of Principal Practice	Monitor Instructional Improvement			**0.36***	

N=38

**. Correlation is significant at the 0.05 level (2-tailed).

*. Correlation is significant at the 0.10 level (2-tailed).

These results provided the best evidence thus far of stronger relationships between the scenario scores and teachers' reports of their principals. Teachers' reports of their principals' expertise in "effective teaching and learning" correlated .56 with the scenario scores, and teachers survey reports of their

principals' activities to monitoring instructional improvement correlated .36 with the scenario results. Both of these correlations were statistically significant. The respective shared variances of .31 and .13 for these correlations offered some evidence that the two different methods are measuring a similar construct.

To summarize the findings for "learning-centered leadership" there were no significant correlations between the scenario scores and principals' self-reports of their expertise or practices. Principals who scored higher on the scenarios did not self-report higher expertise or more frequent practices on the surveys. These results mirrored those of the domain "leadership content knowledge"—these low values provided no evidence that the different measures were capturing a similar construct of expertise. However, stronger correlations between the scenario scores and the teacher surveys demonstrated that for principals who scored higher on the scenarios, their teachers more frequently rated them higher in related areas of expertise or reported that they engaged more frequently in related activities. These results suggested that teacher observations may better capture the leadership expertise constructs that the scenarios measure. This finding matched that of Goldring, et al. (2008) who presented similar correlations between the scenarios and teacher survey measures.

Substudy 3C: Problem-solving Expertise

In Table 45 I have summarized the findings for the principal surveys for the subdomains of problem-solving expertise. As in the two previous domains, I hypothesized that principals higher in expertise in the scenarios would self-report greater expertise and more frequent practices on the surveys. Therefore

those school leaders with higher scenario scores for "gathering information" would report that they have greater expertise in this area. Likewise, principals who showed higher expertise for "planning" would describe themselves as higher in this area and report that they engaged in "planning" practices more frequently.

Table 45. Correlations Between Problem-solving Expertise Scores and Principal Self-reports

		Problem-solving Expertise Subdomains			
		Gather Information	Delegate Authority	Planning	Resolve Conflict
Principal Self-report of Expertise	Gathering Information	**0.24**			
	Planning			**0.01**	
Principal Self-report of Practice	Planning			**-0.14**	

N=36

**. Correlation is significant at the 0.05 level (2-tailed).

*. Correlation is significant at the 0.10 level (2-tailed).

As in the previous two domains, the low correlations (.24 and .01, respectively) offered little evidence that the scenarios and principal reports of expertise captured similar constructs. None of the correlations was significant, and even the highest correlation of .24 for the scenario and self-report expertise scores for "gathering information" had very little shared variance (.06).

I predicted finding higher correlations between the scenario scores and teachers report of their principals engagement in related activities. Teachers

whose principals showed higher expertise in "delegating authority" in the scenarios would therefore report that their principals engaged more frequently in "encouraging teachers to take responsibility" or "distributing leadership in meetings." Likewise, teachers with principals with higher scenario scores in "planning" would report that they engaged more frequently in "planning" activities. Finally, principals with higher scenario scores for "addressing conflict" would have teachers who reported that they more frequently were "open to discussion."

Table 46 shows the results of these analyses. Principals' scores of "delegating authority" had a correlation of .38 with teacher reports of how frequently they "encouraged teachers to take responsibility (this value was statistically significant). The correlation between this same subdomain and teachers' reports of principals' frequency of "distribution of leadership" was lower at .22. The correlation between principals' scenario scores on "addressing conflict" and their "openness to discussion" was .30. Those principals with higher scenario had teachers who were more likely to report that they engaged in related activities. While still relatively low, these higher correlations offered similar evidence to the teacher reports in the previous domain—teachers' reports of their principals' expertise and related practices are more likely to tap the same constructs as those measured in the scenarios.

	Table 46. Correlations Between Problem-solving Expertise Scores and Teacher Reports				
		Problem-solving Expertise Subdomains			
		Gather Information	Delegate Authority	Planning	Address Conflict
Teachers' Reports of Principal Practice	Encourage Teachers to Take Responsibility		**0.38***		
	Distribute Leadership in Leader Meetings		**0.22**		
	Planning			**0.26**	
	Openness to Discussion				**0.30**

N=38

**. Correlation is significant at the 0.05 level (2-tailed).

*. Correlation is significant at the 0.10 level (2-tailed).

Like the domain of "learning-centered leadership," these correlations again provided little evidence for a relationship between the scenario scores and the principal survey scores. Principals with higher scores on the scenarios were not more likely to self-report having higher expertise or engaging in related practices. However, the teachers' reports of principal practices, however, offered greater evidence of their correlation with the scenarios. The correlations were higher (as with the scenario scores for "resolving conflict" and principals' practice of being "open to discussion") and in some cases statistically significant (as between "delegate responsibility" and "encouraging teachers to take responsibility"). These relationships offer additional support for relationships between the scenarios' measures of expertise and teachers' reports regarding their principals.

Summary

To summarize the findings in this criterion validation study, low correlations across all three primary domains offered little evidence that the

scenarios and the principal self-reports of expertise and practice measured similar constructs. Principals with higher expertise as measured in the scenarios were not more likely to self-report having higher expertise on the surveys or to self-report engaging more frequently in related activities. While I predicted that higher correlations between the scenarios and principal survey results would offer evidence that they were both measuring leaders' expertise, the low correlations indicated that the two different measures are capturing different constructs of expertise. Two factors may help to explain the differences. First, as Goldring, et al. (2008) discussed, principal self-reports may certainly be subject to self-report biases that influence each individual's ratings of him- or herself. Second, the differing methods of measurement (scenario versus survey) may explain differences in these scores. I discuss the role of these factors more closely in the final chapter.

Results from the "learning-centered leadership" and "problem-solving expertise" domains supported a final proposal that Goldring, et al. (2008) offered: that the teacher surveys may indeed be better at capturing the constructs of leadership expertise embedded in the scenarios. While the correlations in these two areas did not provide complete support for this conclusion (a number of them are low and statistically not significant), a number of these in both domains suggested that the teacher survey and scenario measures may have shared variances ranging from .13 to .31—preliminary evidence that the two different types of measures are tapping the same construct. As with the principal self-reports, I discuss the implications of these results and consider possible explanations for them in the final chapter.

CHAPTER VI

DISCUSSION

The findings in this dissertation relate to two primary areas: the conceptual nature of educational leadership expertise as it has been defined in the literature thus far and the validity of the measures designed to capture educational leadership expertise. In this final chapter I first review how the findings relate to the central questions I have posed in the study before I discuss how they also inform our broader understanding of the nature of educational leadership expertise and our efforts to measure it.

How the Findings Address the Research Questions

This study explored a series of scenarios that measure educational leadership expertise by looking at their validity and reliability. It focused on three primary questions to examine the content, construct, and criterion validity of the proposed measures of leadership expertise in this study. Overall the analyses showed promising yet mixed results for their validity and reliability and suggested multiple directions for future improvements and uses of these measures. I return to each question of validity to discuss how this study answered each of them, and I discuss the limitations and implications for the findings before specifying future steps that would advance the use of scenarios to capture school leaders' expertise.

Study 1: Content Validity

The first study evaluated the content validity of the proposed subdomain measures by soliciting reviews from a panel of content experts. These reviews asked the experts to examine how well each measure covered the range of meanings included under each domain of expertise, and they solicited recommendations for how to modify the rubrics to better capture these domains of expertise.

Content experts' comments highlighted the need to modify both the rubrics' definitions of content and their examples to demonstrate the different subdomains. Variations in their scores of expert principal responses underscored these comments: the rubrics as initially proposed did not guide content experts to assign scores with high inter-rater reliability. Their varying scores illustrated not only the need to modify the rubrics but also the complex nature of evaluating principals' expertise through more graduated scales that captured "quality of response." The content experts' scores I reviewed in the first study demonstrated the difficulty of attaining high agreement without an arbitration process or further discussions between raters to reconcile differences in understanding or interpretation.

Content experts' comments helped to refine the rubrics and promote raters' shared understandings of the expertise constructs covered in this study. The changes I made to the rubrics were essential to providing more content valid measures for the subdomains. The final results of the first study were rubrics for each subdomain that were clearer and that more fully captured the content of each subdomain. Ultimately, however, experts' differing scores also highlighted

the need for raters to use an arbitration process in which they clarified any misunderstandings or disagreements and resolved differences in the scores.

Study 2: Construct Validity

The second study evaluated the construct validity of these subdomain measures by asking, do these measures of leadership expertise relate to each other as predicted by theory? I first presented qualitative and descriptive evidence from the coding that demonstrated that the rubrics did indeed capture levels of expertise in principals' answers according to the quality and content of their comments—those individuals who provided more elaborate discussions of expertise received higher scores. While these scores derived in part from the number of times they mentioned a concept (or "frequency of mention"), only those who offered more developed explanations of how they employed their expertise scored higher according to the rubrics.

Next I hypothesized that there would be low to moderate correlations but a high alpha reliability coefficient for the subdomains in each domain. Such results would provide evidence that, while the subdomains in each domain were distinct conceptually (and therefore did not correlate highly), as a whole they comprised an internally consistent scale for each domain (and thus returned higher alpha reliability coefficients). Low to moderate correlations between the subdomains supported the first part of my hypothesis, but the lower alpha reliabilities for each of the domains provided only mixed evidence for the second part of the hypothesis. While the results suggested that the subdomains in each domain were distinct, they offered limited evidence that they comprised internally consistent scales for each of the domains. Furthermore, low

correlations for particular subdomains in each main area (such as for "teachers as learners" in leadership content knowledge or for "delegate authority" in problem-solving expertise) called into question how well some of them helped to measure their respective larger domain. These findings raised further questions about the domains as researchers have structured them—*are* all the subdomains that researchers have included in these domains crucial to tapping individuals' expertise in these broader areas? And just how do the domains relate to each other? This led to the final examination of subdomain relationships across the primary domains in the third substudy.

Because researchers have developed these domains as separate ways to view expertise, I predicted that there would be low correlations between the three larger domains. However, analyses showed moderate and statistically significant correlations between the aggregate scores for each domain. The alpha reliability coefficient for these aggregate scores was .74, and it was statistically significant. Although some of this higher alpha coefficient may have been due to the greater number of items included in its calculation, the results suggested there were stronger relationships between the three primary domains than I predicted. Closer analyses of the subdomains helped to explain these broader relationships. Subdomain correlations *between* the domains were often equal to or greater than subdomain correlations *within* the domains, suggesting that these larger domain constructs were not as distinct as researchers have implied in distinguishing these different areas of expertise. Such correlations (as in Table 40) suggested that a) some of the subdomains might be conceptually similar enough that they should be collapsed into a single construct, and 2) there may be

relationships between conceptually distinct subdomains across domains that researchers have not yet examined in full.

Thus while the results of this second study did not fully support all my theoretical predictions, they offered evidence that 1) the individual subdomain measures did successfully capture differing levels of expertise, and 2) particular subdomains (though not all) were central to capturing leaders' expertise in the broader domains. Finally, significant relationships between the different domains offered evidence of much more complex relationships in principals' expertise than what the three main domains in this dissertation covered.

Study 3: Criterion Validity

The third study explored the criterion validity of the expertise measures by asking, how do these relate to other measures of principals' expertise and practice? While I predicted that on surveys principals with higher scenario scores would self-report higher expertise in related areas and more frequent engagement in related practices, analyses showed no correlations between the measures—principals higher in expertise tapped through the scenarios did not self-report higher expertise or more frequent practices. These results indicated that the scenarios and principal surveys may be capturing different constructs, or that one is more subject to biases that influence the scores. However, stronger correlations between the scenarios and teacher surveys indicated that principals who score higher in their scenario responses are more likely to have teachers who rate them as higher in expertise or report that they are more frequently engaged in particular activities related to expertise. These results suggested that the two measures may be tapping similar constructs of expertise. Thus while the principal surveys offered no criterion validation of the scenarios, teachers'

survey reports of their principals' expertise and practices offered stronger evidence for the criterion-related validity of the scenarios.

Limitations of the Research

A close examination of this study's limitations helps to qualify the findings above; I review the most prominent challenges in this section and examine how the limitations might influence the results. Later I discuss how future work might address some of these concerns.

The first challenge for this study focuses on the extent to which the scenarios prompted principals to demonstrate their full expertise in the different subdomains I included. Evidence for this limitation lay in two areas: a) the low scores that principals generated in their written responses to the scenarios, and b) content experts' concerns about the extent to which the scenarios elicited demonstrations of expertise. This issue relates directly to the validity of the scenario measures—did they actually produce valid evidence of principals' expertise? While the low scores might have resulted from the sample principals being low in expertise, principals' low scores across almost all of the subdomains necessitated a closer examination of the scenarios' content. As I summarized earlier, even the "selected average" scores for each subdomain generated low scores with limited variation—the mean scores and standard deviations were at the highest at 1.36 and .78 respectively, on a possible scale of 5 (both these scores came from the subdomain "data-based decision making"). Out of all the selected averages, the highest score an individual received was 3 out of 5 (different principals scored this for the subdomains "data-based decision making," "gather information," and "address conflict"). A second source of evidence for the

limitations is that four different content experts questioned the scenarios' ability to prompt for the different subdomains, even after they read responses from principal experts. A first content expert's comment about the scenarios for "effective teaching and learning" illustrated their concerns: "I'm not sure your prompts lead participants to the level of detail [you were looking for]." A second reviewer's comments addressed the potential limitation more directly:

> If my scores were 'reliable' in the sense of similar to other assessors, then I think your scenarios are not suitable as they stand. The scenarios should yield greater variation in responses; yet these were almost uniformly low. It may be that you need a bit more information in the scenarios…, or alternatively you may need to provide more probes in the interview process to elicit the 'depth' of information that you are looking for. However, as it stands I just don't see the depth of information present that would be needed to meet the upper levels of the standards implied in the rubric.

If true, this limitation raises two issues for the study. First, at a basic level, if the scenarios have not prompted adequately for expertise, then the resulting scores and analyses are based on only partial demonstrations of principals' expertise. While there are measurement errors inherent to using any instrument, such a systematic error as this calls into question the validity of the findings. Second and more specifically, a restriction of range for the scenario scores would most likely decrease their correlations with other variables (Allen & Yen 1979; Nunnally & Bernstein, 1994). Such an attenuation due to restriction of range would affect both the results of studies 2 and 3 which examine the subdomains' correlations with each other and with other criterion measures: the lower correlations we see in both of these studies may be due in part to the restricted ranges of responses that the scenarios generated.

A second limitation of this study focuses on differences between the scenarios and the survey instruments used as criterion measures in this. While I

initially hypothesized that the scenarios and survey items measured similar constructs of expertise, mixed correlations between the measures indicated otherwise. The absence of any correlations between the scenarios and principal surveys suggested large differences between the measured constructs for the instruments, while low correlations between the scenarios and teacher surveys underscored that the constructs were similar but certainly not identical. A number of differences in the instruments could help to explain these outcomes.

First, the surveys were administered to principals and teachers in the spring of the school year. In asking them to report on principals' expertise and related practices the surveys required respondents to aggregate their perceptions of the principals from the past year (or as long as they had known them). Item responses on the surveys thus represented a cumulative perception of the principal over time. On the other hand, the scenarios required principals to respond to particular circumstances with the expertise they possessed at that specific time. The instruments therefore differed in the perceptions they tapped from the respondents, and such differences would decrease the correlations between the measures.

Second, these instruments may also have differed in the type of expertise they measured. The results raised this question for the principal surveys and scenarios in particular. The scenarios required participants to apply the expertise they possessed in addressing the situations, and they arguably tapped the tacit, practical knowledge of each respondent. The principal survey, on the other hand, required individuals to identify or *declare* their expertise in a Likert scale response—this instrument may therefore have measured principals' declarative knowledge, that which they say they know. As emphasized in the literature both

within and outside of educational leadership, researchers have drawn distinctions between the knowledge that individuals say they know and what they use in addressing a situation (Wagner & Sternberg, 1985; Eddy, 1988; Leithwood & Steinbach, 1995). Again, such differences in the constructs measured by these instruments would contribute to lower correlations between the scenario and survey results.

A final limitation of this study centers on the generalizability of the findings to actual principal practices. Because the scenarios required participants to write their theoretical responses to particular situations, it is still not clear how closely the scenarios captured what principals actually do in their work. As Stecher, et al. (2006) argue, previous psychological research has shown that intentions can predict behavior (p. 103), and such work would suggest that intentions embedded in participants' scenario responses reflect their actual behavior. However, while these measures arguably move closer to principals' practical work in the sense that they ask respondents to address specific realistic conditions (and are thus closer to actual events than a principal's survey reports), they are nonetheless written predictions of what principals say they would do. These measures may also favor those principals who are better able to write their plans or reactions on paper—certain individuals may be high in educational leadership expertise and yet may not write much about their strategies or actions in the scenarios. In light of these limitations additional work is still warranted to examine the validity of scenarios to measure principal practice.

Conceptual Implications for the Field

The Complexity of Leadership Expertise

First and foremost, the correlations we see in Tables 39 and 40 in this study demonstrate that as a field we are still wrestling with how to define expertise for educational leaders. The strong correlations and alpha reliability of .74 for the primary domains in Table 39 suggest that individuals who were high in one area tended to be high in the other areas; these broader definitions were thus helpful in identifying individuals who were high or low in expertise across the three areas. However, the results in Table 40 (which showed that subdomain correlations across the three domains were equal to and sometimes higher than correlations within the domains) illustrate numerous conceptual overlaps and relationships between these different domains that have yet to be examined. Taken together, the results offer evidence that while we've identified broader categories that help us distinguish between experts and non-experts in general, we are still uncertain of the specific content of these domains and how they relate to each other. These findings underscore the need to re-visit the specific structures that researchers have used thus far to conceptualize educational leadership expertise. Little if any dialogue has occurred to compare these different lines of research and consider not only if some constructs are the same across domains, but also if alternative domains or groupings of expertise help to explain these different domains' relationships to each other. For example, do principals' skills as "instructional leaders" better explain their expertise, so that we see principals with a greater understanding of "subject matter" also higher in areas such as "effective teaching and learning" and "standards-based thinking"? Do different understandings of principals' organizational roles better explain

why principals who score higher in "data-based decision making" also score higher in "planning" and "monitoring instructional improvement"? These results emphasize the need for content experts to examine these conceptual distinctions and consider if or how we might better define expertise in its complexity.

A return to the literature behind these three domains helps to explain in part the distinctions they have drawn thus far, and it illustrates the need to re-examine their relationships to one other. For "problem-solving expertise" Leithwood, et al. (1986, 1989, 1993, and 1995) based their work on an effort to open the "black box" of administrators' cognitive strategies in completing their work. They argued that a cognitive approach would provide a deeper understanding of the mental analyses that guided principals' practices, and they borrowed heavily from existing research such as Schon (1987) and others that examined individuals' cognitive processes. Their findings focused primarily on the cognitive differences that they saw between expert and non-expert principal responses to various scenarios—they reported differences in the skills that respondents actually possessed and demonstrated in the studies.

The origin of this work contrasts with Stein and Nelson's (2003) research on "leadership content knowledge," which they base heavily on Shulman's (1986) pedagogical content knowledge. Rather than focus on principals' cognitive strategies, Stein and Nelson develop this domain to define what practical knowledge of subject matter and pedagogy principals need in their work as instructional leaders—part of the purpose for their work is to advocate for the expertise that leaders should have. Their purpose differs markedly from Leithwood, et al., and it takes them to significantly different literature to define

leadership expertise. Stein and D'Amico's (2000) discussion of this domain illustrates the efforts in their work to advocate for particular skills in school leaders.

> In order to provide intellectual leadership for instruction, principals and superintendents must understand the manner in which classroom practices and curricular programming differ in mathematics vs. literacy, as well as the different needs that teachers have with respect to each subject area.

Stein and Nelson (2003) echo this effort to define the skill that leaders should have to guide their schools effectively:

> …as demands increase for them to improve teaching and learning in their schools, administrators must be able to know strong instruction when they see it, to encourage it when they don't, and to set the conditions for continuous academic learning among their professional staffs. (p. 424)

This differs markedly with Leithwood, et al., which used an emergent analysis of principal interviews to identify the distinctions they saw in principals' interview responses. Leithwood, et. al's work arguably started with a closer examination of the actual skills that leaders possessed. While work on leadership content knowledge (such as Nelson, et al., 2003 and 2005) has begun to explore the nature of what principals actually possess for this expertise, it has focused primarily on knowledge of mathematics subject matter and beliefs about mathematics pedagogical strategies. Thus while the initial literature for this leadership content knowledge has consisted more of describing the expertise that leaders need to have, it has not examined all three subdomains. Much work remains to understand the nature of leadership content knowledge that administrators and other school leaders actually possess.

The origins of "learning-centered leadership" differ from the two previous domains in that Murphy, et al., (2006) and Goldring, et al., (2009) derive their

different subdomains from extensive reviews of literature regarding the roles of leaders in successful schools. Murphy, et al's. (2006) explanation of their literature search illustrates this.

> If leadership is indeed a hallmark element of school performance, it seems appropriate that we begin by corralling the type of leadership behaviors found in the literature on effective schools and school districts….For the most part, we culled information from empirical studies of effective schools, school improvement, and principal and superintendent instructional leadership (pages 7 and 8).

This particular article that develops "learning-centered leadership" consists primarily of advocating specific subdomains that the literature supports. It demonstrates how this literature depends heavily on studies that tie school leaders to the broader organization and structures of effective schools (as opposed to cognitive processes, or subject matter and pedagogy). This scope focuses the researchers on different literature than that of the other two domains. While the two studies cited above for "learning-centered leadership" draw from research that shows what leaders do in effective schools, only Goldring, et al. (2009) go further to examine the content of this expertise with primary research of principal responses. While work in this domain has thus far documented the expertise that successful leaders possess, additional research is necessary to understand the nature of expertise across a broader spectrum of leaders.

Two primary points emerge from these domains' contrasting conceptual origins and the differing research surrounding each. First, the different conceptual and empirical roots of these three help to explain the theoretical distinctions we see between the subdomains, and they help to explain how these domains may contain subdomains that are conceptually similar but with named differently (such as "data-based decision making" and "gathering information").

As researchers have focused on their own areas, little work has been done to examine just how these relate, and whether or not they possess similar constructs. This study helps to start this discussion by demonstrating some of those conceptual overlaps between the different domains. Second, much less research has been done with "leadership content knowledge" and "learning-centered leadership" to understand the actual structures of what expertise principals possess and use in these areas. Along with Goldring, et al. (2009) and Nelson, et al. (2003 and 2005), future work can help to determine just how much principals possess the expertise that the literature advocates for these two domains. At present, these limitations in the literature help to explain in part the relationships we see between the domains.

What Do Instructional Leaders Need to Know?

As researchers propose different areas of expertise and advocate their importance for school leaders, they must also consider how much principals actually understand and use them. While selected average scores for all the subdomains were low, principals scored lowest in the "leadership content knowledge" domain (see tables 30, 32, and 34). Principals' responses not only generated the lowest scores across this domain, but they also included the fewest discussions or demonstrations of this expertise. While one cannot generalize broadly to other principal populations, these findings nonetheless raise the question of just how much expertise in "leadership content knowledge" the sample principals actually possessed. For all the discussion in instructional leadership of how essential it is for principals to dialogue directly with their teachers about their curriculum and teaching, studies have questioned how

much they actually do this or know how to do this. For example, different works have shown how principals have less expertise in the subject areas than their teachers who teach them (see Bossert, Dwyer, Rowan, & Lee, 1982). As the field has advocated that principals engage in instructional leadership in their schools, questions remain about whether or not they shifted their practices to target learning more directly. Hallinger (2005) wrote in his recent review of instructional leadership studies that "there is little evidence to support the view that on a broad scale at either the elementary or secondary school level principals have become more engaged in hands-on directed supervision of teaching and learning in classrooms" (p. 230) and he noted "the absence of any empirical evidence that principals spend more time directly observing and supervising classroom instruction than they did twenty-five years ago" (p. 233). We thus often see differing pictures between what expertise the field advocates and what practitioners actually possess. The scores in this study for "leadership content knowledge" suggest such a disparity between theory and practice, and they raise the question of whether or not principals possess the more detailed subject matter and pedagogical content knowledge that Stein and Nelson (2003) describe.

There are two possible explanations for principals' limited expertise in this area. First, Stein and Nelson (2003) have only recently published their piece that describes "leadership content knowledge," and Nelson, et al. (2004 and 2005) have just started to examine what levels of content knowledge principals possess. As discussed above, papers in this area have so far primarily argued for the importance of such expertise and presented only initial findings. The authors' advocacy for principals to possess such extensive expertise has thus had limited

time to influence researchers and the dialogue about essential leadership expertise, let alone practitioners and those who train school leaders. Such limited expertise as I found here may be a result of the field not having time to respond to their call to equip leaders with such expertise.

A second explanation comes from those who advocate that instructional leadership includes much broader roles that principals play in organizing their schools as a whole to focus on improved teaching and learning. Hallinger (2005) writes that if one focuses primarily on principals' direct involvement in teaching and learning, "the classroom doors appear to remain as impermeable as a boundary line for principals in 2005 as in 1980" (p. 230). He writes, however, that dimensions such as "defining a school mission" and "creating a positive school culture" have become more deeply integrated into principals' responsibilities and understanding of instructional leadership (2004). Stressing principals' direct connections and involvement in classrooms misses the larger organizational responsibilities that principals have in focusing their schools as a whole in improved instruction. In light of such roles, it may be unreasonable for most principals to possess the more specialized, subject-specific expertise that Stein and Nelson (2003) have proposed. Some principals may bring detailed, subject-specific expertise from their previous positions as teachers in different subjects, but many may not have this detailed expertise (or the time to develop it) in light of the broader organizational roles they must play. While principals may have more general knowledge of effective teaching and learning (as Goldring, et al., 2009 proposed), these individuals may also rely on others with more subject-specific content knowledge to guide efforts to reform curriculum and instruction in their schools.

To summarize, while further research is needed to understand the level of detail that principals understand about particular curricula, these results provide evidence of their limited expertise in "leadership content knowledge." They also push the question of how much of this more content-specific expertise leaders need or possess in light of their broader organizational roles in schools.

Methodological Implications for the Field

There is a constant need for researchers to connect their work to practitioners' realities. Murphy (2006) and others have criticized the field for its past neglect of the problems practitioners faced; such omissions in the research have previously resulted in school leadership theories that missed the complexity of leaders' experiences and were rarely helpful in helping them understand or lead schools (Mulkeen & Cooper, 1989). These gaps between research and practice are no more obvious today than in the debates to reform principal certification and professional development programs. As I discussed in my introduction, the critiques of leadership training programs suffer from a lack of rigorous measures to assess their impacts. Both scholars as well as the program administrators themselves have offered little beyond graduate self-report surveys or curricular analyses of programs. Few if any measures exist to capture the expertise that school leaders obtain through different training programs or through their experiences in schools. The result, as Murphy (2006) notes, is that "there are no research articles in the leading journals in the field over the past quarter century that directly address the skills and knowledge gained in preparation programs" (p. 71).

And yet there is evidence of the field's recent efforts to respond with more

sophisticated measures. Pounder argues that Murphy's 2006 review and above comments overlook changes in administrative programs that use case studies or simulations in their teaching and evaluation as well as broader efforts by the University Council of Educational Administration and the National Association of Secondary School Principals to develop rigorous curricula and assessments for such programs (Pounder response to Murphy in Murphy, 2006, p. 90-91). The Educational Testing Service for a number of years has now offered its School Leadership Licensure Assessment that relies heavily on short and long vignette responses that are scored according to the Interstate Leadership and Licensure Consortium Standards (ETS, 2005). In addition, Goldring, et al. (2009) have reported findings from their uses of scenarios to measure expertise. While few of these initiatives have published findings in peer-reviewed journals, these illustrate ongoing efforts to address the existing gaps in the field.

There is thus a continuing need to examine the roles that new methods can play in evaluating what leaders know. As the field pursues more rigorous measures of leadership expertise to understand what successful leaders know and how they use this information, this study's findings contribute to this research in two ways. First, they provide evidence that carefully designed scenarios can indeed tap school leaders' expertise, and second, they suggest that new insights into expertise may also be gained by viewing the results of these scenarios alongside other measures to obtain a more complete picture of an individual's expertise.

First, the findings in Study 2 demonstrate the scenarios' potential to capture the expertise that leaders use. This evidence comes through variations in the responses that principals provided for the scenarios as well as in

relationships between the scenario scores and other criterion variables. Just as Leithwood, et al's (1986, 1989, and 1995) and Brenninkmeyer and Spillane's (2008) work with scenarios showed differences between expert and non-expert principals, this study demonstrated that principals varied in the expertise they demonstrated through their responses. However, while both the previous two lines of research used dichotomous measures (those studies examined statistical differences in whether or not principals mentioned particular dimensions, not the depth to which they discussed them), this study captured leaders' quality of responses across more graduated scales (0-5) and demonstrated these qualitative differences in high- and low-expertise responses.

While these graduated rubrics helped to tap more nuanced levels of expertise, their results also show the need to tailor scenarios more tightly to the different domains. As I reported earlier, scores across all the rubrics were quite low, which raised the question of whether or not they prompted principals to offer full demonstrations of their expertise. There is certainly promise for scenarios such as these to capture the more practical knowledge that principals use in their work, but additional reviews of the scenarios themselves would be needed to elicit better demonstrations of expertise. Reviews by content experts and principal practitioner experts would help to evaluate and develop new scenarios that more closely measure respondents' expertise. Stecher, et al. (2006) offer guidance for such a process by describing detailed steps to assemble experts who devised and reviewed the vignettes they used to measure reform-oriented instruction in mathematics.

The second methodological contribution of this dissertation focuses on the use of scenarios alongside other measures. Findings on such analyses have come

primarily from outside of the educational leadership field. As discussed in the methods section, two recent studies that examined teachers' knowledge of mathematics and subject matter (Stecher, et al. (2006), and Kersting, 2008) both reported finding high and statistically significant correlations between their subjects' scenario scores and criterion measures such as daily logging mechanisms, observations, and subject matter tests. Respondents who scored higher on scenario measures of teaching mathematics also practiced these concepts more frequently in observations, they reported engaging in them more on logging mechanisms, or they reported higher engagement on self-report surveys. Both studies presented these results as initial criterion validations for the scenario measures they developed, and they offer evidence of the role that additional methods can play in validating scenarios such as those used here.

Results from Study 3 presented lower correlations between the scenarios and surveys than the two pieces above. The scenarios' mixed correlations with teacher surveys provided evidence that the two methods may be tapping similar constructs; use of these two could provide a more complete picture of how expertise influences school leaders' actions. For example, while the scenarios may offer richer insights into how principals differ in the depth and content of their expertise, measures such as teacher surveys of principal practices would help to measure how frequently principals who are higher (or lower) in expertise engage in particular related activities in their schools. Viewed in this way, the greater value for such measures may come when we view them as providing complementary perspectives on expertise rather than assuming that their results will always correlate highly. These results would match at least one previous study's conclusions: when Leithwood and Stager (1989) compared principals'

responses on scenarios to their scores on an instrument that measured principal effectiveness, they theorized that only when results from these measures were examined together did they help to paint a more complete picture of each principal's expertise.

Future Research Suggestions and Directions

One primary objective of this dissertation is to advance the discussion regarding measurements of educational leadership expertise. Current critiques and debates about the current state of certification and professional development program for school leaders have cited the paucity of evidence for just what impacts these different programs have (Smylie and Bennett, 2006; McCarthy, 1999b; Copland, 2000). As the field considers ways to evaluate the effectiveness of certification or professional development programs it must pursue strategies to measure the practical expertise that guides school leaders' actions.

With the findings from this study I am optimistic about the use of open-ended scenarios to provide valid measures of the expertise that educational leaders use in their work. As I discussed earlier in this chapter, this study raises significant questions about the complexity of expertise as a construct and just how well the scenarios prompted individuals to demonstrate their expertise in the responses. These results indicate that scenarios such as these can capture differing levels of expertise, but additional work is needed to refine their structure and content to elicit greater evidence of respondents' expertise. A number of future initiatives would help to address the issues.

First, the results in Study 2 (which showed that subdomain correlations across the domains were equal to and sometimes higher than correlations within

the domains) underscore the need to re-visit the structures that researchers have used thus far to conceptualize educational leadership expertise. Little if any dialogue has occurred to compare these different lines of research and consider not only if some constructs are the same across domains, but also if alternative understandings of expertise help to explain these different domains' relationships to each other. For example, do principals' skills as "instructional leaders" better explain their expertise, so that we see principals with a greater understanding of "subject matter" also higher in areas such as "effective teaching and learning" and "standards-based thinking"? Do different understandings of principals' organizational roles better explain why principals who score higher in "data-based decision making" also score higher in "planning" and "monitoring instructional improvement"? This study emphasizes the need for content experts to examine these conceptual distinctions and consider if or how we might better define expertise in its complexity.

Earlier in this chapter I discussed how findings from studies 1 and 2 questioned the design of the scenarios: multiple content experts expressed concern that they did not adequately prompt respondents to discuss the various areas of expertise, and the low means and standard deviations across the selected average scores supported these concerns. An essential step to advance this research in the future will be to solicit content expert feedback in the creation of the scenarios themselves. Such a strategy would involve asking experts to critique draft scenarios before they are administered to respondents. While results from Leithwood, et al. emphasized the need to use "ill-structured" or open-ended prompts, numerous questions remain about how best to elicit demonstrations of different areas expertise. For example, can one scenario

prompt adequately for more than one subdomain of expertise? What specific contextual details should be included? Study 2 demonstrated that different scenarios generate better evidence of expertise than others, and these findings can guide a deeper examination of the scenarios by a group of content experts.

Study 3 offered initial (albeit limited) evidence that teachers' survey reports of their principals' expertise and related activities tap similar constructs of expertise, but additional work can re-examine the survey questions used as criterion measures. As discussed in the methods section, the survey questions used in this study came from a larger survey asking principals and teachers about a range of conditions in their schools, and these survey measures might be re-examined to ask respondents more closely about principals' expertise in specific areas. As a group of content experts revisits the scenario content, they might also revisit the survey items.

Even as future research might revisit the surveys as criterion measures, it can also evaluate additional measures such as logging mechanisms, observations, or subject matter tests. As summarized above, Stecher, et al. (2006), and Kersting (2008) reported higher correlations between their scenario results and criterion measures such as these. While these two studies used different scenarios to examine teachers' knowledge (and not leadership expertise), they provide evidence that methods beyond surveys may offer better criterion validation of the leadership expertise scenarios.

On a broader level, with additional refinement scenarios such as these could help to address a number of gaps in the field. As I have emphasized in the literature, the need for these measures in educational leadership research is great. First and foremost, they can help to address the field's limited evidence

regarding the effectiveness of different training and certification programs for new school leaders. Scenarios could be administered to candidates before and after their participation in a program to examine if they demonstrate greater or less expertise in scenario responses after completion. Such longitudinal evaluations could provide better evidence of programs' effects on graduates' expertise and practical knowledge than many of the traditional measures such as surveys that have asked for participants' opinions about the coursework or activities. As the field continues to debate the most effective preparation strategies for school leaders, measures such as these can provide additional evidence about what approaches work best to equip these individuals with the skills they need.

Finally, while this study has focused on principals' responses to the scenarios, these measures could certainly help examine the leadership expertise that other individuals throughout the school possess. As I discussed in the literature review, this research used principal responses in part because of the central role that they play in guiding and organizing a school around improved teaching and learning. However, numerous other individuals possess leadership expertise that is key to these goals, and we must ultimately look beyond the principal to understand how expertise is distributed across a school staff and how it shares those resources. The concepts and measures developed here can certainly be used to measure other school members' leadership expertise as well. Scenarios administered to additional members could help to identify those who possess and provide the expertise that faculty rely on to organize their schools successfully.