PITTS, ROBYN LYN THOMAS, Ph.D. Understanding Student Learning Evidence: A Case Study of Evaluation Use and Evaluation Influence for Accountability and Learning. (2017)
Directed by Dr. Jill Anne Chouinard. 235 pp.

Evaluation use is a key construct in evaluation that characterizes the ways in which an evaluation, through its processes and findings, affects people and situations. Through in-depth case study, this research explores the nature of evaluation use, and the related notion of evaluation influence, within the context of assessment in higher education. Despite a historical focus on compliance and accreditation, assessment contemporarily hinges on increasing the use of student learning evidence for decision making across many levels of an educational organization. This shift toward learning has positioned assessment as a context for evaluation theory and practice, one that offers a unique opportunity to study evaluation use and influence relative to various purposes for evaluation (i.e., accountability and learning). The findings suggested three problematics, or dilemmas, that shape the nature of evaluation use and influence in assessment: facilitating sensemaking processes, engaging systemic complexity, and attending to power and information gaps that exist within and between educational program models and their evaluative tools. Findings from this study also suggest that student learning evidence has a profound impact on educational programming, both at the individual student and program levels.

Keywords: evaluation use, evaluation influence, assessment, accountability, learning

UNDERSTANDING STUDENT LEARNING EVIDENCE: A CASE STUDY OF EVALUATION USE AND EVALUATION INFLUENCE FOR ACCOUNTABILITY AND LEARNING

by

Robyn Lyn Thomas Pitts

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements of the Degree
Doctor of Philosophy

Greensboro
2017

Approved by

Jill Anne Chouinard
Committee Chair

APPROVAL PAGE

This dissertation, written by Robyn Lyn Thomas Pitts, has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair Jill Anne Chouinard

Committee Members Ayesha Boyce

Jodi Pettazzoni

John Willse

_____

Date of Acceptance by Committee

_____

Date of Final Oral Examination

ACKNOWLEDGMENTS

It has been a privilege to undertake doctoral studies, and I am thankful for the opportunities offered to me as a member of this vibrant community. I am grateful for the generous financial support of The University of North Carolina at Greensboro, the School of Education, and the Educational Research Methodology department that enabled me to contribute my time to a thrilling and diverse collection of experiential learning projects. I am thankful to my professors, including Randy Penfield and Devdass Sunnassee, who encouraged me to study broadly and think critically about issues of measurement and evaluation, and to Micheline Chalhoub-Deville, who introduced me to the notion of consequential validity and encouraged my academic pursuit of such issues. As I reflect, I appreciate all of the academic meanderings and quirky conversations I have shared with my professors, colleagues, and peers: thank you all for enriching this season of study by sharing your wisdom, support, and humor.

In this specific effort, I am indebted to my committee members for graciously sharing their time and ways of thinking with me. Jodi Pettazzoni spent many hours conversing with me about the current moment in higher education assessment and cheerfully coaching me throughout this process. As professor and program chair, John Willse helped me to explore the intersections between assessment and evaluation, offering insights, encouragement, wit, and advice. I am incredibly thankful for Ayesha Boyce's contributions to this work and to my overall training in evaluation. As professor and mentor, she spent a lot of one-on-one time with me answering questions, providing

guidance, and engaging with me in deep consideration of my career aspirations. My thanks are especially given to Jill Anne Chouinard, my chair and advisor, thought partner, colleague, and friend. I thank her for introducing me to evaluation. The ways in which she has adeptly stretched and enriched my thinking are innumerable. I have appreciated her readiness to invite students into her work while concurrently emboldening us to develop our own interests, and I've grown as a result of her continual encouragement to think deeply, critically, and creatively about the nature of evaluation.

For their role in helping me to arrive where I am today, I am grateful to many: to Beth Cook and Pat Cleino, who nudged me toward a different path in life and urged me to see the world differently; to Jon Yoshioka and Tarin Schmidt Dalton, who encouraged me to pursue doctoral studies and catalyzed my development in my early career; to my teaching colleagues and friends, whose day-to-day work makes the difference for their students; and to my former students, especially the class of 2013, who taught me some of my most valuable lessons and by whom I am continually impressed and inspired.

I lack the words to thank my family sufficiently. My in-laws, Dave and Sally, have been incredibly engaged and supportive throughout this process. I am lucky to be a part of 'Team Thomas'—including James, Kathryn, and Ryan—who 'do life' with me across the varied seasons of our lives. My parents, Jim and Christie, have my deepest admiration for their vision for our family, their unwavering love, and their unconditional support. I share my life with Brian, my partner in all things, whom I adore and respect and with whom I joyfully pursue the adventure that is out there.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

What gets measured gets managed. —Peter Drucker

Evaluation seeks to generate contextually consequential information for such purposes as decision making and social betterment. Instrumental use of evaluation leads to direct and observable outcomes, making the influence of evaluation on a situation and its stakeholders explicit. However, through its processes and findings, evaluation influences people and situations in other ways. Understanding evaluation use, factors influencing use, and how uses of evaluative information align with the overarching purposes of accountability and learning is a longstanding concern within the field that has been researched extensively. Within the last 20 years, the notion of use has been expanded to consider some more diffuse and indirect effects of evaluation, resulting in a line of inquiry into "evaluation influence" (Alkin & Taut, 2003; Henry & Mark, 2003; Kirkhart, 2000; Mark & Henry, 2004). Across these two areas of research, effort has been made to understand how evaluation use and influence manifest within the specific contexts of practice within which evaluation occurs (e.g., education, public health, and public policy, among many others). While evaluation has an established history in education, its role in a specific area of education, the higher education practice of "learning outcomes assessment" (Rickards & Stitt-Bergh, 2016b), is less well-understood. An in-depth case study of the uses and influences of evaluative evidence in

assessment can add to the existing literature on evaluation use and influence, many studies of which have been undertaken in the educational context. Such research can also support evaluators working in the practical context of assessment by positioning the use of assessment-based evaluative information (i.e., student learning evidence) within known conceptual and theoretical frameworks for evaluation use and influence.

Background

Evaluation is a practice that is defined by its purpose. Simply stated, evaluation is "judging the merit or worth of an entity" with the goal of valuing in a systematic way (Alkin, 2011, p. 9). While there exist many connections between evaluation and research, it cannot be overlooked that evaluation, in seeking to generate credible and actionable evidence (Donaldson, Christie, & Mark, 2015), is inherently context-bound (Cousins & Shulha, 2006). Ideally, evaluation findings are generated for specific use by specific users within a specific context. Since the purpose of evaluation is expressed in its use, and since use is informed by important aspects of context, evaluation cannot be separated from those contextual factors that indelibly shape its processes and findings. Compared with researchers, evaluators face the unique challenge of attending to context within the generation, interpretation, and use of the findings it produces (Cousins & Shulha, 2006). However, while evaluation constitutes its own professional field, it is performed within many specific applied domains (hereafter referred to as contexts of evaluation practice or practical contexts; Christie & Vo, 2015). Much research on evaluation is rooted in these practical contexts (e.g., education, public health, public policy, etc.), and evaluators advance evaluation knowledge and practice by studying its nature across various contexts

of practice (Christie, 2015). To these practical contexts, professional evaluators bring a robust, multi-decade tradition of theoretical, empirical, and practical research on evaluation (Mathison, 2005), which, in turn, permits them to improve the quality of evaluation processes and findings within each practical context.

One emerging context of practice for evaluation is that of student learning outcomes assessment in higher education (hereafter referred to as assessment or assessment-based educational evaluation). Though educational evaluation has a long and rich history, the nature of its role in assessment is less well-understood. Assessment began to take hold as a widespread practice in higher education during the 1980s as a means of generating accountability evidence through compliance reporting (Banta & Palomba, 2014; Ikenberry & Kuh, 2015). Suskie (2004) argued that a revolution in higher education occurred in the mid-1990s when Barr and Tagg (1995) introduced a "learning-centered paradigm" and called for a shift in the focus of assessment from teaching to learning: a focus on teaching utilized assessment for grading students, while a focus on learning utilizes assessment for understanding what does or does not work in curriculum or instruction. Historically, schools utilized courses and credit hours as currency, a practice stemming from an earlier use of amount of "seat time" as a measure of achievement (Suskie, 2004). Education was predicated on characterizing a student's educational career in courses (short for "course of lectures," a vestigial remnant from a time when students paid for a specific number of lectures when paying tuition) and credits (which represent a certain number of hours of study; Schneider & Shoenberg, 1999). However, as Schneider and Shoenberg offer, "There is no particular reason why a

bachelor's degree should take four years of full-time study . . . nor is there any particular reason why all bachelor's degrees should take the same amount of time to complete . . ." (p. 33). While these artificial designations support the flow of students through educational programs by tracking acquisition of courses and credits to obtain degrees, this approach lacks a meaningful, substantive anchor on which to evaluate the quality of student learning or the quality of educational programming. Moving away from these somewhat legitimative structures, contemporary educational approaches (i.e., standards or objective-based teaching and learning) require evidence of student learning to substantiate claims of minimal competency based on demonstrated learning.

In the 20 years since the introduction of Barr and Tagg's (1995) learning-centered paradigm, accrediting bodies have called for an increased focus on evidence of student learning within the traditional assessment framework (Ikenberry & Kuh, 2015; Kinzie, Hutchings, & Jankowski, 2015), resulting in two main priorities for assessment practitioners: leveraging the use of assessment for learning and improvement purposes (as opposed to compliance reporting) and doing so by using direct evidence of student learning (Kuh et al., 2015b). A focus on the use of student learning evidence is new to assessment within the last decade, a stark contrast to the more traditional, indirect sources of evidence like course grades, grade point averages, and perceptual surveys (Rickards & Stitt-Bergh, 2016b). This shift represents an increasing focus on the use of direct evidence of student learning for higher order purposes within an educational organization, a vision that is more aligned in scope and effort with that of professional evaluation. Indeed, there is evidence in both professional literatures to suggest a growing

interest in connecting evaluation theory and practice with assessment (in evaluation: Rickards & Stitt-Bergh, 2016b, within a volume that evidences the work of many evaluators working in assessment; in assessment: Ikenberry & Kuh, 2015; Jonson, Guetterman, & Thompson, 2014; Kinzie et al., 2015). While the exact nature of the relationship between evaluation and assessment has not yet been explored extensively within the literatures of the two fields, two bodies of literature that have yet to intersect, it has been suggested that assessment is a context of evaluation practice that provides a worthwhile opportunity for research on the interplay between evaluation use and context (Stitt-Bergh, Rickards, & Jones, 2016).

Use is a key construct within the field of evaluation since the driving purpose of evaluation is to judge the merit or worth of an entity, a value- and context-laden designation. It has been suggested that evaluation use may be one of the most studied areas of evaluation (Christie & Vo, 2015; Fleischer & Christie, 2009) since underutilization of findings remains a sustained concern over the past 40 years and has fueled a number of research studies (see Alkin, Daillak, & White, 1979; Brandon & Singh, 2009; Christina A Christie & Vo, 2015; Cousins & Leithwood, 1986; Cousins & Shulha, 2006; Fleischer & Christie, 2009; Johnson et al., 2009; Loud & Mayne, 2014; Mayne, 2009; Patton, 1997; Preskill & Caracelli, 1997; Shulha & Cousins, 1997). Since underutilization of findings is presently a priority area for assessment (Kuh et al., 2015b), applying the longstanding and extensive research base on evaluation use to the specific context of assessment can lead to new understandings of evaluation use relative to two distinct purposes (i.e., accountability and learning) while simultaneously supporting the

day-to-day efforts of evaluators working in assessment. The research base on evaluation use includes case studies (many of which were undertaken in educational contexts; Brandon & Singh, 2009), systematic literature reviews (see Cousins & Leithwood, 1986; Johnson et al., 2009; Shulha & Cousins, 1997), and surveys of professional evaluators (Fleischer & Christie, 2009; Preskill & Caracelli, 1997). Evaluation use has also been researched as it relates to "research knowledge use" (i.e., the use of findings from research studies) suggesting that, while evaluation use and research knowledge use are related, the constructs are fundamentally unique since the former is inherently context-bound (Cousins & Shulha, 2006). Case study of evaluation use in the practical context of assessment thus poses a new inquiry that can add to a long tradition of research on use. This contextualized study can address contemporary issues facing evaluative efforts in the assessment context, specifically the call to increase assessment use by leveraging student learning evidence.

Interestingly for evaluators, assessment utilizes evaluation within program contexts that have an explicit learning focus (Rickards & Stitt-Bergh, 2016b). Unlike most evaluation contexts, the primary methods of gathering direct evidence in assessment (i.e., the tests, projects, portfolios, presentations, performances, and other assessment methods by which student learning is measured) have their own unique utility: while student learning evidence may be used for the assessment purposes of program improvement and organizational learning, its primary role within educational organizations is to evaluate individual student performance for formative and summative purposes (e.g., identifying individualized areas for student improvement, determining

whether students have attained sufficient mastery of learning goals). Thus, from a bottom-up perspective, student learning evidence can be used simultaneously for two distinct purposes: as evidence of learning mastery and as evidence for accountability. Furthermore, these purposes occur at two distinct levels: the student level and the program level. Issues surrounding evaluation uses, types, and purposes suggest a variety of possible questions for study: What current and future roles exist for evaluators working in assessment contexts? How is evaluation used, and what factors specific to assessment affect those uses? How might knowledge of evaluation use support and extend evaluators' capacity for developing quality assessment-based educational evaluation? How can evaluators maximize and leverage student learning evidence?

Out of the multi-decade evaluation use literature has emerged an expanded notion of use. These new notions emerged from the constraint felt by many evaluators in response to typologies of use that predominated the evaluation field at the turn of the century, ones that largely consisted of result-oriented conceptions of "findings use" (Shulha & Cousins, 1997). Importantly, Patton (1997) suggested that the process of an evaluation itself (i.e., "process use") affects people and situations in a way that is distinct from findings use. To bring attention to the diffuse and unexpected ways in which evaluation affects its local context and program stakeholders beyond notions of findings or process use, Kirkhart (2000) argued for consideration of the influence of an evaluation by analyzing additional dimensions of use (e.g., intention and time). Her dimensional framework of "evaluation influence" prompted the generation of additional theoretical frameworks for influence (Alkin & Taut, 2003; Henry & Mark, 2003; Mark & Henry,

2004) which in turn prompted empirical study of evaluation influence in specific contexts of evaluation practice (see Herbert, 2014). One such empirical study (Gildemyn, 2014) used case study to better understand the influence of a monitoring and evaluation (M&E) approach on a Ghanaian governmental health program, suggesting that a similar study within the practical context of assessment will be valuable not only to evaluators working as embedded evaluators in assessment contexts, but also to evaluation researchers and academic faculty responsible for undertaking assessment. Such study has similarly been suggested by Jonson et al. (2014), who drew on Kirkhart's (2000) framework to develop a conceptual framework of evaluation influence for assessment in higher education. Bridging notions of evaluation influence to evaluative practices in assessment suggests many interesting questions for evaluators and educators alike: Given the unique learning-centered and evidence-generating context within which assessment occurs, how might analysis of evaluation influence contribute to understandings of evaluation use? In what ways does evaluation influence manifest in this context of evaluation practice, specifically regarding student learning evidence? How might analysis of the influence of student learning evidence in assessment supplement the existing theoretical, empirical, and practical research on evaluation use and influence?

Thus, to position assessment as a context of evaluation practice and to expand upon current considerations of evaluation use and influence, an in-depth, contextualized case study of the influences of student learning evidence within real-world assessment work is needed. As a context of evaluation practice, assessment provides a worthwhile opportunity for researching evaluation use and influence in at least five specific ways.

First, assessment facilitates evaluation iteratively within a specific context, permitting research on the relationship between evaluation and context more deeply and longitudinally (Rickards & Stitt-Bergh, 2016b). Second, accreditation requires documentation of assessment efforts for compliance reporting (Kuh et al., 2015b), providing an interesting and unique opportunity for evaluators to study the reported uses and influences of evaluation. Third, since documentation of assessment is required, analysis of intended and actual use may contribute to discussions within both fields surrounding intention and consequences. Fourth, assessment occurs in educational organizations that are focused on tiers of learning (i.e., at the student, program, and organizational levels) permitting the study of use across levels of an educational organization as well as across stakeholder groups embedded in these levels. Finally, like evaluation, assessment is facing the challenge of negotiating accountability and learning purposes, meaning that research conducted in assessment may be useful to inform this tension more broadly within the field of evaluation.

Purpose

This study seeks to understand the uses and influences of evaluation within the practical context of assessment. Given the complexity of educational evaluation, this study will focus on the uses and influences occurring from a single source of a specific type of evaluative evidence (i.e., student learning evidence generated by a single testing system in a clinical science training program within a medical school). The study will address the use and influence of evaluative information relative to the two key particularities of evaluation as it manifests in the context of assessment (i.e., an explicit

focus on learning (Rickards & Stitt-Bergh, 2016b) and an increasing focus on deriving evidence from course-embedded assessments (Kuh et al., 2015b). The use of student learning evidence for these purposes requires collaboration between faculty and evaluators to design course-embedded assessments that produce evidence in alignment with program and organizational goals. Through a case study of one such intentionally designed testing system, evaluation use and influence can be positioned to demonstrate the value of aligning accountability and learning purposes at the program level with student learning mastery purposes at more micro-levels (e.g., student or program levels). It is hoped that such research will contribute to the existing literature regarding evaluation use and influence while also providing guidance for evaluators working in the practical context of assessment. The findings might also be meaningful for educators who are responsible for conducting evaluation and/or using evaluative information.

Research Questions

The following research questions have been developed to guide this inquiry: Research Question 1. Based on a specific case of an innovative clinical skills examination system (CSE system) in a medical school, what is the nature of evaluation use and influence in assessment-based educational evaluation?

1.1 What types of use (instrumental, conceptual, symbolic, process, etc.) are made of student learning evidence within the educational organization?

1.2 To what extent are uses oriented toward accountability and learning?

1.3 How are these findings affected by expanding the notion of evaluation use to consider dimensions of evaluation influence?

The overarching research question guiding this inquiry considers evaluation use and influence in assessment through an interpretive analysis of use types (1.1), the purposes for use (1.2), and how dimensions of evaluation influence affect these findings about use (1.3). These questions suggest a qualitative case study approach, one that draws on a rich tradition of exploratory study of evaluation use and influence as they manifest in practical contexts (Christie & Vo, 2015; see Alkin et al., 1979; Gildemyn, 2014). Findings from this study position the use of student learning evidence within the traditional framework of instrumental, conceptual, or symbolic forms of findings and process use (Alkin & King, 2016), existing frameworks of factors influencing use (Cousins & Leithwood, 1986; Johnson et al., 2009), and a contemporary challenge in evaluation regarding how an evaluator can address the tensions that exist between accountability and learning purposes for evaluation.

Relevance

The driving goal of this work is to develop a greater understanding of evaluation use and influence within the practical context of assessment. Findings from an in-depth case study of these concepts may provide further support for pre-existing conceptual frameworks surrounding use and influence; alternatively, the study may suggest a modified conceptual framework for understanding use and influence the specific context of assessment. This study provides an opportunity to "examine the ways in which information—evaluative and otherwise—is packaged, diffused, understood, and utilized to serve decision-making purposes" (Christie & Vo, 2015, p.xiii). In a report on the use of evidence in public policy, it is suggested that a gap exists in understanding the systems

and structures within educational organizations that facilitate research use and evaluation use (Prewitt, Schwandt, Straf, National Research Council (U.S.), & Committee on the Use of Social Science Knowledge in Public Policy, 2012). Similarly, there have been calls for more micro-level research on the use of data within the everyday practices of educational organizations (Spillane, Reiser, & Reimer, 2002). Given the primacy of the use construct within the field of evaluation, there is a continued interest in research on evaluation use and influence, specifically regarding the use of evaluation for decision making within specific contexts of evaluation practice (Christie, 2015). Studying use and influence in the practical context of assessment offers evaluators the chance to study evaluation use more deeply and longitudinally (Rickards & Stitt-Bergh, 2016b), a somewhat unique opportunity among contexts of evaluation practice since assessment includes an explicit goal of "closing the loop" (Ikenberry & Kuh, 2015) by documenting use and its associated follow-up responses.

Additionally, this study responds to the call for an expanded consideration of use as influence. Since evaluation influence frameworks are relatively new within the literature, few empirical studies have been conducted to explore these frameworks in the various contexts of evaluation practice (Herbert, 2014). Much like one such study, Gildemyn's (2014) case study of evaluation influence in a multi-site M&E approach, this case study will result in increased awareness of evaluation influence within a practical context in which evaluation is used. This work provides guidance for embedded/internal evaluators working in assessment contexts as they seek to support educators through processes of "assisted sensemaking" (Julnes, 2012) across various levels of an

educational organization (Porteous & Montague, 2014) in order to facilitate negotiation of meaning and values when making sense of student learning evidence. This is particularly relevant given that the use of student learning evidence for assessment-based improvement across levels of the educational organization is currently receiving a great deal of attention from practitioners (Ikenberry & Kuh, 2015; Jonson et al., 2014; Kinzie et al., 2015).

Finally, findings from this in-depth case study will illustrate conceptual connections between the theories and practices of evaluation and the practices and nuances of the context of assessment, relationships that have been suggested in both evaluation and assessment literatures (Ikenberry & Kuh, 2015; Jonson et al., 2014; Kinzie et al., 2015; Rickards & Stitt-Bergh, 2016b). This study responds to calls for research on the use of evaluation in assessment, such as that of Jonson and colleagues (2014), who argue for the utility of evaluation influence for addressing the current assessment climate. These authors suggest that qualitative study can be utilized to characterize how practitioners perceive the assessment process and findings (Jonson, Thompson, Guetterman, & Mitchell, 2017), key considerations of evaluation use and influence.

Key Constructs

Multidisciplinary work oftentimes necessitates clarification of terms that may be used differentially across various theories, literatures, and contexts of practice. In some cases, this is necessary because fields use language with specialized meaning. Other times, fields that share many of the same terms may use them to describe slightly different notions. “Assessment” is one such term: evaluators consider assessment to be an

evaluation of individuals (e.g., state assessment tests or national assessment programs; Alkin, 2011, p. 10) while, in higher education, assessment refers to tools measuring student learning (e.g., tests, projects, portfolios, presentations, performances, etc.) as well as to the process of evaluating academic programs for improvement. To clarify the meaning of such terms for the purposes of this study, the remainder of this section serves to define some key terms in evaluation and assessment.

Evaluation. Evaluation is an effort that is defined by its purpose. Simply stated, evaluation is "judging the merit or worth of an entity" with the goal of valuing in a systematic way (Alkin, 2011, p. 9). Evaluation constitutes its own professional domain of social science inquiry. It is distinct from research in the sense that its findings are crafted for a specific use by a specific set of users within a specific context (Cousins & Shulha, 2006). This highlights an important characteristic of evaluation: it is inextricably context-bound and must be designed and implemented anew in each specific set of circumstances. This need for customization has propagated many approaches to evaluation (e.g., utilization-focused, theory-driven, culturally responsive, etc.) that helpfully guide evaluators as they perform evaluations within the unique contexts and circumstances of a given evaluation and of the many practical contexts (e.g., education, health care, etc.) in which evaluation is used.

Student learning evidence. In education, assessment is a term that commonly refers to two distinct concepts: a measurement tool (product) is an assessment and a mechanism for program-level improvement (process) is called assessment. As Alkin (2011) frames, evaluators consider these uses of the term "assessment" as distinctive

concepts: while assessments and evaluation are both evaluative processes (i.e., they both judge merit or worth), evaluation refers to the evaluation of programs while assessment refers to the evaluation of the clients of a program. From this perspective, evaluation of educational programs is insulated from assessment of students. By contrast, this study focuses on the use of assessment-generated evidence in evaluation. Evidence generated by assessments (hereafter referred to as student learning evidence) evaluates *what students know and can do* as a result of participation in an educational program or course of study.

This type of evidence is generated through several mechanisms (e.g., tests, projects, portfolios, presentations, performances, etc.). If evaluators collaborate with educators when designing assessment tools, student learning evidence generated from these sources may be used at higher order levels of the educational organization (e.g., course, cohort, track, department, policy, or organization levels) as well as by program stakeholders (i.e., program beneficiaries (students, families, etc.), internal program stakeholders (i.e., faculty, staff, administrators, and other educators facilitating the program), and/or external program stakeholders (i.e., educational administrators, funders, accrediting agencies, etc.). Using this evidence across various levels of the educational organization can serve to anchor and animate improvement using a currency of learning. Within this study, specific mechanisms of gathering student learning evidence will be referred to directly (e.g., tests, projects, etc.); and the term “assessment” will be used to describe the specific context of educational evaluation in which program-level accountability and learning purposes are addressed.

Assessment. While testing tools provide discrete measures of 'what students know and can do,' the process of assessment leverages information from both measurement and non-measurement sources to make value judgments about the quality of teaching and learning (Miller & Linn, 2013). From an evaluative perspective, *assessment-as-process* can be considered "utilization-focused outcome evaluation with goals that include program/organization improvement and an integration of evaluative thinking in the program and organization" (Rickards & Stitt-Bergh, 2016a, p. 7). In this sense, assessment is a practical context in which evaluation occurs. Unlike other forms of educational evaluation, assessment has specifically emerged over the course of the last thirty years to address accountability and learning at the program level. In an authoritative text on the essentials of assessment published in 1999, two leaders in the field, Catherine Palomba and Trudy Banta, defined assessment as "the systematic collection, review, and use of information about educational programs undertaken for the purpose of improving student learning and development" (p. 4). Reflecting the ongoing evolution of assessment as a niche within educational evaluation, Banta and Palomba (2014) offered a more holistic definition of assessment in the second edition of their text: "Assessment is the process of providing credible evidence of resources, implementation actions, and outcomes undertaken for the purpose of improving the effectiveness of instruction, programs, and services" (p. 2). The newer definition is more comprehensive, acknowledging that assessment findings will apply to student learning, academic programs, students support services, and administrative services.

In the updated conceptualization of assessment, Banta and Palomba emphasize that the primary goal of assessment is "to help faculty and staff improve instruction programs, and services, and thus student learning, continuously" (p. 3). This goal of learning and improvement aligns well with the overarching purpose of assessment, which is "to understand how educational programs are working and to determine whether they are contributing to student growth and development" (pp. 9–10). Since assessment is a relatively new field that has not yet consolidated its terminology (Suskie, 2004), it may also be referred to as "student learning outcomes assessment" and "institutional effectiveness" (Banta & Palomba, 2014). Assessment is distinguished from the related processes of grading (which is more holistic than assessment), testing (which is but one process for gathering assessment evidence), research (which seeks to generate knowledge while assessment seeks to inform practice), and institutional effectiveness (of which assessment is but one component; Suskie, 2004). In this sense, learning outcomes information is a direct measure of student ability that should be understood and analyzed separately from grading, testing, research, and/or institutional effectiveness uses of the same information.

Like evaluators, assessment professionals also draw a distinction between assessment and evaluation, positioning evaluation as being broader than assessment since the latter focuses explicitly on student learning goals; however, evaluation is also a constituent element of the assessment process since stakeholders use professional judgment to make sense of assessment outcomes (Suskie, 2004). Drawing distinctions between evaluation and assessment may be less important as assessment pursues the use

of direct student learning evidence since such a direction mitigates its distinction from evaluation. As assessment increasingly utilizes learning outcomes (i.e., direct indicators of student learning goals) to facilitate decision making across all levels of an educational organization (Ikenberry & Kuh, 2015; Kinzie et al., 2015), assessment will begin to resemble professional evaluation more closely. Indeed, in a recent volume of *New Directions for Evaluation* (Rickards & Stitt-Bergh, 2016a) evaluators working in assessment contexts directly align assessment with evaluation, suggesting the field of evaluation has much to offer assessment efforts. Thus, for the sake of clarity and the purposes of this study, assessment will be considered a specific practical context in which evaluation is undertaken, sometimes described as an 'applied domain' within the evaluation literature and hereafter referred to as assessment-based educational evaluation. Additionally, professionals working in assessment will be referred to as evaluators, though it should be recognized that not all professionals undertaking assessment will have received formal training in evaluation.

Educational organization. There are many terms to describe the sites within which assessment takes place. Such sites may be considered schools, colleges, universities, or institutions, as well as courses, programs, departments, or organizations. For the purposes of this study, the term "educational organization" will be used to describe a single site within which assessment occurs. While there exists some distinction in the literature between program-level learning and organizational learning (e.g., Porteous & Montague, 2014), such a distinction will not be made in this study. Within in

this study, the use of student learning evidence will be discussed at the student and program levels.

Evaluation use. A key concept in evaluation theory and practice, evaluation use is a term that comprises the ways in which the process, products, or findings from evaluation are applied to produce an effect (Johnson et al., 2009). A key construct in professional evaluation for the last 40 years, evaluation use has been discussed in research literature by describing factors that facilitate or hinder use. Evaluation use research has also produced various typologies to describe use, both in terms of findings (i.e., resulted-based conceptions of instrumental use, conceptual use, and legitimative use) and processes (i.e., process-based conceptions of instrumental process use, conceptual process use, and symbolic process use; Alkin & King, 2016; Alkin & Taut, 2003). These concepts are elaborated, and other types of use are described, in the section on evaluation use in Chapter II.

Evaluation influence. An expansion of the concept of use, evaluation influence describes "the capacity or power of persons or things to produce effects on others by tangible or indirect means" (Kirkhart, 2000, p. 7) to address "multidirectional, incremental, unintentional, and instrumental" (p. 7) effects. Multiple frameworks for evaluation influence have been generated and are described further in the section on evaluation influence in Chapter II. Kirkhart's (2000) initial framework focuses on three dimensions of influence: source, intention, time. Alkin and Taut (2003) modified Kirkhart's intention dimension (e.g., intended, unintended) by adding awareness (e.g., intended/aware, unintended/aware, unintended/unaware) to distinguish use and influence

conceptually. Two additional frameworks have been developed to characterize influence processes and outcomes (Henry & Mark, 2003; Mark & Henry, 2004).

Program. As described by McDavid, Huse, and Hawthorn (2013), programs "are means-ends chains that are intended to achieve some agreed-on objective(s)" (p. 10). Despite the conception of programs as manifestations of policy, it is important to recognize that those responsible for implementing a program must translate policy mandates into action. Within this study, the relationship of policy to the program (and thus to the testing system that animates it) will be explored, permitting an analysis of the ways in which student learning evidence generated by the testing system aligns with the purposes for educational improvement. For the purposes of this study, a program will be comprised of multiple discrete courses that are connected by a set of learning goals, or program outcomes, that students are expected to achieve as a result of undertaking and completing the program. Within the program in which this case study is conducted, the term "program model" will be used to describe the inputs, activities, outputs, and outcomes of a program that might be depicted as a logic model or program theory, and the term "evaluative tool" will be used to indicate all of the evaluation-related materials produced by and/or used within the program model.

Manuscript Organization

This manuscript consists of six chapters. Chapter I offered an overview of the study and its rationale. Chapter II consists of two sections that serve as a primer on the contemporary notions of evaluation use and influence (section 1) and positions inquiry into evaluation use and influence within the domain of assessment (section 2). In Chapter

III, methodology is presented, followed by a thematic analysis of findings in Chapter IV. An in-depth discussion of findings and conclusions is offered in Chapter V with implications, limitations, and future directions presented in Chapter VI.

CHAPTER II

LITERATURE REVIEW

This dissertation study focuses on the uses and influences of evaluative information in assessment-based educational evaluation. The literature supporting this analysis is primarily positioned within evaluation, a field with a robust understanding of use and influence across many decades of research. Since this study seeks to understand student learning evidence, special attention must be paid to the context in which this study takes place. Assessment is a unique context for educational evaluation with reference to its explicit focus on learning, its obligation to document interventions using evaluative information, and its current focus on increasing assessment use (evaluation use) by leveraging student learning evidence (evaluative information). Accordingly, this chapter provides an overview of research on evaluation use and influence (section 1) before positioning these concepts in educational evaluation (section 2).

Positioning Use and Influence in Evaluation

A focus on improving the use and quality of evidence is not a new concern in the social sciences; rather, it constitutes a consistent area of research across many fields in which scientific evidence is used for social science purposes. Nutley, Walter, and Davies published a 2007 text titled *Using Evidence: How Research Can Inform Public Services* that outlined the current state of research utilization in practical contexts and that highlighted the need for future research in this area. Five years later, in 2012, a

multidisciplinary committee commissioned by the National Research Council published a comprehensive report on how to improve the quality and use of social science research as evidence for policymaking (Prewitt et al., 2012). Paralleling these comprehensive studies of 'research use/utilization' and 'knowledge use/utilization' is the notion of 'evaluation use/utilization.' In deference to its unambiguously context-bound nature, evaluation use has been positioned as a related, yet distinctive, construct from knowledge or research use, and the field of evaluation has conducted extensive study on the use of evaluative information (i.e., information that has been generated for specific uses by specific users within a specific context; Cousins & Shulha, 2006). Just as findings produced by evaluation are distinct from findings produced by research, use is distinct from utilization. As Carol Weiss argued, the term "evaluation utilization" connotes a methodological orientation while the term "evaluation use" suggests broader consideration of the ways in which evaluative information is used (Kirkhart, 2000). The rich theoretical, empirical, and practical evaluation literature includes a multi-decade inquiry into evaluation use (e.g., Alkin et al., 1979; Cousins & Leithwood, 1986; Fleischer & Christie, 2009; Johnson et al., 2009; Preskill & Caracelli, 1997; Shulha & Cousins, 1997). More recently, the notion of 'use' has been expanded to that of 'influence' in order to emphasize the more indirect and diffuse ways that evaluation processes and findings affect people and situations (Alkin & Taut, 2003; Henry & Mark, 2003; Kirkhart, 2000; Mark & Henry, 2004).

Attending to use in evaluation. Evaluation is pervasive, both in professional practices and in ordinary, everyday endeavors. Given its expansive nature and the lack of

a consolidated, overarching theory of evaluation (Mathison, 2005), evaluation can be undertaken through one or more of the many approaches that have been well articulated and debated in the literature (Alkin, 2013; Mertens & Wilson, 2012). Across these theories exist a handful of concepts with which research on evaluation remains attentive, including such issues as: approaches to evaluation and their primary orientations, the teaching of evaluation (including capacity building, competencies, and value orientations of both evaluators and stakeholders), the role of the evaluator, the roles of various stakeholder groups (i.e., as participants and/or collaborators), the process and findings use of evaluation within the learning organization, contextual complexity (including multiplicities of purpose and issues of power), the influence and treatment of culture, the translation of evaluation theory to practice, and continued research on the nature of evaluation. Among these enduring issues are two concepts that, having been well researched within the evaluation literature, may be particularly useful for situated study within the context of assessment in higher education: evaluation use and evaluation influence. Evaluation use is a concept that comprises the ways in which the process, products, or findings from evaluation are applied to produce an effect (Johnson et al., 2009) while evaluation influence expands this conception by considering the dimensions, processes, and outcomes of use. In an effort to overcome the construct underrepresentation that is consequential to a results-oriented focus on evaluation use, Kirkhart (2000) offers three dimensions of evaluation influence: source (findings and process), intention (intended and unintended consequences), and time (immediate, end-of-cycle, and long-term). By exploring these concepts within a specific context of

evaluation practice, this study can add to the existing literature on evaluation use and influence while also supporting current and future evaluators working within this context. Findings from an in-depth case study of these concepts may provide further support for pre-existing conceptual frameworks surrounding use and influence or may suggest the need for new ways of perceiving the use and influence of student learning evidence.

Defining evaluation use. Evaluation is an effort that is defined by its purpose. Simply stated, evaluation is "judging the merit or worth of an entity" with the goal of valuing in a systematic way (Alkin, 2011, p. 9). While there is little agreement in the field regarding the overarching purpose of evaluation, evaluators agree that how an evaluation is designed, implemented, and used has a substantive effect on the findings it produces (e.g., Donaldson, Christie, & Mark, 2009; Donaldson et al., 2015). As previously stated, evaluation use is a concept that comprises the ways in which the process, products, or findings from evaluation are applied to produce an effect (Johnson et al., 2009). Use is a key construct within the field of evaluation since the driving purpose of evaluation is to judge the merit or worth of an entity, a value- and context-laden designation. It has been suggested that evaluation use may be one of the most studied areas of evaluation (Christie & Vo, 2015; Fleischer & Christie, 2009) since underutilization of findings remains a sustained concern over the past 40 years and has fueled a number of research studies (see Alkin et al., 1979; Brandon & Singh, 2009; Christina A Christie & Vo, 2015; Cousins & Leithwood, 1986; Cousins & Shulha, 2006; Fleischer & Christie, 2009; Johnson et al., 2009; Loud & Mayne, 2014; Mayne, 2009; Patton, 1997; Preskill & Caracelli, 1997; Shulha & Cousins, 1997). The remainder of the section presents a brief description of the

theoretical, empirical, and practical research on evaluation use. Key studies on use are presented, including a description of factors known to influence use, and are followed by a discussion of use types, limitations of a focus on use, and calls for an expanded notion of use.

The history of research on evaluation use. Research on evaluation use emerged from a desire to better understand how evaluations were being used and what factors supported or hindered use. Alkin and colleagues (1979) used longitudinal, multi-site educational case study to characterize the influences of evaluation on decision making and program operation. This seminal study produced an analytic framework of eight properties influencing evaluation use: evaluation boundaries, user orientations, evaluation approach, evaluator credibility, organizational factors, extraorganizational factors, information content and reporting, and administrator style. As Patton (2012) summarizes, Alkin's longitudinal research in this area resulted in his generating four categories of factors associated with evaluation use: evaluator characteristics (i.e., approach to setting priorities, involving stakeholders, and addressing credibility), user characteristics (i.e., user interests in evaluation, willingness to be involved, and position of influence), contextual characteristics (i.e., size of an organization, political climate, and competing information), and evaluation characteristics (i.e., timing of the evaluation report, relevance of its information, methods utilized, and quality of the data produced).

Cousins and Leithwood (1986) conducted a systematic review of 15 years' worth of empirical literature across the fields of education, mental health, and social services. Their work resulted in a conceptual framework of 12 factors influencing use, including

six pertaining to evaluation implementation (i.e., evaluation quality, credibility, relevance, communication quality, findings, and timeliness) and six pertaining to the decision or policy setting (i.e., information needs, decision characteristics, political climate, competing information, personal characteristics, and commitment and/or receptiveness to evaluation). This study was updated by Johnson and colleagues (2009), bringing the preceding work to date through 2005. Within evaluation implementation category of Cousins and Leithwood's framework, Johnson and colleagues added the factor of evaluator competence. They also added a category to the overall framework to reflect emergence of stakeholder-engaged evaluation approaches (e.g., utilization focused, participatory, collaborative, empowerment, democratic, and culturally responsive) in the late twentieth and early twenty-first centuries (Shulha & Cousins, 1997). The new category added by Johnson and colleagues consists of nine factors influencing use related to stakeholder involvement (i.e., their involvement with credibility, relevance, communication quality, findings, information needs, decision characteristics, personal characteristics, commitment or receptiveness to evaluation, and direct involvement. Overall, their findings suggest that use is influenced by interpersonal engagement, interaction, and communication between evaluators and stakeholders, factors addressing the intersection of stakeholders' unique thoughts and contributions with other factors known to affect evaluation use.

Two additional key developments in the evaluation use literature were the awareness, first, that the experience of an evaluation affects people and situations and, second, that context has a substantial effect on use, more so than method selection or

evaluator identity. The realization that the process and experience of evaluation influences those who engage in it. Patton (1997) called this “process use” in effort to describe those effects that emerge not from the use of evaluation results, but rather from the process of engaging in the act of evaluation itself. In this seminal work on process use, Patton identifies four consequences of process use: enhanced communication, data collection as intervention, stakeholder engagement in the substantive elements of the evaluation, and organizational development (Shulha & Cousins, 1997). While it may seem that use can be fostered based on evaluation design or evaluator qualities, Contandriopoulos and Brouselle (2012) found that the context in which evaluation is undertaken can “explain in large part the level and nature of results use” (p. 71). They offer a framework of relationship between evaluation choice and types of fit, suggesting that appropriate model-context fit is necessary to promote sue of results.

From these efforts to study use, evaluators began to realize the complex nature of evaluation use, one that mirrors the complexity of the programs and contexts in which evaluation occurs (Shulha & Cousins, 1997). As Shulha and Cousins describe, trends emerging from research on evaluation use in the late twentieth century prompted scholars to begin to approach use more holistically by pushing the use dialogue beyond the selection of appropriate approaches and methods in accordance with evaluation questions. These new research efforts began to address the centrality of context in considerations of use, the identification of uses resulting from the experience of engaging in evaluation activities, the consideration of use at the individual and organizational levels, and the facilitator, planner, and trainer roles of evaluators seeking to promote use

(Shulha & Cousins, 1997). One salient product of this transition was the development of a robust area of evaluation theory, research, and practice known as evaluation capacity building (ECB; e.g., Preskill & Russ-Eft, 2016). By focusing on ECB, evaluators could harvest organizational coherence in the design and use of evaluation within programs (Shulha & Cousins, 1997), producing a "culture of evaluation" (Loud & Mayne, 2014) within the program or organization.

In addition to these systematic reviews, two surveys of professional evaluators were used to characterize evaluators' conceptions of use. The first was conducted in 1996 (Preskill & Caracelli, 1997) just as the field began to explore broader notions of use and as evaluative work began to gravitate toward ECB and evaluation culture building. Survey results indicated the primary purposes for evaluation as facilitating organizational learning, providing information for decision making, improving programs, and determining merit or worth. Suggesting a growing importance for participation and organizational learning for evaluation practice, results also identified some key strategies for facilitating use: planning for use from the outset of an evaluation, prioritizing intended use by intended users, honoring resource limitations, involving stakeholders, and communicating often as part of an established plan. Ten years later, in 2006, this effort was repeated (Fleischer & Christie, 2009), with a key finding suggesting the growing prevalence of organizational learning in discussions of evaluation use.

Criticism of research on evaluation use. In the empirical literature on evaluation use, educational case studies, systematic literature reviews, and surveys of evaluation practitioners have been used to characterize use. The value of this corpus of research

literature is that it provides conceptual guidance for evaluators (and practitioners in field using evaluation) regarding steps that may be taken to enhance evaluation use. Despite extensive study, research on use has failed to generate convergence, which may be unsurprising given the complex nature of evaluation and the many contextually unique factors influencing its use. Indeed, findings across studies have generated multiple sets of factors impacting use, and, to date, little effort has been made toward consolidating these findings into a working theory of evaluation use, nor has extensive empirical research been conducted to test the nature of these influences or the strength of their relationships and influences on use (Henry & Mark, 2003). While this lack of convergence is perhaps to be expected given the inherently context-bound nature of evaluation, and given the centrality of use to evaluation theory and practice, continued research in this area is warranted.

Concerns about the quality of methods used to study evaluation use were highlighted in a 2009 study by Brandon and Singh that analyzed the strength of the methodological warrants for five predominant literature reviews on evaluation use, including two of the three systematic reviews presented in this literature review: Cousins & Leithwood, 1986; Shulha & Cousins, 1997. Brandon and Singh found that the studies that served as primary sources for the literature reviews were undertaken using the methods of narrative reflection, case study, survey, and simulation, thus providing insufficient scientific evidence to address the relationship between identified factors and the construct of evaluation use. While noting these methodological deficiencies in the sourced empirical research, Brandon and Singh argued that such deficiencies may be less

important when notions of evaluation use are placed within a broader context of research on evaluation. They offer that the linkages built between evaluation use (evaluation) and knowledge utilization (research) in works by Hofstetter and Alkin (2003) and Cousins and Shulha (2006) contribute greatly to the evidentiary quality of the present body of research on evaluation use. Both of these works emphasize that a key difference between knowledge use and evaluation use occurs as a result of the nature of evaluation as being inherently context-bound. Unlike traditional research, evaluation findings are produced for a specific use and specific set of users within a specific context, positioning evaluation use as unique enough from knowledge use as to justify its distinction as a unique construct (Cousins & Shulha, 2006).

Typologies of evaluation use. One of the primary outputs of the research on evaluation use is the creation of use typologies (i.e., frameworks classifying varieties of use distinguished within the literature). The four most predominant types of use are instrumental, conceptual, symbolic, and process. Instrumental use describes situations in which findings from an evaluation are used directly, typically either for decision making and problem solving (Shulha & Cousins, 1997) or program improvement (Loud & Mayne, 2014; Nutley et al., 2007). Instrumental use is direct and observable (Christie, 2015), results from a rational process, and constitutes the type of use most anticipated by evaluators and stakeholders (Mayne, 2014). By contrast, when evaluation use is less direct, less predictable, and more diffuse, it is referred to as conceptual use (Christie, 2015). Conceptual use occurs when evaluation findings influence how people think about an issue in a more general way (Alkin & Taut, 2003; Rossi, Lipsey, & Freeman, 2004). In

this sense, conceptual use can be viewed as serving an educative function (Shulha & Cousins, 1997). The third main type of use is symbolic use, though Alkin argues this is not "real use" (2011, p. 207). Symbolic use describes situations in which evaluations are conducted to fulfill a requirement (Mayne, 2014) and can thus be considered "token use" (Patton, 2008, p. 112). Shulha and Cousins (1997) position symbolic use as serving a political function, warning evaluators against circumstances in which evaluation findings may be used for "calculated actions" (Cousins & Shulha, 2006, p. 268). In addition to these three findings-oriented types of use, the final type of use, process use, describes the ways in which the experience of an evaluation affects people and situations (Patton, 1997). The addition of process use to the main types of evaluation use represented a paradigmatic shift regarding use, one that resulted from the growing integration of stakeholders into the process of an evaluation argued for by utilization-focused and participatory approaches to evaluation (Shulha & Cousins, 1997).

In response to a call for expanded notions of use in the late 1990s, evaluation researchers explored new ways of describing use (Cousins & Shulha, 2006). The evaluation use literature is now replete with additional conceptions of use, many of which are permutations of instrumental, conceptual, and symbolic use types. Conceptual use, for example, may be considered a facet of 'enlightenment use' which, alongside 'reflective use' (retrospective evaluation for future efforts), can be used in both short-term and long-term efforts (Mayne, 2014). Symbolic use of findings may be considered 'legitimative' when used to rationalize earlier decisions (Alkin & King, 2016) or 'persuasive' when used to legitimize or criticize a current or future intervention (Mayne, 2014; Rossi et al.,

2004). In a similar sense, Weiss suggested the term ‘imposed use’ to describe evaluation findings that result from situations in which higher levels of government require evidence as a requirement for lower level agencies to receive funding (Weiss, Murphy-Graham, & Birkeland, 2005). Importantly, the “shadow side of use” (Patton, 2015a), misuse, has received ongoing attention in the literature (Alkin et al., 1990; Cousins & Shulha, 2006; Patton, 2005).

Evaluation influence. Though many systematic research studies have been conducted on evaluation use (e.g., Alkin et al., 1979; Cousins & Leithwood, 1986; Fleischer & Christie, 2009; Johnson et al., 2009), an understanding of how to foster and sustain use remains elusive (Christie & Vo, 2015). While this may suggest a need for further research on evaluation use and its relationship to decision-making, it may also indicate a need to reframe the dialogue on use. To this end, evaluation scholars have offered additional conceptual frameworks and initial empirical analyses of an expanded notion of evaluation use as one of evaluation influence. The remainder of this section outlines three frameworks and findings from a recent literature review of the literature surrounding evaluation influence.

At the turn of the century, Kirkhart (2000) called for an expansion of the notion of ‘use’ by suggesting that the term itself artificially constrains consideration of how evaluation affects people and situations. She argues that ‘use’ is an awkward term due to its inability to describe effects of evaluation that are not based on findings, are unintended, or unfurl over time. To expand representations of the use construct, Kirkhart offers the concept of ‘influence’ as “the capacity or power of persons or things to produce

effects on others by tangible or indirect means" (p. 7) so as to address effects that are "multidirectional, incremental, unintentional, and instrumental" (p. 7). Kirkhart then offers a framework for influence composed of three dimensions: source (whether the effects are based on evaluation findings or evaluation processes), intention (whether the effects are intended or unintended), and time (whether the effects are immediate, end-of-cycle, or long-term). Using this "integrated theory of influence" (p. 5), evaluators are able to reconsider debates on evaluation use, map influence within a specific evaluation, track patterns of influence over time, distinguish between use and misuse, support theory building, and further study influence and evaluation theories.

In response to Kirkhart's (2000) call for evaluators to consider the effects of evaluation more holistically, Alkin and Taut (2003) suggested a framework to clearly delineate use and influence. Taking Kirkhart's model as a starting point, they reframed the second dimension, intention, as a synthesis of awareness and intention. The intention dimension is framed as three essential divisions: aware/intended, aware/unintended, and unaware/unintended. Alkin and Taut are thus able to parse use and influence within Kirkhart's framework: the notion of use comprises the aware/intended and aware/unintended dimensions of process and findings use within the immediate and end-of-cycle timeframes, and influence occupies the remaining dimensions (i.e., all process or findings effects that are long-term and/or unintended/unaware). In addition to illustrating how use and influence complement one another, Alkin and Taut highlight that the typically short-term nature of evaluation tends to prohibit evaluators from detecting long-term and unintended/unaware effects. Use and influence are argued to be complementary,

yet distinctive, elements of understanding the ways in which evaluation affects people and situations.

Also in response to Kirkhart's (2000) call, Henry and Mark (2003) generated a theoretical framework to drive comparative and empirical research on evaluation by delineating typical processes and outcomes of evaluation. While they position the use construct as a "handy, informal" (p. 309) way to refer to evaluation consequences, their framework supports three imperatives: first, to provide a shared language for considering evaluation influence, second, to offer an alternative framework to organize influence across multiple tiers (individual, interpersonal, and collective levels), and, third, to create an opportunity to connect the concept of evaluation influence with the research literatures of other fields (provided in their work). Additionally, while traditional notions of use are focused at the level of the individual involved in an evaluation, Henry and Mark characterize processes and outcomes at three distinct levels of influence: influence at the individual level (e.g., attitude change, salience, elaboration, priming, skill acquisition, and behavioral change), influence through interpersonal interactions between individuals in the evaluation (e.g., justification, persuasion, change agency, social norms, and the influence of minority opinions) and, influence that occurs at collective levels (e.g., agenda setting, policy oriented learning, policy change, and diffusion). By structuring the notion of influence in this way, processes and outcomes can be considered as 'outcomes chains' that can be tested, providing a means for researching evaluation influence empirically. Henry and Mark position this approach as an effort to address the

intrapersonal, interpersonal, and societal changes that connect evaluation with what they view as its ultimate purpose, the goal of social betterment (p. 294).

Continuing this line of research, Mark and Henry (2004) provided a more nuanced list of processes and outcomes relative to evaluation influence, including a crosswalk of their work with traditional use typologies. This expanded model crosses the levels of analysis from their previous work (individual, interpersonal, and collective) with various types of processes and outcomes (general influence, cognitive and affective, motivational, and behavioral), depicting mechanisms for each type at each level. Mark and Henry highlight that while instrumental, conceptual, and symbolic use can be identified within a model, the nature of process use prohibits its inclusion: process use is positioned as a trigger for influence, either from the single evaluation instance or the accumulation of evaluation findings over time. Mark and Henry also offer a general logic model for evaluation (i.e., a visual depiction that facilitates a better understanding of theory and meaningfully influences theory-based practice; Gargani, 2013) as a "useful starting point for future theoretical and empirical work, rather than as a final product" (Mark & Henry, 2004, p. 50). Finally, the authors warn that evaluators need not focus on increasing influence, but rather on the improvement of the quality of influence.

These foundational theoretical frameworks have spurred research on evaluation influence, efforts summarized in a literature review spanning 2000-2014. Within the literature review, Herbert (2014) consolidates the three available frameworks and consolidates the available empirical literature on evaluation influence. Herbert found three types of empirical studies, including nine descriptive, 15 analytic, and four

hypothesis testing studies of influence. Across these 28 studies, it is important to note the inconsistency of definitions and of applying theoretical lenses related to influence. Herbert found that the nine descriptive studies characterized influences identified in real-world contexts without analysis of factors that may have hindered or contributed to influence. While these studies contribute little to a grow understanding of the influence construct, they do highlight how evaluation influence manifests in practical contexts, demonstrating its value for guiding evaluation practice (Herbert, 2014). The bulk of studies (15 of 28) were analytic in nature and sought to explain how identified influences unfolded in specific cases using a retrospective, exploratory, qualitative methodology. By contrast, the final type of study, hypothesis testing, sought to test specific mechanisms of evaluation influence, typically using the construct of evaluation influence as the basis for study of an evaluation or evaluation users. Herbert concludes his review of the literature by offering four unique definitions of evaluation use: as an expansion of the use construct; as a consideration of various levels of effects; as a framework of mechanisms that parallel use types, and as pathways of influence that seek to connect mechanisms and outcomes. Given this multiplicity of notions of influence, Herbert underscores limitations of existing research on influence and suggests further empirical study.

One analytic study from Herbert's (2014) review researched processes and outcomes of influence within a multi-site monitoring and evaluation (M&E) of a Ghanaian governmental health program (Gildemyn, 2014). In this study, Gildemyn utilized a multi-site case study design to better understand the influence of a single evaluative process (an interface meeting) by comparing influence across sites that either

had or did not have such meetings. In addition to suggesting that interface meetings created spaces within which evaluation influences could occur, results from the Gildemyn's study identified additional influence processes and outcomes that could be mapped to generate a theoretical framework of influence for the M&E for health programs. This study demonstrates the value of using a case study approach to analyze influence within a specific type of evaluation (M&E) and a specific applied context for evaluation (health programs).

Overall, while the field has been receptive to the suggestion of expanding the notion of evaluation use to one of evaluation influence, research on use is still considered valuable (Mark, 2011) and efforts to position the constructs of use and influence as complementary notions have been offered (Alkin & Taut, 2003; Mark & Henry, 2004). Authors of the most recent systematic review of the literature on use (Johnson et al., 2009) suggest that influence provides a unifying construct to facilitate a more nuanced understanding of the consequences of an evaluation. Theoretical frameworks and studies on evaluation influence position empirical studies of influence as efforts to more comprehensively address the ways in which evaluation affects situations and people.

Positioning Evaluation Use and Influence in Assessment-based Educational Evaluation

The second section of this literature review positions study of use and influence within a specific context of evaluation practice, that of assessment-based educational evaluation. While evaluation has a long history in education (Rickards & Stitt-Bergh, 2016b), a specific context of educational evaluation has emerged over the last 30 years

within which evaluation practice is less well-understood: assessment. Assessment is a specific form of educational evaluation that has traditionally focused on accountability and compliance reporting (Banta & Palomba, 2014; Ikenberry & Kuh, 2015; Kinzie et al., 2015). Since the turn of the century, assessment mandates have grown to include a focus on learning and improvement purposes through the use of direct evidence (e.g., student learning evidence; Kuh et al., 2015b). Evaluators have much to offer and to learn from the study of evaluation use and influence in this context. Educational evaluation leverages contextually influenced information about *what students know and can do* to make decisions across many levels of the educational organization. These decisions range from student-level determinations of individual student achievement and academic progression to program-level decisions regarding how to best use evidence to support and improve program quality. Since many evaluators work in educational evaluation and in assessment, investigating the nature of evaluation use and influence in this nuanced context of evaluation practice is an important effort. In this section, the context of the current assessment moment is described and positioned relative to the opportunity it offers for research on evaluation use and influence. The section concludes with a description of the benefits of in-depth case study of use and influence in assessment-based educational evaluation, offering research questions to guide this inquiry.

The history of assessment in higher education. From a historical perspective, assessment is rooted in educational measurement as well as in calls for public accountability for educational quality. Kinzie and colleagues (2015) trace the historical roots of assessment to studies in the 1930s that sought to measure and interpret cognitive

gains. As they describe, interest in assessment shifted over the next 40 years to focus on two main targets: the cumulative outcomes associated with college attendance and scholarly research into student learning mastery. By the 1980s, accountability pressures emerging in K-12 education began to influence higher education, positioning the focus of assessment on compliance with external expectations from accrediting bodies (Banta & Palomba, 2014). This trend that has persisted and been reinforced by federal and state policymakers causing assessment work to focus primarily on providing accountability and compliance evidence to satisfy accreditation standards (Banta & Palomba, 2014; Ikenberry & Kuh, 2015).

From the 1980s until the turn of the century, accreditation standards largely focused on the sufficiency of institutional resources to deliver educational programming with an emphasis on program inputs (i.e., those assets supporting program implementation such as faculty credentials, curricular coherence, library resources, and fiscal integrity; Ikenberry & Kuh, 2015). While assessment practitioners acknowledge the importance of this function (Kuh et al., 2015b), a focus on compliance has traditionally edged out opportunities to use assessment processes and evidence for organizational learning and improvement, mostly due the necessity and urgency that cause the prioritization of compliance reporting over improvement or learning efforts (Ikenberry & Kuh, 2015; Kinzie et al., 2015). As previously stated, applying a paradigm of learning to assessment constitutes a revolution in the field (Suskie, 2004). Barr and Tagg (1995) introduced a “learning-centered paradigm” and called for a shift in the focus of assessment from teaching to learning: a focus on teaching utilized assessment for grading

students, while a focus on learning utilizes assessment for understanding what does or does not work in curriculum or instruction. Following a teaching-centered paradigm, indicators in assessment were historically predicated on forms of indirect evidence like course grades, grade point averages, or perceptual surveys (Rickards & Stitt-Bergh, 2016b). These types of indirect measures of learning fail to provide evidence capable of informing improvement. Encouragingly, accreditation requirements have been revised gradually to mandate the use of assessment evidence within the educational organization, resulting in measurable increases in the use of assessment for accreditation, program review, curricular modifications, and institutional improvement (Kinzie et al., 2015). Accreditors have also introduced requirements that mandate the use of direct evidence of student learning, resulting in the increased capacity of educational organizations to produce and use such evidence for assessment (Ikenberry & Kuh, 2015). This shift toward student learning evidence is non-trivial given that accreditation is the driving force behind assessment efforts and given that the bulk of assessment work is typically undertaken in conjunction with accreditation site visits (Kinzie et al., 2015).

The current moment in assessment in higher education. Though these increases in actual use and organizational capacity for use of evaluative evidence are encouraging, assessment is negatively impacted by two significant vestigial remnants of its compliance-oriented past—a mindset of compliance culture and a perception of assessment as a compliance-oriented process. First, as leaders in the field of assessment point out, assessment is still oftentimes undertaken within a “culture of compliance” (Kuh et al., 2015b) that distances assessment from the day-to-day experiences of faculty

and students (Ikenberry & Kuh, 2015). When assessment is positioned as an accountability and compliance structure facilitated by assessment professionals, it fails to understand and contribute to the lived experiences of internal program stakeholders. This can result from the sheer complexity of stakeholders involved in assessment across many levels of the educational infrastructure, including: accreditors and government officials, university and college presidents and provosts, assessment professionals and other administrators in addition to faculty and students (Ikenberry & Kuh, 2015). Another factor that contributes to the distancing of assessment from the program is when student learning evidence is not used for decision making or improvement efforts (Kinzie et al., 2015). As distance between assessment processes and program stakeholders increases, faculty "tend to adopt a role of passive resistance and often become a barrier rather than a pathway to consequential assessment work" (Ikenberry & Kuh, 2015, p. 16). It is believed that a focus on student learning evidence collapses this distance by positioning assessment as a mechanism for addressing faculty concerns about teaching and learning based on measures of student performance (Kuh et al., 2015b). This can be achieved through meaningful involvement of faculty and students throughout all phases of the assessment process (akin to participatory, transformative, and culturally responsive approaches to evaluation) and by grounding assessment in direct evidence of student learning.

Second, traditional compliance-oriented processes associated with assessment pose a challenge to using assessment for learning and improvement. At present, accreditation requirements tend to entail the generation of assessment plans with a

description of their intended uses, but fall short of requiring educational organizations to provide evidence of actual use (Jonson et al., 2017). While creation of assessment plans is relatively easy, later phases of the assessment process (i.e., collecting data, reporting findings, designing next steps) become increasing challenging to complete (Kinzie et al., 2015). This highlights a hallmark of assessment for compliance purposes: the assessment process tends to conclude when data is collected and measurements are recorded (Kinzie et al., 2015), not after the results have been used through some sort of intervention to produce an effect. Unsurprisingly then, assessment has tended toward a mechanistic focus (Blaich & Wise, 2011; Ikenberry & Kuh, 2015; Kinzie et al., 2015), one centered on collecting and reporting data with less emphasis on subsequent use of results (Ikenberry & Kuh, 2015). Furthermore, quality of evidence is not a consideration for compliance reporting, possibly furthering a mentality that any use of evidence is more critical than appropriate use of evidence. As a result, a broad trend in assessment is that actionable evidence is generated with little to no tangible follow-up response (Blaich & Wise, 2011).

The future of assessment in higher education. Contemporarily, assessment is a process of gathering and using evidence of student learning for decision making across all levels of the educational organization (Ikenberry & Kuh, 2015). To actualize this aspirational vision of what assessment can and should be, some next steps have been offered that will help practitioners close the gap between compliance- and improvement-oriented mindsets. First, engagement of faculty throughout all phases of assessment engenders ownership and fosters assessment use for improvement and learning (Kinzie et

al., 2015). Second, effective communication structures and practices are necessary to bring assessment efforts into alignment (Jankowski & Cain, 2015) so that assessment may successfully bridge classroom, course, program, department, college, and organizational information needs, priorities, and stakeholder groups. Engagement throughout the entire process and strong communication structures facilitate alignment of assessment needs and interests across various levels of the educational organization. Most use of assessment has tended to take place at the program level (Kinzie et al., 2015) where it tends to be used to answer questions about individual student needs or about areas of faculty interest. Use of student learning evidence can at higher order levels within the educational organization serves to strengthen connections between assessment and institutional goals, strategic planning, institutional decision making, accreditation processes, institutional assessment planning, improving student engagement and success, building a culture of teaching and learning, enhancing faculty collaboration, and reflecting on current assessment processes (Kinzie et al., 2015). To this, Ikenberry and Kuh (2015) add refining learning goals, courses, and curricula, considering technology, informing the budget, improving retention and graduation rates, improving American higher education, and improving the prospects of graduates. Alignment of uses across these levels is necessary to address the constantly evolving needs of information users (Jonson et al., 2017) as well as to develop shared understanding of assessment processes and products (Jankowski & Cain, 2015).

Overall, assessment faces a critical juncture as it seeks to expand the traditional scope of assessment beyond accountability and compliance by leveraging student

learning evidence for organizational improvement. As Ikenberry and Kuh (2015) argued, accountability is a necessary purpose for assessment, but without alignment with local needs, assessment constitutes a missed opportunity and a waste. Kinzie and colleagues (2015) frame this transition in an exploratory light, stating there is still a great deal that needs to be learned about how to do this work well and posing the following questions to direct this line of inquiry: How can practitioners transition from "doing assessment" to using assessment findings? What strategies influence use? What principles can be used to guide efforts to increase use and improve educational organizations? Thus, as with evaluation, the current moment in assessment is poised in tension between accountability and learning goals, but optimistic about approaches to integrate these purposes by using student learning evidence as a key lever.

Assessment as a practical context of educational evaluation. As a structure for compliance reporting, assessment was essentially distinct in nature from evaluation. However, the emerging focus in assessment on increasing the use of its processes and findings, as well as on using evidence of student learning in these efforts, positions assessment as a context of evaluation practice and a specific form of educational evaluation. Efforts to use assessment for improvement and learning leverage stakeholder involvement by building partnerships between evaluators and stakeholders, positioning student-level evidence for use within higher order learning and improvement efforts, and documenting evaluation results that are consequential to teaching and learning as well as to faculty and students. It is hoped that these efforts will support the creation of a "culture of assessment" (Ikenberry & Kuh, 2015; Kinzie et al., 2015), a goal not unlike calls in the

evaluation literature for a "culture of evaluation" (Loud & Mayne, 2014; Mayne, 2009). Like evaluators working in internal evaluation units who are seeking to integrate use into the fabric of an organization's culture (Loud & Mayne, 2014), internal evaluators working in assessment (and assessment practitioners using evaluation) are seeking to gain "consequential use" (Ikenberry & Kuh, 2015) of assessment-based educational evaluation by grounding assessment in student learning evidence.

While the exact nature of the relationship between evaluation and assessment has not been explored extensively, the editors of a recent issue of *New Directions for Evaluation* entitled "Evaluating Student Learning in Higher Education: Beyond the Public Rhetoric" position assessment as "utilization-focused outcome evaluation with goals that include program/organization improvement and an integration of evaluative thinking in the program and organization" (Rickards & Stitt-Bergh, 2016a, p. 7). Noting the lack of intersection between the assessment and evaluation literatures, the editors suggest that assessment offers new opportunities for evaluators, especially those interested in the use of evaluation within developing and complex institutions. As they outline, while educational evaluation has a longstanding history, a recent focus on evidence of student learning outcomes has made some traditional approaches to assessment obsolete, including the use of course grades, grade point averages, and surveys of student perceptions (Rickards & Stitt-Bergh, 2016b). They frame assessment-based educational evaluation as one grounded in the generation of student learning evidence within a recursive process of teaching, collecting data, and making sense of the data. Within this iterative cycle, faculty engage in evaluation design, data collection,

analysis, reporting, and follow-up action. Accordingly, engagement of faculty in these quality improvement and quality assurance processes is becoming a standard expectation within typical faculty expectations for teaching and learning (Kuh et al., 2015b).

Stitt-Bergh and colleagues (2016) typify the challenge facing evaluators in assessment as requiring the development of systems that generate evidence of student learning in ways that are useful to faculty and institutional improvement while also creating required documentation for compliance reporting. They argue that evaluators with technical knowledge, interpersonal skills, cultural awareness, and contextual knowledge of higher education are needed to meet quality demands, leverage "big data," facilitate "closing the loop," meet documentation requirements, negotiate communication, and partner with faculty. Rickards and Stitt-Bergh (2016b) also suggest that deepening connections between the fields can be mutually beneficial: the participatory and collaborative approaches of evaluation can improve the quality of assessment, and research on evaluation in the assessment context can inform the evaluation theory and practice relative to specific contexts of practice.

Considering evaluation use and influence in assessment. The use of student learning evidence is currently receiving a great deal of attention as a key lever in the effort to shift the focus of assessment beyond accountability and compliance and toward improvement and learning (Ikenberry & Kuh, 2015; Jonson et al., 2014). As a context of evaluation practice, assessment provides an interesting opportunity for researching evaluation use and influence in at least five specific ways. First, assessment facilitates evaluation iteratively within a specific context, permitting research on the relationship

between evaluation and context more deeply (Rickards & Stitt-Bergh, 2016b). Second, accreditation requires documentation of assessment efforts for compliance reporting (Kuh et al., 2015b), providing an interesting and unique opportunity for evaluators to study the reported uses and influences of evaluation. Third, since assessment interventions require documentation of intended and actual uses of student learning evidence, analysis of intended and actual use may contribute to discussions within the field surrounding intention and consequences. Fourth, assessment occurs in educational organizations that are focused on tiers of learning (e.g., at the student, program, and organizational levels) permitting the study of use across levels and stakeholder groups. And fifth, like evaluation, assessment is facing the challenge of negotiating accountability and learning purposes (Ikenberry & Kuh, 2015; Kinzie et al., 2015; Rickards & Stitt-Bergh, 2016b), meaning that research conducted in assessment may be useful to inform this tension more broadly within evaluation.

Interestingly for evaluators, assessment is an applied form of evaluation within contexts that have an explicit learning focus (Rickards & Stitt-Bergh, 2016b). Unlike most evaluation contexts, the methods of data collection for assessment (e.g., the tests, projects, portfolios, presentations, performances, among others, in which student performance is measured) have their own unique utility. That is, while student learning evidence may be used for the assessment purposes of program improvement, its primary use is to evaluate individual student for formative and summative purposes: to identify individualized areas for student improvement and to determine whether students have attained mastery of learning goals to a sufficient degree. From a bottom-up perspective,

student learning evidence can be used for three distinct purposes: as evidence of student learning mastery (used for decision making about individual students), as evidence for accountability (used for documenting compliance for audiences outside of the immediate educational unit), or as evidence of organizational learning (used for improving the quality of educational programming).

This multiplicity of purposes, and the appropriateness for using measures of student learning as evidence for such work, suggest a line of inquiry investigating connections between assessment and evaluation. How are the theories and practices of the two fields similar? On what characteristics or issues might they be distinct? What issues and tensions, challenges and barriers to practice exist within each field? How are research and evaluation leveraged in educational organizations for generating, interpreting, and using evidence for decision making and action taking? How might research on evaluation support assessment efforts? Preliminary investigation into this area suggests that scholars in both fields are interested in increasing connections between the fields (in evaluation: Rickards & Stitt-Bergh, 2016b; in assessment: Ikenberry & Kuh, 2015; Jonson et al., 2014; Kinzie et al., 2015).

A case study of evaluation use and influence in assessment. Inquiry into evaluation use and influence in assessment offers potential benefit to evaluation practitioners working in assessment as well as to increasing understanding of evaluation within specific contexts of practice and across organizational boundaries (Rickards & Stitt-Bergh, 2016b). The driving goal of this work is to develop a greater understanding of evaluation use and influence within the practical context of assessment. The study will

seek to generate a more nuanced understanding of the influence of student learning evidence on assessment-based educational evaluation. Furthermore, this study provides an opportunity to "examine the ways in which information—evaluative and otherwise—is packaged, diffused, understood, and utilized to serve decision-making purposes" (Christie & Vo, 2015, p. xviii). In a report on the use of evidence in public policy, it is suggested that a gap exists in understanding the systems and structures within educational organizations that might facilitate research use and evaluation use (Prewitt et al., 2012). Similarly, there have been calls for more micro-level research on the use of data within the everyday practices of educational programs (Spillane et al., 2002). Given the primacy of the use construct within the field of evaluation, there is a continued interest in research on evaluation use and influence, specifically regarding the use of evaluation for decision making within specific contexts of evaluation practice (Christie, 2015). Studying use and influence in the practical context of assessment offers evaluators the chance to study evaluation use more deeply (Rickards & Stitt-Bergh, 2016b). This opportunity is somewhat unique among contexts of evaluation since assessment includes an explicit goal of "closing the loop" (Kuh et al., 2015b) by documenting use and its associated follow-up responses.

This study responds to the call for an expanded consideration of use as influence. Since evaluation influence frameworks are relatively new within the literature, few empirical studies have been conducted to explore these frameworks in the various contexts of evaluation practice (Herbert, 2014). Much like Gildemyn's (2014) case study of evaluation influence in a multi-site M&E approach, this case study will facilitate an

increased awareness of evaluation influence within the practical context of assessment. Findings from this in-depth study will contribute to a greater understanding of the dimensions of influence (Alkin & Taut, 2003; Kirkhart, 2000) that are closely associated with using student learning evidence for assessment-based educational evaluation. This work provides guidance for evaluators working in assessment in their efforts to support stakeholders using student learning evidence across various levels of an educational organization (Porteous & Montague, 2014). This is particularly relevant given that the use of student learning evidence for assessment-based improvement and learning is currently receiving a great deal of attention (Ikenberry & Kuh, 2015; Jonson et al., 2014; Kinzie et al., 2015) and evaluators are being called on to fill this gap (Rickards & Stitt-Bergh, 2016b).

In addition to this theoretical relevance, findings from this in-depth case study generate further conceptual connections between the theory and practice of evaluation and the practical context of assessment in which evaluation is used, a relationship that has been suggested in both evaluation and assessment literatures (Ikenberry & Kuh, 2015; Jonson et al., 2014; Rickards & Stitt-Bergh, 2016b). This study responds to these calls for research on the use of evaluation in assessment. Jonson and colleagues (2014) argue for the utility of evaluation influence for addressing the current assessment climate, furthermore suggesting qualitative interviews can be utilized to characterize how practitioners perceive the assessment process and findings (Jonson et al., 2017).

Research questions. Deriving from these broader areas of interest, the following research questions have been developed to guide this inquiry. In addressing these

questions, this research study will provide an indication of how student learning evidence can be used for various purposes for assessment-based educational evaluation. The effects of student learning evidence will be considered at the student and program levels for a single case. In effort to bound the contextual complexity of educational evaluation, the case will focus on a single source of student learning evidence: a clinical skills examination system (CSE system) that generates student learning evidence for formative and summative uses at the student and program levels. The effects of evaluative information will be analyzed for accountability and learning at the student and program levels according to the following research questions:

Research Question 1. Based on a specific case of an innovative clinical skills examination system (CSE system) in a medical school, what is the nature of evaluation use and influence in assessment-based educational evaluation?

1.1 What types of use (instrumental, conceptual, symbolic, process, etc.) are made of student learning evidence within the educational organization?

1.2 To what extent are uses oriented toward accountability and learning?

1.3 How are these findings affected by expanding the notion of evaluation use to consider dimensions of evaluation influence?

Summary

A key construct in evaluation, use comprises the various ways in which evaluation affects people and situations. Use has been an enduring focus of evaluation research for the past 40 years, producing a variety of use typologies (chiefly instrumental, conceptual, symbolic, and process) and a robust description of factors influencing use. At

the beginning of the twenty-first century, and grappling with issues about the sufficiency of research on use, evaluators suggested that a focus findings-related use artificially limited one's ability to understand use. A notion of evaluation influence emerged to address the less direct and measurable ways in which evaluation exerts influence on people and situations. While some empirical research on evaluation influence has been conducted, continued research on the complementary notions of use and influence is warranted, especially within specific applied domains or contexts of evaluation practice.

While educational evaluation has been well-researched, assessment in higher education is an educational context in which evaluation has not yet been sufficiently explored. Given a historical focus on accountability and compliance purposes, assessment for learning and improvement is a relatively new and emerging practice in education. Research into the relationship between education and assessment is only recently emerging. Given its embedded mechanisms for data collection through traditional educational activities like testing and grading, and given its iterative processes for continuous evaluation, assessment provides a unique and interesting context for research on evaluation use and influence, especially regarding the tension between accountability and learning purposes for evaluation.

CHAPTER III

METHODOLOGY

The primary objective of this study is to develop a greater understanding of evaluation use and influence within the practical context of assessment. Findings from an in-depth case study of these concepts may provide further support for pre-existing conceptual frameworks surrounding use and influence; findings may also suggest additional considerations that may be unique to assessment-based educational evaluation. As such, a single instrumental case study approach is appropriate to facilitate this in-depth study. Since multiple definitions, types, and purposes for case study exist (Merriam, 2002; Simons, 2009; Stake, 1995; Yin, 2009), the remainder of this chapter serves to clarify the nature of the case study methodology and the nuances of its implementation in this analytic study of evaluation use and influence in assessment-based educational evaluation.

As a methodology, the case study approach provides a set of principles and values to guide the process through which the researcher gains and maintains access to undertake, analyze, and interpret a case (Simons, 2009). In alignment with the purposes of for this research on educational evaluation, the overarching motivation for undertaking case study is to generate a thick description and portrayal of specific events, circumstances, or people in which the uniqueness of the single case is of interest. In telling the story of the single case, the researcher focuses attention on the idiosyncrasies

of a ‘specific, complex, functioning thing’ (Stake, 1995), a particular process (Merriam, 2002), or a single policy (Simons, 2009). As an approach, case study is useful in situations in which researchers cannot control events that unfold within a real-world context of study (Yin, 2009). Creswell offers guidance for the selection of a case study approach along five criteria (1998), dimensions along which this study of the uses and influences of assessment-based educational evaluation is well-suited: in-depth analysis of the particularity of the uses and influences of educational evaluation is of interest, the approach is rooted in the discipline of evaluation (as opposed to knowledge-oriented social science research), multiple methods will be used in data collection, data analysis will proceed through the generation of description, themes, and assertions, and the narrative form will be used for reporting. Using qualitative case study methods will permit an in-depth exploration of student learning evidence use and influence within the educational organization.

Case study for educational research and evaluation emphasizes a constructivist nature and a focus on the particularity of the case (Simons, 2009): “Case study is an in-depth exploration from multiple perspectives of the complexity and uniqueness of a particular project, policy, institution, programme or system in a ‘real life’ context. It is research-based, inclusive of different methods and is evidence-led” (p. 21). Case study methodology permits detailed exploration of singular manifestation of larger grain sized constructs like policy or professional practices. In this study, the uses and influences of assessment-based educational evaluation will be researched for a single case. By seeking to explore this specific, pre-determined issue in a real-world scenario, the case study will

be instrumental in nature (Stake, 1995). The study design can be summarized as an interpretive case study that followed an emergent design so that findings from each phase refined subsequent analyses through progressive focusing (Simons, 2009). Through layered analysis (Creswell, 1998), this approach permits comparison between an interpretive lens of evaluation use (research questions 1.1-1.2) and one of evaluation influence (research question 1.3).

Bounding the Case

In order to investigate the uses and influences of a single source of evaluative information within a single curricular domain, the case will be bound in five ways. First, while there are more than 150 educational organizations conferring medical degrees within the United States, the case study will consist of a single educational organization. Second, while medical education is composed of diverse training across six Accreditation Council for Graduate Medical Education (ACGME) core competencies (i.e., medical knowledge, practice-based learning and improvement, patient care, systems-based practice, professionalism, and interpersonal and communication skills; Association of American Medical Colleges, 2008), the case will comprise a single domain of medical education (i.e., clinical science) through which medical students are trained on the core competencies of patient care and interpersonal and communication skills. Third, while clinical science involves many instances of assessment-based educational evaluation, the case study will be bound to consider the uses and influences of a single source of evidence (i.e., a clinical skills examination system that produces formative and summative score reports that include quantitative and qualitative feedback on student

performance). Fourth, while the testing system serves many purposes for clinical science training (e.g., across the medical school, residency, fellowship, and physician training and continuing education programs), the case study will be limited to the uses and influences of student learning evidence within pre-clerkship clinical training (i.e., the training of first and second-year medical students). Fifth and finally, while the CSE system is used for workshop-based training, self-driven practice, formative testing, and summative testing, the case study will only include evidence for which official score reports are available (i.e., formative and summative student learning evidence).

The case is thus bound and defined as a single source of student learning evidence for formative and summative purposes within a single setting in which assessment-based educational evaluation occurs. While this source is a testing system, a similar approach could be taken to look at the evaluative evidence produced by other methods that are employed to evaluate student performance (e.g., projects, portfolios, presentations). The CSE system has been designed for longitudinal use across multiple courses in pre-clinical medical education in alignment with internal and external frameworks that shape the content of its individual tests and the quality, frequency, and specificity of the evidence it generates. Evidence of student learning produced by the performance assessment system is used for student improvement, for decisions regarding the promotion, remediation, or retention of students, for improvement of the CSE system, for improvement of the curriculum and its implementation within the program, for reporting within the educational organizations, and for accountability and compliance reporting beyond the educational organization. The overarching purpose for the CSE system is to provide

quantitative and qualitative feedback that supports student development toward mastery of the learning goals established by the educational organization, toward demonstrating mastery on a high stakes examination given at the end of the third year of medical school (the United States Medical Licensing Examination STEP 2 Clinical Skills examination, hereafter referred to as the STEP 2 CS exam), toward demonstration of minimal competency to transition to the clerkship portion of medical school (third and fourth years), and toward preparation for future practice as a professional physician. The evidence generated by the CSE system is used as a measure of these overarching goals, specifically relative to specified student learning outcomes subsumed under these overarching goals.

Case Selection

In two ways, the case selection was based on convenience sampling. First, I selected an educational domain (clinical science) with which I had previous professional experience and a personal working knowledge. Second, approval for a research study was facilitated through relationships I had previously established while working in clinical science. These considerations facilitated the selection of the case and research site. However, this justification for case selection is somewhat misleading. The case and research site were chosen because of the perceived utility of the evaluative information it produces for apprising student learning while simultaneously generating student performance evidence applicable to organizational learning and accountability purposes. The system is thoughtfully designed in alignment with many external frameworks of quality and is updated in real-time (i.e., from week to week as well as from year to year)

through quality improvement processes predicated on the evaluative information it generates. Furthermore, the CSE system is longitudinally implemented along a developmental continuum, thus providing informal, formative, and summative score reports that consist of quantitative scores and qualitative commentary on areas or specific skills for student improvement. In this sense, the case was selected because it is an intensity case (i.e., an information-rich case that intensely manifests a wide variety of uses and influences of the student learning evidence it generates; Miles & Huberman, 1994).

Researcher Position

Within the qualitative case study approach, the researcher is the primary instrument for data generation, interpretation, and use, prompting the need for explication of how the researcher's values and actions shape the process as well as how the study impacts the researcher (Simons, 2009). Situational subjectivity (i.e., the differential influence of self on the research through paradigms, values, and meaning-making influences elicited by various situations in which the researcher may find him/herself; Peshkin, 1988)—is thus an inherent element of the research frame, one that must be readily acknowledged and monitored (Simons, 2009). The importance of analyzing one's subjectivity (Peshkin, 1988) is intensified by the political nature of evaluation as a means for determining the allocation of resources and opportunities in response to evaluative information. There is no doubt that my professional identity influenced my perception of this research. Overall, this study serves as a continuation of my ongoing interest in educational improvement, interests that I developed over the past ten years working as an

educator, educational administrator, and doctoral student of evaluation. My perspective is informed by this practical lens and has prompted my interest in studying research on evaluation to investigate these constructs in the specific context of assessment-based educational evaluation as facilitated by internal evaluation units. As Loud and Mayne (2014) suggest, these efforts can further our understanding of evidence use and influence in evaluation while concurrently facilitating the use of student learning evidence in educational organizations.

My academic interests in evaluation cannot be separated from my past professional experiences in education as a student (before and after being trained in teaching and learning), a teacher of high school and undergraduate students, a high school administrator responsible for externally mandated standardized testing, a clinical science coordinator facilitating educational programing and seeking to improve it under practical constraints (time and necessity), and as an evaluator interested in supporting educational improvement efforts. In these efforts, one must learn to balance constructivism and pragmatism, seeking to continually increase effectiveness while improving program quality. My training as an evaluator disposes me to view assessment-based educational evaluation as a manifestation of the larger tensions and issues facing the field of evaluation. Such issues include the information and power gaps within and beyond organizational structures that result from mechanistic approaches to evaluation; the struggle stakeholders face in trying to navigate the subjective interpretation processes *in situ* that influence the greater objective meaning of information; and the need for methods that engage systemic complexity by negotiating and collaborating across the

traditional boundaries of professional domains. Research in real-world contexts, appreciative inquiry, and an assets-based orientation can address dilemmas regarding evaluation and information needs in specific contexts in which evaluation takes place. Rationality and sensemaking processes can facilitate the design, implementation, and use of evaluation, acknowledging the interplay of meaning and valuing in developing contextually relevant understandings. Research on evaluation use and influence can serve efforts that seek to better understand and use student learning evidence in educational improvement efforts.

Data Collection

This qualitative case study strategy utilized multiple methods, specifically document review and three sets of interviews, to facilitate in-depth analysis of the particularity of the case (Simons, 2009). Overall, data for this study was collected through the review of 11 program documents, an initial round of interviewing (24 interviews), a second round of interviewing framed as a reflective exercise in which students interpreted score reports (five interviews), and a final follow-up interview with the administrator most directly responsible for the program (one interview). As previously mentioned, a single round of interviews was initially planned; however, following the first round of interviews, the reflective exercise was developed to engage students in sensemaking processes so as to observe how they responded to actual score reports. While observations are typically included within qualitative case study, the method of observation would not have been suitable for characterizing and analyzing the uses and influences of student learning evidence in assessment for two reasons: first,

observations of testing instances (students participating in testing procedures) would do little to inform the study and second, sensemaking and subsequent use of student learning evidence tend to be internal processes that unfold over time outside of any routinely scheduled activity.

Interviews. As seen in Table 1, a total of 30 interviews were conducted to explore the uses and influences of evaluative student learning evidence: 24 interviews of program stakeholders (including eight program administrators, seven current students, and nine former students), five reflective exercise-based interviews of current students, and one follow-up interview with the program administrator most responsible for the program.

For each of the first two rounds of interview questions, protocols were reviewed with program administrators to adjust wording and refine intentionality. As a result, language was refined in attempt to improve clarity. This effort served as a pilot of the interviewing protocols to anticipate how questions would be understood and to support interviewees' engagement and reflection on their personal experiences with evaluative information (Maxwell, 2005). The protocol for these in-depth, semi-structured interviews can be seen in the appendix. These in-depth, semi-structured interpersonal interviews were implemented as described by Simons (2009), and the interviews were audio recorded for selective transcription. Throughout the course of the interview, participants were asked to explain their own use of evaluative evidence, shifting the focus of the dialogue and some level of control over the conversation to participants.

The purpose of the first interview was to gather a rich understanding of the uses and influences of the evaluative information produced by the CSE system, including a

characterization of the ways in which and the purposes toward which evaluative information could be used. The eight program administrator interviews ranged from 20-90 minutes in length, while the 16 student interviews lasted 15-30 minutes. All student interviews were conducted within a short time frame, providing a snapshot of evaluation use and influence at a specific point in time. From the first round of interviews, the idea of asking students to interpret a real score report emerged, resulting in the development of a second round of interviews that were framed as a reflective exercise that lasted 5-15 minutes. The second interviewing protocol was developed and framed as a reflective exercise in which five study participants interpreted the meaning of score reports to address this internal process of sensemaking. The goal of the reflective exercise was to investigate how students use and are influenced by actual score reports of evaluative information. In this interview, five students accessed summative score reports immediately after the reports were released and engaged in a self-directed talk aloud by voicing both their understanding of the meaning of the evidence and their plans for action in response to the evaluative evidence. To conclude the reflective interview, students responded to questions about the relevance of the information to their future development in clinical science training.

Finally, following data collection from the first two rounds of interviews, a 90-minute phone interview was conducted with the administrator most responsible for the program. While the primary purpose for this follow-up interview was member checking, additional data and examples were offered during the course of the dialogue. This

interview was not audio recorded, but was selectively transcribed throughout its duration, and the selective transcription was analyzed alongside other data sources.

Table 1

Interviews of Various Stakeholder Groups

	Interview 1		Interview 2		Follow-up Interview	
	Number	Length	Number	Length	Number	Length
Program Administrators	8	20-90 mins	--	--	1	90 mins
Current Students	7	15-30 mins	5	5-15 mins		
Former Students	9	15-30 mins	--	--		
	24		5		1	

Document review. The use of student learning evidence in educational evaluation is intrinsically related to policies, rules, and regulations that govern educational organizations. To gain a clear and comprehensive understanding of these overarching representations of the organization, a thematic analysis of documentation was undertaken. Documents were collected through an analysis of the available clinical science training literature, then categorized into two main categories: quality frameworks (i.e., standards and objectives from external governing bodies as well as those internal to the educational organization) and information use protocols (i.e., standards and required processes for evidence use and reporting). Documents were then subcategorized as either external (imported) or internal (developed locally) to the educational organization. Documents

were reviewed with program administrators to determine which policies and procedures influenced the department and/or the CSE system. This process of document review by exclusion of irrelevant documentation, with subsequent addition of local documentation, served to ensure as many documents as relevant were included and to bring awareness to the program stakeholders about other frameworks that exist to guide clinical science education. Relevant documentation was then analyzed to determine the significance of each document to the study (Miles & Huberman, 1994). As a sample, these documents collectively represent the purposes and values influencing student learning evidence use from the CSE system, including the intended and actual uses of this evaluative evidence in practice.

Sample selection and characteristics. The participant sample included program administrators as well as students. The sample notably excluded instructors since, in the context of medical education, the many physicians who teach individual class sessions or a series of classes are typically not responsible for the formal evaluation of students. Instead, traditional teacher responsibilities are shared across instructors and educational administrators, with the latter being responsible for educational evaluation activities. Accordingly, all administrators within the educational organization related to the program were interviewed. These individuals were identified through collaboration with an administrator directly in charge of the program.

Another interesting characteristic of the sample is the inclusion of student participants. Since assessment-based educational evaluation typically targets the program level, student perspectives are typically not solicited. However, Ikenberry and Kuh

(2015) suggest that, even though they are oftentimes overlooked, students can provide helpful insights that benefit assessment-based educational evaluation. Furthermore, consideration of the uses and influences of student learning evidence would be artificially truncated if students were excluded from the sample since measures of student performance are created for the explicit purpose of being used by teachers and students. With regard to the notion of evaluation influence, consideration of the intersectional dimension of awareness and intention (Alkin & Taut, 2003) can be more fully understood by including program beneficiaries in the sample.

To select participants from the program beneficiary (student) stakeholder group, a stratified, purposeful sample (Miles & Huberman, 1994) was used. Stratification of the student sample was conducted along two dimensions: extent of participation in the program (current or previous students) and typical performance (high, average, or low relative to cohort group). These efforts were undertaken to ensure the participant sample included diversity on these characteristics. Current students explained their recent experiences with evaluative evidence while previous (post-program) students shared their experiences from a more distanced, retrospective lens. The sample was also designed to include a students of varying typical performance levels (i.e., students of high, average, or low typical performance based on their past performance data). A program administrator was responsible for selecting participants along this dimension, and this designation was hidden from all others involved in the study. Recruitment of students who had finished the program was based on the availability and willingness of the students to participate. For all groups, recruitment for participation was undertaken by an

administrator using scripted information. Recruitment protocols strongly emphasized that participation in the study was not required and that choice to participate would in no way affect students' relationships within the organization or their academic standing since program facilitators and administrators would not be made aware of their participation status.

Data Analysis and Quality

Merriam (2002) typifies case study according to the way in which the findings will be reported: descriptive, interpretive, or evaluative. This interpretive study was undertaken using an emergent design such that ongoing analysis took place between rounds of interviews and document review. This process of responsive data collection and analysis permitted me to address and check my developing notions of uses and influences, and their interconnections, across data collection methods and stakeholder groups. Following selective transcription of all interviews, I developed preliminary interpretive themes related to evaluation use and influence while using ATLAS.ti MAC (version 1.6.0) to read transcripts and compose memos, marginal notes, and annotations. In my second reading of transcripts, I compiled these themes into three categories based on dilemmas that emerged from stakeholders' descriptions of the challenges facing the use of evaluative information. Each dilemma consists of subthemes that provided a more nuanced perspective of specific dimensions of that dilemma. These dilemmas and subthemes were organized from preliminary themes through a process of "dancing the data" (Simons, 2009, p. 117) in which themes and quotes were printed to note cards to organize and make sense of the data through conceptual mapping.

Four primary strategies were utilized to ensure the trustworthiness and credibility (Lincoln & Guba, 1985) of data analysis: rich data, triangulation, respondent verification, and reflexive memoing. To facilitate trust and confidence in the findings on the part of stakeholders, the process of the study was made transparent, including the means of connecting the findings to the collective judgments of stakeholder groups. Richness of data considered data quality and corroborating sources of data (Maxwell, 2005). This in-depth study consisted of 30 interviews that sought diversity of perspectives across and within stakeholder groups. Multiple methods (document review and three sets of interviews) were utilized in attempts to capture the case through different approaches, to cross-check the relevance and significance of findings, and to overcome the shortcomings of any single data collection method. Methodological triangulation permitted the comparison of findings across document review and interviews, while data triangulation permitted the comparison of findings across data sources (Simons, 2009). Respondent verification was utilized for member checking purposes, typically within the interview (as well as after the interview for key administrators). Active, iterative engagement in dis-identification (i.e., observation of aspects of self without emotional entanglement Simons, 2009) was integrated throughout data analysis in effort to foster reflexivity (i.e., a critical examination of the assumptions underlying one's actions, the impact of one's actions, and how one constitutes reality and identity relationally; Cunliffe, 2004). Reflexivity was of importance given my pre-existing conceptions of clinical science education from past professional experiences and my pre-existing relationships with some of the program stakeholders from this program. Stratified sampling sought to ensure a variety of program

beneficiary interviews and an additional round of interviewing was added in response to initial findings to increase the quality of information gleaned from the interviews. Importantly, for the purposes of case study and this specific case, triangulation may not generate convergence; rather, the use of multiple methods supports the goal of understanding different perspectives, what factors impact those perspectives, and how those factors play out in the local context, permitting the possibility that divergence may be a more accurate and meaningful outcome of methodological and data triangulation (Simons, 2009).

Summary

A single instrumental case facilitated this in-depth study so as to generate a thick description and portrayal of how evaluative information affected people and situations connected with the CSE system for first- and second-year medical students at the research site. A constructivist approach and use of qualitative case study methods permitted an in-depth exploration of student learning evidence use and influence within the educational organization. The case was selected as an intensity case (Miles & Huberman, 1994), and data collection methods included document review and three distinct types of interviews. The participant sample included program administrators as well as students. In this case, program administrators are responsible for evaluative responsibilities that traditionally belong to instructional faculty. All program administrators related to the program were interviewed, and a stratified, purposeful sample (Miles & Huberman, 1994) was used to diversify the student sample along two

dimensions: extent of participation in the program (current or previous students) and typical performance (high, average, or low relative to cohort group).

An emergent design positioned findings from each phase to refine subsequent data collection and analyses through progressive focusing (Simons, 2009). Data for this study was collected through: an initial round of interviewing (24 interviews) to characterize the lived experience of evaluative information produced by the CSE system through stakeholder reflection (an indirect measure); a second round of interviewing (five interviews) framed as a reflective exercise to observe students' active interpretation of score reports (a more direct measure); a final follow-up interview with the administrator most directly responsible for the program (one interview); and a document review (11 documents) to investigate the quality frameworks and information use protocols that contributed to the design and purpose of the CSE system. After selective transcription of all interviews, interpretive themes were developed (first round of formal analysis) and then compiled into three categories based on dilemmas that emerged from stakeholders' descriptions of the challenges facing the use of evaluative information (second round of formal analysis). Layered analysis (Creswell, 1998) of program documentation and over 200 pages of interview transcripts permitted comparison of findings based on two distinctive interpretive lenses (i.e., one of evaluation use and one of evaluation influence). Finally, themes and subthemes related to each dilemma were organized and refined by "dancing the data" (Simons, 2009, p. 117). Data quality was fortified by using rich data, triangulation, respondent verification, and reflexive memoing.

CHAPTER IV

RESEARCH FINDINGS

The purpose of this chapter is to discuss the findings of this study regarding the use and influence of student learning evidence in assessment-based educational evaluation. Student learning evidence is a type of information produced through evaluative processes. The CSE system produces student learning evidence relative to a variety of standards and objectives that is used for accountability and learning purposes at the student, course, program, department, and institutional levels. As previously indicated, inquiry into the uses and influences of this evaluative information was guided by the following research questions:

Research Question 1. Based on a specific case of an innovative clinical skills examination system (CSE system) in a medical school, what is the nature of evaluation use and influence in assessment-based educational evaluation?

- 1.1 What types of use (instrumental, conceptual, symbolic, process, etc.) are made of student learning evidence within the educational organization?
- 1.2 To what extent are uses oriented toward accountability and learning?
- 1.3 How are these findings affected by expanding the notion of evaluation use to consider dimensions of evaluation influence?

Overall, three rounds of analysis of a rich dataset comprised of stakeholder interviews, reflective exercises, and document review of educational quality frameworks and

information use protocols suggest that the use of student learning evidence in assessment-based educational evaluation is subject to three ongoing dilemmas that influence the applicability of evaluative information.

As described in greater detail Chapter III, data were analyzed through rounds of sequential thematic analysis. In the first round of analysis, I identified seven preliminary themes: (a) the interpretation of evaluative information relative to a developmental continuum anchor, (b) the influence of personal factors on prioritization and operationalization of evaluative information, (c) the struggle to balance and consolidate meaning across sources of evaluative information, (d) the systemic complexity of stakeholders' perspectives and information needs in educational contexts, (e) the systemic complexity of inter-related educational activities within the program model (e.g., planning lessons, training preceptors, instructing learners, designing evaluations, and using evaluative information), (f) the need to attend to positions of power and influence affecting the use of evaluative information within the educational organization, and (g) the need to attend to the partitioning of, and limited access to, evaluative information that constrains the utility of evaluative information for learning. Subsequent thematic analysis of transcripts and documentation through two additional rounds of analysis (the last of which involved "dancing the data" to conceptually map themes; Simons, 2009, p. 117) led me to organize these seven interpretive themes around three main dilemmas.

These dilemmas result from higher order influences that originate beyond the local environment. Such higher order influences include: the nature of competency-based

or standards-based education, the directive of education to meet students' diverse learning needs, the multi-level nature of student learning evidence use for assessment across tiers of the educational infrastructure, and the tacit tension between meaning and values within the deconstruction of information. In her work on institutional ethnography, Smith (1987) describes such dilemmas as "problematics," or micro-level issues that manifest within the local setting due to more macro-level forces (e.g., institutional, political, and/or social). While problematics influence everyday experiences of people and situations, it must be recognized that problematics cannot be resolved within the local setting since they originate from external, higher order entities. Despite this limitation (of a constrained locus of control), attentiveness to these problematics by local evaluation practitioners and program stakeholders can inform the evaluative thinking underlying use of evaluative information for program improvement.

In this chapter, three problematics are used as an organizational framework for relating my findings about the use and influence of evaluative information in assessment-based educational evaluation. These problematics are summarized in Table 2. In the remainder of this chapter, each of the three problematics is introduced and elaborated using findings from document review and interviews. Themes are presented that illustrate some salient elements of the assessment context that seemed to contribute to the development of the problematics. Direct quotations from interviews of program stakeholders illustrate these problematics and are drawn from stakeholders' reflections on how evaluative information affected them and the learning environment. Findings from the reflective exercises, which are based on direct observation of students' interaction

with evaluative information, are presented separately to permit comparison between the indirect (interview-based reflections) and direct (reflective exercises) data sources. The chapter concludes with a summary of research findings.

Table 2

Problematics Facing Assessment-based Educational Evaluation

Facilitating Sensemaking Processes	Engaging Systemic Complexity	Attending to Power and Information Gaps
Developmental Anchor Personal Factors Balancing Competing Information Tiered Sensemaking Through Intentional Design	Inter-relationships Perspectives & Boundaries	Program Model Information Needs Evaluative Milestones

Problematic 1: Facilitating Sensemaking Processes

Student learning evidence does not speak for itself (Ikenberry & Kuh, 2015) but rather must be understood relative to considerations that are not necessarily indicated with a score or rating scheme. While the purposes for student learning evidence related to individual student performance are straightforward (e.g., whether the student should pass or fail an assignment, exam, or course), the use of student learning evidence for learning purposes, both at the student and program levels, involves more holistic consideration of its meaning and implications for action (Ikenberry & Kuh, 2015). In deference to the influence of values on meaning making, revelatory processes have been described in the assessment and evaluation literatures as a process of ‘sensemaking’ more than one of

interpretation. Use of the term sensemaking draws attention to the ways in which diverse participants contribute to a shared community and culture in which academic activities, including the use of student learning evidence, take place (Rickards, Abromeit, Mentkowski, & Mernitz, 2016). Indeed, as a professional field, evaluation can be considered "assisted sensemaking" (Julnes, 2012).

One who adopts a lens of sensemaking perceives decision making as a social construction that is influenced by the values of the individuals engaged in decision making, by social interactions between those individuals engaged in decision making, and by the organizational values shaping the decision context (Jonson et al., 2017). These influences are not easy to anticipate or address, causing the processes of obtaining and using credible evidence appropriately to be neither simple nor rational (Hutchings, Kinzie, & Kuh, 2015; Jonson et al., 2017). Stakeholders may disagree about what constitutes evidence (Jonson et al., 2017), a common occurrence given that faculty engaged in assessment may perceive its methodological quality as insufficient according to the standards of practice for the disciplinary research they conduct (Blaich & Wise, 2011). Furthermore, it may be challenging for research professionals to accept that the technical adequacy of assessment design is less important than ensuring that the evidence it produces is useful for decision making (Kinzie et al., 2015).

Within the case study, the need for sensemaking to address the use of evaluative evidence at the student- and program-level manifested around four main themes that indubitably affected use. First, the evaluative tools that constituted the CSE system were anchored along a developmental curriculum that influenced students' understanding of

their experiences and feedback with an additional influence on the ways in which students decided to respond to evaluative information. Evaluative information was perceived as a momentary snapshot of proficiency relative to established developmental milestones with inherent assumptions of growth and improvement over time to a point of minimal competency. For each developmental milestone, two key elements were provided: for the program model, specific standards and objectives clarified the required knowledge, skills, and mindsets that indicated minimal competency; and for the evaluative tools, this vision of minimal competency was translated into a rating scheme that indicated student performance relative to those standards using a single, composite measure of performance outcomes (e.g., multiple instruments produced scores that were combined using a weighting scheme and adjusted using cohort-level statistics to produce a single overall score for each student).

Second, personal factors seemed to affect how stakeholders made sense of evaluative information, and this subsequently affected the extent to which evaluative information was used. The influence of these personal factors was detected at four points in the evaluative process: immediately following the experience of the evaluation (process use), upon receipt of evaluative information, after full analysis of evaluative information, and longitudinally across multiple evaluations.

Third, use was affected by the ways in which stakeholders considered multiple sources of evaluative information. Stakeholders were required to balance and combine evidence for each evaluation across their experience and multiple sources of feedback, to consider evidence relative to cohort-level performance, and to address longitudinal

consolidation of evaluative information across courses and milestones. Fourth, since assessment-based educational evaluation is predicated on the use of evaluative information to inform both student learning and the improvement of teaching, themes in this chapter are described at the student and program levels. This is necessary in order for student learning evidence to be meaningful for both student learning and program improvement: the primary utility of information as providing an indication of *what students know and can do* must be addressed. As discussed in Chapter II, overlooking these elements facilitates separation between faculty members' responsibilities by taking these important design elements as given without meaningfully connecting them with higher order outcomes and information needs. By exploring use at the student and program levels, separation between accountability and learning purposes can be reduced. The first three subsections address the sensemaking experience at the student level (i.e., for student learning) while the final subsection explores the need for sensemaking at the program level (i.e., for program improvement).

A developmental continuum anchor. Inherent within all stakeholder dialogue about the use of evaluative experiences and feedback, the developmental continuum onto which the series of courses is built had a noticeable effect on sensemaking processes. This dynamic was evident in the formative and summative uses of evaluative information within each course: mid-course formative testing was systematically kept distinct from accountability decisions which resulted in differential use of formative and summative evaluations by stakeholders. The developmental anchor was also evidenced across courses: past, present, and future performance was anchored at specific evaluative

milestones along the developmental continuum. Former students of the program interacted with current students of the program, enlightening current students of the extent to which the training program set them up for future success. Taken altogether, aspects of the CSE system connected to the developmental continuum anchor served to push the interpretation of evaluative findings away from absolute measures of competence or achievement and toward a more malleable perception of accountability and learning. In these ways, the developmental continuum anchor grounded the program model and its evaluative tools along a spectrum that supported students from novice to increasingly more competent student physicians.

Formative and summative uses. Stakeholders made sense of evaluative information by comparing it not only with their past performance but with their expectations of future performance. Furthermore, the developmental continuum anchor underlying the CSE system connected program experiences and evaluations (within and across courses) with post-program expectations. As one administrator explained:

> We encourage students to realize that they are at an appropriate developmental level—to give them reassurance that, "It's okay, all first-year students struggle with this. You're going to be fine with this. You'll keep practicing and we're going to help you." So just being able to put their performance in context of the developmental level and our track record of being able to get people of that developmental level up to the developmental level they need to be in. I think that is an encouraging component of [the CSE system].

This sentiment was echoed by a student who offered a rationale for the relevance of evaluative feedback: "*Passing* is just a note that says, 'This is where you're supposed to

be right now based on what we expect of a first-year student, or a second-year student.' It just means that you are on your way."

At the course level, each evaluative milestone included both formative and summative evaluations. Evaluations were directly aligned so that students had a formative evaluation prior to each summative evaluation. The evaluations were conducted using the same evaluative tools (i.e., checklists and rubrics) and parameters (e.g., setting, timing, type of case, tasks, etc.). Since formative evaluation in no way influenced accountability decisions at the student or program level, its evaluative information was utilized solely for improvement and learning. For students who performed far below expectations during formative evaluation, administrators offered additional time and resources to help develop the students' abilities prior to the summative evaluation. Accordingly, from the student perspective, formative evaluations were seen as "just for me," as one student shared, "The mid-block tests don't affect our grades . . . every time we ask how this affects our grades they say, 'It doesn't, this is just for you. It's just to give you feedback and help you know how you can improve.'" Evaluative information generated for summative purposes was compared against minimal competency expectations to make decisions about each student. The minimal competency designation was determined by institutional policy to indicate that the student, on the basis of demonstrating the expected level of competency on prior learning, was developmentally prepared for future learning.

The influence of past performance. Students and administrators made sense of new evaluative information relative to a student's past evaluative performance.

Developmentally sequencing evaluations across courses created milestone events against which stakeholders could visualize student performance and reflect on student change over time. All administrators described how anchoring performance along this developmental progression is an important element of the CSE system: though student physicians typically begin the program with variable levels of prior experience working with patients and little to no experience filling the physician role in a patient encounter, all American medical students are required to take a standardized performance assessment at the end of their third year of study in which they must conduct 10 consecutive patient visits. The developmental continuum anchor on which the program model and its evaluative tools were structured also oriented students to the relevance of performance expectations at various points in time, both within and across courses. Indeed, all students shared a personal story of growth over time using evaluative milestones as indicators of their past performance and overall growth. Many students highlighted how the CSE system created a safe space for their development. As indicated in one student's remarks, anchoring sensemaking along a developmental continuum encouraged a growth mindset by emphasizing growth over absolute achievement: "I find feedback very helpful all of the time. I look for things to improve because I know that I can improve . . . seeing feedback helps me know exactly what I can improve on. It is helpful for future experiences." Furthermore, most students described how the CSE system enabled them to appreciate their growth over the course of the entire program. As one student reflected, though some early learning goals may seem trivial in hindsight, they were essential at earlier points along the developmental continuum:

> That first week of medical school we were all so nervous when all we had to do was introduce ourselves [to the patients]. In hindsight, that seems so silly, but at the time, it was everything. Even though it seems so silly now, it definitely wasn't at some point, so it was pretty necessary.

Since students developed competencies incrementally through time across related courses using related evaluations, all present and future clinical work with patients was influenced by the students' experiences with the CSE system.

The influence of present (concurrent performance). In the present moment, students related their experience of the CSE system with concurrent clinical experiences (i.e., shadowing physicians, volunteering in free clinic settings, practicing under physician supervision, etc.). Students' experiences within the CSE system and the evaluative findings it generated helped them to apply program-based learning to real-world contexts of practice. Many students described feeling much more comfortable working with real patients as a result of the CSE system since they are able to call on these experiences when they are asked to perform in real clinical settings. As one student expounded:

> When I'm doing [clinical work] in the emergency department and my physician says, "Do you want to do an abdominal exam in room 22?" and I'm like, "Yeah, absolutely." I step away and think, "Okay, abdominal exam, what do I need to do for this?" The more patients you see, the more you'll get comfortable rolling with it. The more students know that, and the earlier and the more they are reminded of it, the better off they will be.

Students combined learning across these concurrent experiences, transferred skills practiced in the CSE system in real clinical settings, and felt more comfortable performing these tasks due to the CSE system. As one student explained: "This has

helped me get a lot better. I can go out into the clinics and shadow physicians feeling very comfortable taking histories and doing physicals there and that's because of the [CSE system]." Another student emphasized how the CSE system prepares students for practice in real settings:

> I feel [the CSE system] is really important . . . I've talked to friends at other institutions and some of them do not get the exposure we do right off the bat. Sometimes they "get thrown to the wolves" by seeing real patients without having the experiences that we do.

Real-world clinical experiences influenced students' perceptions of the CSE system over time by confirming its relevance to their future practice. As one student reflected:

> I think it's okay to say outright that these experiences will help you and you will have patients who are just like these standardized patients. I really saw that more in my [concurrent clinical] experiences—it started to click there and I saw more merit in the [CSE system]. In the beginning, everyone says, "This isn't real at all," but I've definitely had encounters exactly like standardized patients.

Corroboration of the CSE system with real experience was important since, generally speaking, students are not well-positioned to understand how real patient visits compare with simulated ones or which types of learning goals will be prepare them for future professional work.

The influence of expected future performance. In addition to reflection on past performance and confirmation with concurrent experiences, the developmental continuum anchor supported the extrapolation of evaluative information to future performance. This was a common theme across most interviews, suggesting that, even if evaluative information is not designed (or validated) to predict future performance,

stakeholders used it as a 'loose predictor' of future success in real-world clinical settings. This perception was conveyed subtly: while many stakeholders recognized that evaluative findings did not generate context-independent endorsements of ability, most interviews indicated that stakeholders expected evaluation results to provide an indication of future performance in three specific contexts: on the STEP 2 CS exam, in future training programs (residency, fellowships), and in the student's future career. This current-future performance tension was evidence in stakeholders' ideas about the implications of evaluative information. One student framed the CSE system as an integral element of job training and emphasized how completing the training program (i.e., going through the experience of the program and its integrated evaluations) should ensure future success:

> Coming into medical school from working a few years, I see all of medical school as job training . . . You get back into the student mindset but I try to keep reminding myself and separate myself from my grades. It doesn't really matter. I want to be the best physician and be the best for my patient. This is 100% job training. So I see [the CSE system] as just that. A snapshot of where you stand right now. That's how I make preparations. I view practice sessions as the real thing too. It's just like preparing for any type of job you have. It's not that you have to know this for a few weeks then take the test and you're good to go. It's not like that anymore . . . [in] clinical specifically, I view [my learning] as the processes and tools that will make you a good professional.

By contrast, many stakeholders emphasized that evaluative information should be used merely as a reflection of a student's current ability with a specific patient case and encounter. In this sense, evaluative information provided a single 'snapshot' in time of student ability. One student stressed the need to integrate evaluative information into a broader consideration of one's ability:

> I think everyone uses the information pretty directly, maybe more so some than others—some use it too literally. With more and more patient encounters people realize that it is just a snapshot. You take the feedback. You employ the feedback.

Another student highlighted the developmental aspect of the training program in discussing the relevance of evaluative information as a snapshot in time:

> So I take [the feedback] and say, "Okay, that's awesome, but that was just a snapshot in time." I'll be doing these for the rest of my life, day in and day out, and each will vary so I see this as a benchmark.

The notion of evaluative feedback as a snapshot was also expressed by administrators, one of whom explained that strong performance merely suggest that a student can conduct a certain type of patient encounter at the appropriate developmental level without providing a broader endorsement of the student's ability. Another administrator described the CSE system as providing evidence of student performance at a given milestone along the developmental continuum anchor, emphasizing that the student must undergo the entirety of the training experience to be prepared for future clinical experiences. As another administrator explained,

> The scores are used to identify students who have deficiencies in their knowledge base and clinical skills to provide—whether it is formal remediation nor just feedback—and ultimately with an aim for getting them to be as ready to be a clinician as possible. I don't necessarily see the scores as having a specific ideal upon which to say yes you can move to third year or no you can't move to third year. It is a program that at its core if it is delivered and if a student is undergoing all of the aspects of it, we would argue that student would be ready to start the third year of medical school.

Though students largely seemed aware of the limitations of evaluative information in their descriptions of needing to avoid feeling satisfied and to keep their clinical skills "honed" and "fresh," the developmental continuum anchor onto which their experiences are mapped implicitly suggests that strong evaluations should lead to strong performance in future clinical practice. Students expected favorable evaluations indicate that they "are on track for success," that they are "prepared for the next stage of training," that they "have a strong baseline of skills," that they are "where they need to be now to be successful in the future," and that they can pinpoint "where exactly we are as student physicians." This aspirational expectation for evaluative information was evident in two student reflections. One student expressed wanting the CSE system to confirm that the student would be prepared for working in clinical settings:

> I hope these scores somehow reflect our abilities to perform in the clinic or in a hospital setting since ultimately that what we have to do. We want to know if we can do that or not. But I don't know how well these do measure that. That's what I hope these scores somehow determine whether—okay this score means the student passed the block which means that person should be ready for the hospital. I hope that is how they are using it. (Tracking your progress toward an established vision of preparation?) Yes exactly.

A second student emphasized this same notion relative to competency in relating to patients:

> Ultimately your end of course exam matters because you want to pass and . . . as strange as it is, you know your real patients won't grade you but they will see people similar to you at a similar stage of medical education and you don't want to be the one that they are like, "Oh that girl, she shouldn't come back in here." You want your patients to like you, to have a good reaction with you when you're in there with the team. So [the evaluation] gives you a good gauge. If my standardized patients are giving me a 90% then my real patients might like me

> too. It's all a foreshadow, or at least you hope it is, for when you're alone and expected to know things on your own.

Despite this tension in viewing evaluative information as concurrently reflective of demonstrated learning and future performance, stakeholders recognized the limitation of the CSE system in predicting future performance. One administrator described the artificiality of the system:

> I don't know if [the CSE system] is necessarily going to set [students] up for success in [the clinical] environment. I feel like real-world, real people situations are the best way of learning . . . so this is the closest we can get to mirror what they will do in the real world.

A student emphasized that the relevance of the evaluative information was not in the score itself, but rather the expectations that scores are perceived to represent:

> If you are basing your entire experience on these scores, I'm not sure that transfers to the wards. But if you are trying your best to make sure you meet expectations, I think that would be good enough for everyone and everyone would succeed.

Though the CSE system is designed to be as true to reality as possible, it may necessarily fall short of providing an experience that directly transfers to future clinical practice. As one student stressed, all forms of measurement and evaluation invariably measure more than just the construct of interest:

> I think it is probably pretty easy to misinterpret scores because it is an objective way of grading oneself but it doesn't paint the whole picture. Getting an 80% on something doesn't necessarily mean you aren't good at something, just that you missed it at that time. Misinterpretation of your own score could lead to you

> feeling really bad or inflate your ego if you got lucky that what you studied the night before was on the test.

The tension between using evaluative information to provide an indication of student learning mastery or of anticipated future performance highlights an important consideration for designing programs and their evaluative tools: though instruments used for evaluation might not be validated for predicting future performance, as administrators in this case argue, the program model itself may be designed to ensure students are adequately prepared for future performance. The temporal dimension of evaluation in assessment-based educational evaluation requires consideration of evaluative information relative to the expectations and implications of past, concurrent, and future performance.

Personal factors. Many students described the effects of subjective influences on their interpretations and uses of evaluative information, both in response to the testing experience (process use) and score reports (findings use). Subjective responses differed based on personal factors, including age, previous professional experience, and typical performance. Students also identified that past experiences and personal values influenced how they made sense of evaluative information. One student described how past experiences and one's background shape perceptions of evaluative information:

> Personalities and past experiences shape how you interpret a score. My background and how I grew up is not focused on getting the highest score but on making small improvements. [In my former career], too, and that's how you look at problems. You don't jump in right away to making something great and big—you just figure out what small things you can improve to get a little better. So if I get a 15/20, it's not the end of the world, it's like, "Okay, let me work on getting 1 extra point or 2 extra points the next time." I don't really talk about scores with others too often but I can imagine that some people would not be comfortable

> receiving a score that low and would start to worry, so I think it depends on our past experiences.

Another student emphasized how having past professional experiences shapes perceptions:

> I know from having conversations with students in our class that there is a good-sized group of students fresh out of undergrad who don't know any other way than: study for test, take the test, study for next test. I think that factors into their perspective on being assessed. [The CSE system is] obviously a different type of assessment than we've ever had before, but I think they are more to the side of "you've got to get these checkboxes marked off" rather than getting the skill correct . . . It would be interesting to see the trends based on how long you've been out of school versus coming back to school.

Differences in perception and sensemaking manifested at four different phases in the evaluation process: during the evaluative experience, upon receipt of evaluative information, after reviewing evaluative information in its entirety, and across courses longitudinally.

During the evaluative experience. Students emphasized how the experience of the evaluation itself shaped their perceptions. Many students described being unable to quell anxious feelings in the moments leading into an evaluation. One student underscored how anxiety faded as the evaluations begin:

> Everyone wants to get the skill correct but I try not to think about the [evaluative checklists and rubrics] when I'm in the suite itself. Everything changes. All the fear you have . . . I always get butterflies. I get nervous, really for no reason. As soon as you open the door, it all fades. When I see the patient, I get comfortable, then I let myself subconsciously check the boxes while I focus on working on the skills, not just going through the motions.

Another student reiterated the same experience of pre-evaluation stress:

> I have this [Fitbit] that tracks your heart rate. You can look at it. My heart rate would spike high right before—as I'm opening the door—then after that, I was fine. There's so much going on in your head—all the things I need to cover, all of these things I have to remember like telling the patient that you are a medical student. So right when you open the door, you're very aware of all the things you need to do and you know that in two years these patients will actually be sick and you'll have to know what to do and figure out the diagnosis. So up until the moment of opening the door, it was stressful, then, as soon as I walked in, it was fine and it felt fine but it was definitely an interesting phenomenon. Even as you tell yourself that you've learned all of these things, you've practiced this . . . it is really hard not to feel stressed out about it. There's not a great way around that . . . I have not had a good history of being able to think of this as "just a test."

Completing the evaluation experience conferred a short-lived sense of relief for students. They described engaging in justificatory reflection immediately following the encounter as they "vented" with their peers. These processes seemed to lead students to differential feelings about their evaluation experience based on the individual student, their past experiences, their expectations, their values, and their perception of the performance. One student described connecting with peers following evaluations to commiserate and process an emotional response: "I have discussed scores with other students in my class, primarily after exams. We might be venting with each other more than analyzing the [implications of the] scores." Many students discussed their need to touch base with peers on the experience as a "pulse point" for framing their own performance, conversations in which they asked each other questions about feelings and perceptions of fairness.

In reconciling one's experience with how one anticipated being evaluated, students prioritized various elements based on personal factors. One student explained

routinely prioritizing relating with the patient over attempting to complete the evaluative checklist:

> I was always personally concerned with the patient conversation and making it feel like a conversation, making it feel very comfortable . . . There were times when I would probably forget to ask the patient small things like, what reaction they had to a drug allergy, that kind of thing. In my mind I was thinking, if I hit on everything [on the evaluative checklist] then that's great, but I was more concerned about sort of the interaction.

Another student discussed prioritizing certain evaluative elements based on past and future evaluations as well as based on the amount of effort expended in preparation for the evaluation:

> If I practice the interview skills [in a formative evaluation] and do them well and the next [formative evaluation] adds a physical exam [to the part I already practiced] . . . if I then don't do so well on the interview skills, I'm not as focused on [the interview] because I probably spent more time trying to work on the physical exam. Since the physical exam part would be the new part, I'd pay more attention to that because that is where I was putting in the effort. I'd be hoping to get a better score on that so that's what I'd be taking into consideration.

Students feelings about the evaluation and decisions about prioritizing aspects of evaluative information thus varied based on individual-level decisions. As a conceptual process use of evaluation, mid-evaluation feelings and immediate reflection on performance created a baseline against which future use of information was compared. As one student recounted, feelings about past performance caused students to try harder in future evaluations:

> There were certain times I didn't perform as well on the exam. I felt that way coming out [of the evaluation], too, so [the evaluation] was pretty accurate. I

> could use that to find ways to improve in the future and trying to apply that for future exams to make sure I got a higher score and could improve on the things I hadn't done well in the past.

For students, the experience of the evaluation served to confirm or refute the extent to which students trusted the reported feedback. One student offered that the amount of time between an evaluation and receiving results influenced the student's degree of receptivity:

> I think the time delay makes it easier for you to get defensive or to blow it off. Instead of thinking, "Alright, I'll deal with that whenever I have time in the future," an immediate turnaround forces you address it right there, which I personally like.

Many students also described a preference for immediate, face-to-face feedback, an element of the CSE system that is incorporated into some formative evaluations. Evaluative information was considered more trustworthy when reported in person; delayed reporting caused students to question the accuracy and appropriateness of evaluative information with which they disagreed.

Receiving evaluative information. Personal factors influenced student perceptions of evaluative information upon receipt of evaluation reports. These emotional responses seemed to directly relate to accountability aspects of the CSE system, since most students responded differently to formative and summative evaluations. As one student explained,

> The end of block is all that matters. That last day that you have your patient encounter and then you [take the other clinical exams] and those are all a percentage: add them up and that's your grade. I've found that to be fairly

> representative of how I've been doing if not better because I've had time to improve in the weeks leading up to the exam.

Importantly, the CSE system is based on demonstrating minimal competency by the conclusion of the course, so formative testing does not influence decisions regarding promotion and retention. Formative testing experiences are thus positioned squarely on learning purposes, while summative testing experiences address both accountability and learning. Just as students' emotions influenced their conceptual process use of evaluative information, feelings and personal factors influenced students' experiences analyzing the evaluative information, as shown in the varying reflections of the following three students. One student explained how making sense of scores was anxiety-ridden if scores were anything less than perfect:

> I would open [the score report] and see the total number—whatever out of 50—and regardless of whatever it was, 49 or 0, I would panic. What did I miss points on? Then seeing our rubrics in class or meeting with [a program administrator] was a way to focus on what you are working on. The bulk numbers are stressful. They don't tell me a lot other than you are not perfect which not in the moment should be expected and isn't a bad thing, but it doesn't do a lot for me until there is a breakdown. Then you can be like, "Oh shoot, did I forget to check reflexes?" So next time you practice you have a checklist of what you normally do and the one I've bolded in my head because I keep getting marked down on it and need to keep practice doing so I don't forget it next time. I look at it more as an all or nothing interaction: either I got them or I didn't. If I didn't, I'd been made, and that was something I had to focus on.

By contrast, another student described how a pattern of performance was more important than the quantitative representation of the evaluation:

> As long as the score starts with a 7 or above, I'm fine with it. I say, "Okay, that's fine." My scores have tended to be fairly high so I usually see high 80s or 90s and

> I think, "Okay, I did well, great. I'll keep doing that." One time I got a 74% on something and I thought, "Oh man, what did I do?" because it was so abnormal. So I went into [the feedback] and I realized that I skipped an entire part of the checklist for the interview, so I though, "Yeah, that's fair." For me I know it is not totally objective, the way that they grade us, but I take it as an objective fact, like yeah, you did do 74% of this.

The last student emphasized that emotional responses amplified differences in evaluative feedback, possibly beyond what might be useful for learning or even accountability purposes:

> [How I interpret scores] is mostly just a personality thing. Since I've grown up in very competitive environments in high school, college, and here in medical school, I do tend to be driven by numbers. So they might be meaningful to me in the sense of I got an 83%, is that good or bad? Okay, I got a 92% that's a lot better. The real difference between those two scores might just be something you blanked out on that day and it wasn't such a big deal. I think that was my main takeaway from that—that I may have drawn meaning from numbers in an emotional sense because they appear a certain way but when you actually think about it there probably isn't all that much of a difference.

Students thus seemed to make sense of evaluative information in ways that were impacted by past experience, expectations of performance, and perceptions of amount of time and effort needed to improve on formative evaluative information prior to summative evaluation.

Evaluative information from the CSE system was maintained in a separate software repository that required students to log into the software platform to access evaluative information. Students' decision to access this information seemed to be based on two important factors: the level of concern they felt about their overall score (which was connected to their expectations based on the experience itself) and the requirement of

having to be on-site to access the software platform. To understand the influence of these factors, it is necessary to consider how scores were reported. For formative evaluations, evaluative information was accessed directly through the software platform: students logged-on and had immediate access to all evaluative information (e.g., evaluative checklists and rubrics). For summative evaluations, however, students received a separate report that included the student's overall score, sub-scores on each component of the CSE system (3-7 scores), the weighting scheme for the sub-scores, and cohort-level performance statistics. All scores were reported as percentages. This report was produced by a centralized reporting system through which evaluative outcomes from other "departments" in the medical program were also reported. As such, the student received a synopsis of their evaluative information: if the student was satisfied with this synopsis information and/or felt unsurprised by the quantitative representation of their performance, the student might have stopped their analysis of evaluative information without ever accessing the more detailed feedback that was available in the repository. Thus accountability-based reporting at the student-level curtailed the use of evaluative information for learning purposes when such analysis was not required.

The majority of students interviewed indicated that their use of evaluative information was critically affected by these considerations of access to evaluative information. Since passing the course by achieving the minimal cut score is "all that matters," some students felt no need to analyze evaluative information for learning purposes. The limited access to evaluative information created a situation in which a student's concern about the score (either in its magnitude or discrepancy with

expectations) had to be severe enough either to prompt them to travel to campus to investigate the more detailed score reports or, alternatively, to make time for such analysis during their next regularly scheduled trip to campus (which fell at the start of new courses, when analyzing score reports may seem less urgent or important). While a delay of a few days may seem inconsequential, as one student shared, once students knew they received a passing score, they perceived evaluative information differently:

> When you know you passed but then you open the feedback and read the comments, you're like, "Why are you saying this? I am clearly doing okay." Once you are proficient enough, you think, "That's alright. It's fine." You have to self-motivate to take that next step. You have to get past that "It's fine, I'm proficient, quit judging me" mentality.

Taken together, the factors impacting this access-timing dynamic created a missed opportunity for learning from evaluative information. This missed opportunity is relevant and important given that the CSE system is structured around a developmental continuum so that competency at each milestone will be re-assessed in future tests or clinical experiences.

After reviewing evaluative information. Review of evaluative information in its entirety seemed to be affected by the amount of time a student had participated in the program, by students' perceptions of the fairness of the CSE system, and by the severity of the information. Early in the program, students used evaluative information at greater rates for two identified reasons: the newness of the CSE system required students to spend time becoming familiar with it, and students were required to submit assignments in which they reviewed their performance. First year students analyzed multiple sources

of evidence about each formative evaluation independently before they submitted these assignments and discussed their outcomes in small groups of faculty and peers. Use of evaluative information tapered off in later courses when students were not required to submit such assignments or engage in such dialogues. From discussions with administrators, it seemed to be assumed that students would continue to use the evaluations to improve their skills through personal motivation and independent practice.

In addition to the influence of time on evaluation use, increased familiarity with the CSE system prompted students to grapple with the issues of fairness related to the CSE system. Almost all students described negotiating their interpretation of elements of the evaluation that they felt were more objective (such as completing certain tasks or conducting physical examination maneuvers correctly) or more subjective (such as building rapport with patients or responding to patients' emotions appropriately). Expectantly, students often externalized poor marks if they felt the rating was subjective, undercutting the utility of this evaluative information for learning purposes. It seemed common for students to utilize the robust evaluative information, which consisted of multiple sources of evidence, to justify the inaccuracy of their evaluations. Encouragingly, all students expressed that they not only knew how to dispute evaluations but felt comfortable doing so. Students stated that they only initiated these responses if discrepancies in evaluation were egregious enough, a threshold that varied based on students' personal factors. Though accurate and appropriate evaluations are ideal, all human rating systems will necessarily involve some degree of error. While it is somewhat encouraging that students accepted these small areas of discrepancy in their evaluations,

this dynamic can be troublesome: students are not always capable of identifying when "missed points" represent having forgotten to complete an element versus performing the element incorrectly. Indeed, students often describing missing points for a skill they did, attributing the missed points to grading errors though the missed marks might have represented areas that needed improvement.

When reviewing evaluative information in its entirety, the severity of the feedback influenced the extent to which students could mobilize it for future learning: if severe, students expressed a long-lasting influence while, if inconsequential to the overall score, students failed to make considerable enough effort to improve. Two students offered descriptions of each end of this spectrum. The first student explained that feedback was long-lasting and had a large impact on future performance:

> I remember a few specific things that I missed and I never missed those again. It was basic things that the first time doing it and being nervous, not really thinking . . . For a chief complaint of shortness of breath, I didn't listen to the heart and lungs, so I remember that encounter to this day. It definitely sticks with you.

By contrast, another student explained that if the information was perceived as insignificant in nature it had minimal impact on future performance:

> The only thing I could say about [using scores for improvement] is that if you fail you have to redo it. Other than that, I don't think the grading provides that much feedback. Personally, I'd see a list of those nitpicky things and I'd say, "Oh, I'm never going to miss that again"—but I probably did.

Overall, the students' experience with the CSE system over time, their perceptions of fairness, and the severity of information they received seemed to influence how they perceived the relevance of evaluative information.

Across courses longitudinally. Finally, factors related to personal identity also appeared to influence students' use of evaluative information across developmental milestones and courses. All students spoke of how the evaluation experience and information helped them feel comfortable assuming the role of a physician and interacting with patients in clinical encounters, with most students crediting the CSE system for their feeling confidence about making the transition to the clinical setting of clerkship training in the third year of medical school. Students described the development of their professional demeanor, their improved ability to ask for targeted help from instructors, and their readiness to undertake the standardized test. As one student offered:

> The biggest thing for me at least is that I like to think of myself as a fairly outgoing person and competent person but I remember my first [experience in the CSE system] . . . it was the most stressful moment in my entire life. I was so incredibly nervous just to say my name and ask how the patient was doing. It was so far out of my comfort zone, more than I expected, so it was so important for me to just get comfortable going into the [patient room] . . . just being very confident in that role because I am going to be younger than a lot of my patients . . . asking them to move the way I need whether it causes them pain or not, really just being comfortable getting the information that I need. It is easy to shy away from things but that does a disservice to yourself and to your patient because you could be missing something when you're not being complete [in your approach].

Together, the experiences and feedback associated with the CSE system provided instrumental support to student physicians as they learned to engage patients appropriately and to think critically within a time-compressed clinical encounter.

Many students addressed their typical pattern of performance as a factor influencing their use of evaluative information. For some students, feedback tended to be mostly positive and fairly consistent from one testing encounter to the next, leading the student to find the evaluation less useful than students who received targeted and meaningful areas for development. These higher performing students expressed a desire for more guidance or feedback on areas for extended development; however, since the CSE system is designed to ensure minimal competency, these information needs were not necessarily addressed. Other students had experiences in which they performed differentially across evaluations. While longitudinal patterns of typical performance influenced how students made meaning of evaluative information, patterns of performance were tricky for students to navigate: if student performance was stable over time, students tended to externalize any variation across evaluations as relating misevaluation, an emotional response that could hinder students' learning if the variation was due to the student's performance.

This phenomenon occurred at the student level as well as for cohorts of students. As one administrator reflected, cohorts of students respond differentially to particularly challenging patient encounters based on personality:

> [For a particularly challenging patient encounter], the total score on [the patient medical record assignment] was 20 points and the lowest score was 6, so it was a big wake-up call for several students. What was interesting for this particular class, during [a focus group], they said, "Thank you so much for taking the time to give us this feedback because it opened our eyes and gave us an idea of what a huge diagnosis we missed because we jumped to a conclusion [about our patient] and we failed to ask some very important interview questions." And they were extremely appreciative of having the deficiency revealed to them and discussed in detail so that they felt more comfortable moving forward. The cohort of students

> from the prior year, however, based on their personalities, did not like getting such negative feedback, and they argued whether it was very common to see this [type of patient presentation] and raised other concerns about fairness. I've found that the flavor and the personality of the individual students [in a cohort] and [focus group] representatives causes feedback from students to vary year to year even if the curriculum is identical.

Furthermore, despite the explicit focus of the CSE system on standards-based improvement over time, many students still internalized their performances as an indication of the "type of student I am" instead of as a judgment of their current demonstrated competency relative to the individual abilities necessary for a student physician. Many students identified themselves as "strong" or "mediocre" students, indicating they affixed evaluative information to their self-identity as opposed to their burgeoning skillset. This suggested a missed opportunity for the appreciative use of evaluative information that results when students engage with evaluation findings with openness and positivity. This notion was implicit in many statements made by students in their reflections that differentially centered on improving evaluative performance or on improving their learning for future patient encounters. Internalization of evaluative information in this way supports an "entity view of intelligence" that perceives intelligence as a fixed, internal characteristic of one's identity (Dweck, Chiu, & Hong, 1995). By contrast, the developmental continuum anchor and learning purpose of the CSE system support an "incremental view of intelligence" (Dweck et al., 1995), a perspective of intelligence as malleable and affected by effort. An entity perspective of learning parallels accountability-oriented mindsets regarding the use of evaluative

information in education, while an incremental perspective views the same information as useful and action-oriented.

While these personal factors were somewhat unique to this specific performance assessment system, student-level emotional responses to testing experiences and score reports are, to some extent, typical representations of the influence of students' feelings on their reception and use of evaluative information. Given the nature of the CSE system as broad preparation for the type of interactions in which students can expect to engage multiple times per day in their future careers, these emotional responses may be considered developmentally appropriate. As students transition from novice to increasingly competent student physicians, they need to address feelings of nervousness and anxiety that might impede their ability to perform in clinical settings. Accordingly, the CSE system provided a safe space in which students could engage and overcome these developmentally appropriate feelings.

Balancing competing information. The CSE system provided students with practical experience and extensive evaluative information on their performance. As previously discussed, students needed to balance their experiential perceptions with their evaluative feedback. Additionally, students described challenges around three key areas: around receiving conflicting messages across raters of a single encounter, around considering their performance relative to the performance of peers, and around determining how to prioritize areas for improvement.

Conflicting messages within a single encounter. First, since students received extensive feedback for each evaluation, students experienced challenges related to

balancing competing information within a single evaluation. Though these tensions manifested in the reporting of evaluative information, they originated from discrepancies in instruction or previous evaluations or from varying rater identities. Discrepancies between instruction and evaluation created challenges for students' sensemaking processes. Students were sometimes taught to do skills in ways that conflicted with the evaluative expectations (anchored in their current developmental level), positioning feedback that corrected those discrepancies as frustrating and poorly received by students. Likely due to perceptions of relevance and long-term value, students deferred to the instruction they received from veteran physicians instead of adhering to the expectations identified for their developmental level. As one student clarified:

> Where we usually end up with the greatest degree of variability is that for the workshop sessions there will be local physicians—emergency room and other various medical specialties from the hospital and different parts of town—to teach us whatever maneuver we'll be learning at that time. They'll all have a slightly different way of doing it for the benefit of 5, 10, 15, 20 years of experience in which they've found that this is a little better, like this angle is a little better—so depending on who you work with for that first bit of hands-on training, you will adapt the way you do it in the [evaluation] that first time. I would argue that presumably most of the class has had it happen at least once when you learned to do it one way in the instructional session and then you get corrected on it or docked a few points here or there—not the difference between passing and failing but you lose points for improper technique. They say, "That's not what we covered in class"—well, it depends on which class you're referencing.

Students were not well-positioned to understand when deviations from standardized expectations was appropriate. As one administrator explained: "Physicians become able to take shortcuts. Medical students do not have that privilege because they are just learning. They haven't seen or practiced enough." As such, variable instruction by

seasoned physicians and eagerness by students to do a skill the "best way" was sometimes prioritized over performing the skill at the appropriate developmental level, undercutting the utility of evaluative information for learning and improvement.

While students responded amenably to variation from standardized expectations originating from physicians, they did not always accept deviations from other types of raters. This was particularly true for specific elements of the training program (physical examination maneuvers), as one student offered:

> If a standardized patient tells you a specific way you should do this—"You moved my arm to the left when it should have been moved to the right."—I see that feedback like, okay, that's cool, but when it comes down to it I'm going to do it the way that I think works.

Students also struggled when raters' evaluations seemed to contradict one another, as one student described:

> I can recall from my own experience . . . I remember a physician saying that they preferred having the patient lying down [for a specific physical exam maneuver] so then you don't have to worry about looking up the patient's gown . . . When I went into [my formative exam], I did that maneuver with the patient lying down and during evaluation the standardized patient said, "It is a lot easier when the patient is sitting up, so that's how we usually do that. It's also a time saver. In the future, it would be better if you do it that way." I said I was doing it this way because that's the way I was taught, it makes no difference to me . . . Then a couple of times later, when I've incorporated that skill into a full exam, the standardized patient came back and said, "Well a lot of people are doing this maneuver laying down . . ." and I think, "Well, the last time I tried that, I was told I couldn't do it that way." So it's confusing. It's never the same person giving you this feedback. I think it would be much more frustrating if it came from one person.

By contrast, some students placed more emphasis on feedback from standardized patient raters than feedback from physicians, though this occurred typically with regard to other elements of the training program (i.e., interpersonal and communication skills). Students suggested that knowing the identity of the rater influenced how they used evaluative information. Students received feedback from standardized patients and faculty with some students deferring to standardized patient feedback and other students deferring to faculty feedback. For example, one student explained why physician feedback was prioritized:

> Mostly the comments that the physicians give—that's what I put the most credence in because I knew those evaluators were physicians and ones we'd worked with in clinical. After a certain point, based on the tone of the comments, you could guess who wrote them. I would take somebody's comments in a certain way depending on who that clinician was and how I heard them saying that in my head.

This dynamic was amplified when the instructors and standardized patients who rated student performance on summative evaluations were different people than the instructors and standardized patients who rated formative evaluations earlier in the course, a practice undertaken to expose students to a variety of standardized patients and to reduce the academic burden on individual physicians.

Comparison with peers. Second, the majority of students expressed a need to understand their personalized evaluative feedback relative to their peers and greater cohort. Perhaps due to the competitive nature of professional schools that have intense admissions processes, students suggested that they were not always capable of understanding the meaning of their scores until they were able to put it in the context of

peers' performance. As discussed in a previous section, many students discussed their need to touch base with peers on the experience as a "pulse point" for understanding their own performance, conversations in which they asked each other questions about feelings and perceptions of fairness. Two students described that, in situations where patient encounters were particularly challenging or evaluations were particularly low, students were more receptive to poor outcomes if peers had similarly poor outcomes. As the first student described:

> We don't talk about numerical scores so much as everyone saying, "Are we good? Thumbs up or thumbs down?" If we passed, then we don't talk about it, or if we didn't pass, we still don't talk about it. When you review things in class, it's impossible not know how people did: you see everyone's reaction around you and you can't not feel the vibe . . . Even though we have a great environment of, "Oh you didn't do well? Let's all help you do better." I don't think anyone would look poorly on somebody if they didn't do well, but you don't want to be that person because then you have to accept being that person and that's never easy or fun.

A second student emphasized non-verbal communication about evaluation experiences:

> We don't really talk about scores. We'll come out and say, "Oh, I don't feel good about that" or "Oh, that went well." We would all come out and look at each other's faces and see that recognition and say, "Okay that happened to you, too."

Upon receipt of lower-than-outcomes scores, students expressed feeling unable to process the meaning of scores until they compared their experiences and outcomes with other students. For a few students, this resulted in a mutual venting process in which validation of the student's perceived frustrations with fairness permitted the student to overcome emotional responses and determine next steps. Comparison of evaluation experiences and feedback seemed to occur naturally on a student-to-student level.

Comparison between students also occurred when evaluative feedback was compared with cohort-level descriptive statistics. This level of interpretation permitted cohort-level performance to influence how students perceived feedback, as one student reflected:

> I've very type A. I'm a perfectionist. I don't demand perfection from others but from myself. I am my biggest critic which I'm sure you hear a lot when you interview medical students. When I see a score, I initially either am happy with it or not. It is truly dependent on the class average. I want to know where I fall. I like to be at the top because I am a type A medical student. If I am not in that top range, I'm very disappointed.

Cohort-level trends also impacted perceptions of fairness, particularly when most students had lower-than-normal outcomes. One student seemed to expect that the CSE system was flawed if students did not perform well. Such a mindset (i.e., connecting the perceived quality of the CSE system to group-level performance) could be problematic in some instances, specifically when evaluations are designed to be particularly challenging for students at a given developmental level. A formative testing encounter designed around cultural competence, for example, was designed to provide a meaningful learning opportunity by illustrating how cultural biases may impact patient encounters and result in inaccurate differential diagnoses. While lower cohort-level performance in this context is indicative of the relative rigor of the evaluation, some students perceived the outcomes as evidence of unclear expectations and failure to provide adequate opportunity to learn.

Additionally, a few students described feeling comfortable with lower scores from this CSE system than they would accept from other types of tests, an interesting phenomenon likely related to performance assessment. Such a mindset suggests that

students may perceive that perfect scores on performance assessment are not always feasible, a factor influencing how they make meaning of scores. As one administrator highlighted, there is an implied precision in reporting evaluative information quantitatively that suggests differences represent important variation when, in reality, a difference of a couple percentage points does not represent meaningful variation.

Prioritized areas for improvement. Third, since extensive feedback was provided for each evaluation, students sometimes struggled to prioritize areas for improvement in light of the extensive feedback. One of the perceived strengths of the CSE system was its ongoing appraisal of student performance over time: through a longitudinal series of evaluative experiences and outcomes, the system provides frequent and robust evaluations of student performance. Since it is unlikely that all of this evaluative information can be used, either at the student or program levels, the excessive amount of information influences its use. As one student summarized:

> [The CSE system] is great because of the constant feedback, but it is a double-edged sword: because we get constant feedback we don't always know what to focus on . . . until you get a bad score—then there is a rude awakening at that point. But if you are middling along people don't take it as seriously as they could.

Students described feeling challenged to consolidate a bigger picture understanding of their performance and to negotiate the meaning of their scores through time. Students found it easier to understand their performance relative to each one-off evaluation than to connect their experiences and outcomes to a great understanding of their development of clinical skills. While all students recognized the importance of the system in helping them

feel assured they are on track for future clinical experiences, student descriptions of evaluative information use failed to address direct connections between their experiences/feedback and their overall competency on specific components of the clinical science training program.

While sensemaking around longitudinal trends was not supported by the evaluative information provided, students felt that the curricular and evaluative materials they received clarified what they were meant to learn from each specific evaluation. This included an understanding of how elements of the evaluation were combined to make overall judgments of satisfactory performance. As one student explained:

> [For the summative evaluation], you want to feel confident. You've actually prepared hard for this and you want to feel that you are pretty competent in each category. I do prepare for each of these components, so to just receive an overall pass or fail, I wouldn't feel comfortable knowing how much I passed. I want that number. Everyone would like that number especially when they put in a lot of work to try to pass . . . That's the part that I like about actually having numbers—understanding that combination of [various elements of the evaluation].

Since students used these resources to guide their preparations for evaluations, point values and weighting of elements within evaluative tools influenced student preparation and prioritization of their learning. This implies that the development of evaluative tools must address how evaluative elements are weighted and combined in order to reflect the alignment of prioritization of specific standards and objectives across various elements of the program model. Ultimately, each individual score was interpreted relative to other scores (i.e., concurrent scores as well as those reported in past evaluations). Since each score contributed to a bigger picture understanding of one's performance, students

consolidated the meaning of the scores relatively, though this may have occurred without much intentional thought or guidance. Indeed, students sometimes failed to extract any learning value from evaluative information, as one student explained:

> I don't really think about [my individual scores] very much. It's just a thing that happens. I do fine with patient encounters and I move on. There are too many other things going on in medical school. You can't spend time focusing on any one area or else you'll drown in another one.

In this sense, evaluative information served solely as a means of accountability (i.e., grading).

Tiered sensemaking through intentional design. The preceding sections have focused on the need for sensemaking at the student level. The program administrators demonstrated a strong understanding of the student perspective, a factor that seemed to position them well to refine and improve the program for its beneficiaries. Awareness of sensemaking of evaluative information at the student level is integral to developing an educational program model and its evaluative tools since students are a key stakeholder group in all educational programming. Pushing past accountability-centric orientations regarding the use of data, educators can leverage evaluative information for student-level learning and improvement. By understanding the influences of the developmental continuum anchor, personal factors, and balancing competing information, program administrators can leverage elements of the program to ensure appropriate and meaningful use of evaluative information, a response to the call for greater use of student learning evidence across all tiers of the educational organization for learning and improvement. However, these three themes do not merely describe issues of sensemaking

at the student level, but rather across all levels of student learning evidence use within the educational organization. At the program level, for example, the developmental continuum anchor is useful for identifying possible inconsistencies across curriculum, instruction, and testing as well as cohort-level trends in performance. While these levels may be addressed more extensively elsewhere, higher order uses of student learning evidence can only be considered at the program level within this case study given the constraints of its design. At the program level, sensemaking occurred relative to two orientations: internal or external.

Internally, the influence of sensemaking on the use of student learning evidence suggested that the program model and its evaluative tools must be designed in such a way as to address and support sensemaking processes. In the design of evaluative tools specifically, weighting, timing, and alignment were issues that affected evaluation use. As one administrator explained, evaluative tools must be weighted toward various prioritized standards and objectives as well as according to opportunity to learn and time allocation. Furthermore, scores and feedback must be aligned meaningfully to communicate clearly with students about areas of strength and development. As one student offered, discrepancies between quantitative allocations of points and overall qualitative feedback was sometimes problematic:

> [For one evaluation], I got a certain score and it felt like the comments on that evaluation were very, very positive and the things that were areas of improvement were out of proportion to what the score was. It was a good score—an 80 something—but it seemed out of proportion to the comments. So I asked [an administrator] to elaborate . . . the administrator watched my patient encounter video and gave me a more detailed report of where I could improve and where I did well. Regardless of the score, I thought that was very helpful. The

> administrator didn't end up changing my score, which didn't matter any way because I'd passed, but it was a valuable experience in the end because I did get feedback from the administrator as well as the original feedback.

Administrators also recognized that the CSE system measured various skills unrelated to clinical science (i.e., testing taking skills, time efficiency, subjective interpersonal characteristics, and typing abilities/speed). As such, administrators emphasized that care had be taken to ensure that issues of time management would not result in a failing score. For example, if the evaluation focused solely on that final product of the patient encounter (e.g., a written note or accurate differential diagnosis), student performance that otherwise demonstrated sufficient proficiency might produce a failing score based on issues of time management or technology (e.g., having a keystroke error when copying/pasting). To address these issues, the evaluative schema had to be designed, first, to identify and address such issues, and second, to provide a mechanism for responding fairly, consistently, and meaningfully. All administrators communicated the importance of alignment between the various elements of the program model and its evaluative tools, a theme addressed further within the last section of this chapter.

In the context of this case study, one administrator was largely responsible for making sense of evaluative information. For formative testing, the administrator treated quantitative indicators as warning flags to signal areas for deeper inquiry into student performance. Discussions with other program stakeholders, consideration of opportunities to learn across multiple workshop instructors, and targeted interviews and focus groups with students served to inform the meaning and relevance of the evaluative information. The administrator made in-the-moment adjustments to the program model to

address such hotspot areas prior to the summative evaluation, thus using information towards learning and improvement at the student and program levels within each specific course. Similar analyses and follow-up actions were undertaken using evaluative information from summative testing, again at the student and program levels:

> Before we develop [an evaluation in the CSE system], we look at prior years. We consider whether the curriculum was the same or how it had been modified. We look at the resources, the assignments, and what students were asked to do to prepare for it. Then we look at the item analysis, the statistics from the prior years for any given item. We consult [an assessment and evaluation administrator] to request suggestions on item quality, how should we interpret the data, if we may need to consider adjusting or dropping the item, and if there is any content that we have covered that isn't evaluated. Depending on the course, that process can take anywhere from two to eight hours.

In addition to student learning evidence, information from student evaluations of the course were considered for the improvement of the program across temporally adjacent courses (i.e., one course to the next course in the training program for a single cohort of students) as well as across academic years (i.e., one course to its next iteration with a new cohort of students, a first-year course to its counterpart, topically, in the second year, etc.). Student learning evidence was also used by administrators responsible for training standardized patients and evaluating their performance (as raters) over time. Taken altogether, these efforts drove program-level improvement using evidence of student learning.

Externally, the influence of sensemaking on the use of student learning evidence suggested that the program administrator most responsible for the program was best positioned to contextualize evaluative information. In the educational organization, this

administrator served as a conduit through which all student- and program-level information passed. This administrator was held accountable for cohort-level outcomes and ongoing program improvement through a number of administrative systems and structures through which the administrator presented and contextualized the meaning of evaluations to other stakeholders in the educational organization. The administrator positioned these 'reporting up and out' systems and structures as useful for accountability and improvement at the program level. Working with administrators whose roles focus on testing, evaluation, and institutional effectiveness, the administrator was held responsible for developing five to ten student learning outcomes and/or administrative outcomes plans for each academic year. Interestingly, and as suggested by the literature, these processes were divorced from the day-to-day use of student learning evidence for program learning and improvement that occurred naturally as the administrator used data to address the needs of internal program stakeholders. This trend that is discussed in greater detail in Chapter V.

Problematic 2: Engaging Systemic Complexity

While making sense of student learning evidence for assessment purposes is challenging, it is not the only dynamic that holds a plurality of stakeholder perspectives in tension: student learning evidence is used at different levels of the educational infrastructure for a variety of decision making purposes (Ikenberry & Kuh, 2015; Kinzie et al., 2015). Many stakeholders are engaged in educational decision making, including accreditors and government officials, university and college presidents, provosts, assessment professionals, other administrators, and faculty members (Ikenberry & Kuh,

2015). At the most macro level, accreditors and government officials, who serve to ensure compliance, are neither responsible for the use of student learning evidence nor well-positioned to contribute to local improvement efforts (Ikenberry & Kuh, 2015). At the most micro level, assessment work becomes distanced from the daily lived experiences of faculty and students when it is positioned as an accountability and compliance structure that is facilitated by assessment professionals (Ikenberry & Kuh, 2015) as well as when decision making or improvement efforts do not consider student learning evidence (Kinzie et al., 2015). As a result, faculty may "adopt a role of passive resistance and often become a barrier rather than a pathway to consequential assessment work" (Ikenberry & Kuh, 2015, p. 16).

To overcome obstacles resulting from a complexity of purposes and stakeholder groups, it is paramount that use be considered at the onset of assessment to increase the likelihood of use (Kinzie et al., 2015). Similarly, it is necessary to distinguish between the different uses of student learning evidence and the levels at which they will occur when designing mechanisms for collection of student learning evidence since use characteristics will influence selection of evaluation approaches and mechanisms through which student learning evidence is generated, will determine to whom outcomes should be communicated, and will suggest follow-up actions that might be taken in response to various possible outcomes (Kinzie et al., 2015). According to Kinzie and colleagues, examples of the ways in which student learning evidence might be suitable for these higher order purposes include: programmatically, to answer questions about student learning or questions of interest to faculty members; and institutionally, to connect

assessment with institutional goals, for strategic planning and institutional decision making, for accreditation processes and institutional assessment planning, for improving student engagement and success, for building a culture of teaching and learning and enhancing faculty collaboration, and for reflecting on and improving current assessment processes. To this, Ikenberry and Kuh (2015) add refining learning goals, courses, and curricula, considering technology, informing the budget, improving retention and graduation rates, improving American higher education, and improving the prospects of graduates. Furthermore, while educational programs are held accountable to these higher order levels of the educational infrastructure, which include departments, colleges, universities, systems), educational programs are also ultimately responsible to society since education promotes social betterment through the improved quality of citizens' post-education contributions to the public good. This is particularly relevant for medical education and healthcare training programs since they produce healthcare practitioners who must be competent to address the medical needs of the population.

While student learning evidence can be informative to a complex variety of higher order purposes, it has a primary obligation to inform the needs of teachers and students (hereafter referred to as internal program stakeholders with other stakeholders referred to as external program stakeholders). Internal program stakeholders need student learning evidence to serve accountability and learning purposes at the student level and within the program. Since education is a dynamic construct that varies iteratively with changes to curriculum, standards, technology, and student demographics, the responsiveness of program models to these dynamics must be addressed through ongoing evaluation and

improvement. Such is the goal of assessment-based educational evaluation. Since educational programs already produce evidence of student learning, this information can be leveraged for program-level purposes if evaluative tools are designed to inform learning and improvement at the program level. These efforts require tedious attention be paid to the alignment of various elements of the program model and its evaluative tools. Indeed, within the field of assessment, much attention is paid to the notion of alignment (i.e., the idea that various elements of teaching, learning, and testing must be held in symphony with one another for educational programs to function efficiently and effectively). Taken together with a multiplicity of stakeholder groups, a network of elements held in dynamic tension produces systemic complexity, manifesting as varying partitions of relevance, both in education and its evaluation.

Systems thinking has been suggested as a useful tool for describing and exploring issues of complexity. In evaluative systems thinking (Williams & Hummelbrunner, 2011) complexity is characterized using three concepts: inter-relationships, perspectives, and boundaries. Inter-relationships depict connections, that is "how things are connected, by what, to what, and with what consequence" (Reynolds, Gates, Hummelbrunner, Marra, & Williams, 2016, p. 667). Such connections are not neutral, implying that perspectives result in differential interpretation of inter-relationships. Like inter-relationships, perspectives are also not neutral since issues of power impact the generation of boundaries that determine "what is relevant and what is not, what is included and what lies outside" (Reynolds et al., 2016, p. 667). This framework of inter-relationships, perspectives, and boundaries is used to organize findings regarding systemic complexity

in assessment-based educational evaluation. First, inter-relationships are described within and beyond the program. Then, the perspectives and relevance boundaries of internal and external stakeholder groups are presented.

Inter-relationships. Stakeholders identified a large number of inter-relationships between various elements of the CSE system. Expectantly, the degree of alignment between evaluative information and the program model was paramount and manifest through inter-relationships between elements of the curriculum, session-level objectives and instruction, practice workshops, instructional deliverables, and session pacing/ordering. Similarly, stakeholders described the importance of close alignment relating how current expectations and outcomes align with past (pre-professional training) and future (residency, fellowship, or career) expectations. Understanding inter-relationships between past experiences and the current program model manifested primarily in a need to ensure that students selected into the program did not struggle due to their preparation (e.g., selection characteristics) and that the program model could address a wide variety of "starting points" for students' initial clinical knowledge and skills through scaffolded programming. Being designed to ensure minimal competency, the CSE system demonstrated the importance of aligning evaluative information with four key sets of future expectations: third year clerkship rotations in clinical settings (i.e., medical school rotations in hospitals and out-patient clinics), evaluative expectations for the STEP 2 CS exam, evaluative expectations for future training programs (i.e., residency, fellowship), and expectations for professional practice as a career physician. Inter-relationships between various elements of the program model were equally as

essential, particularly for ensuring that students were prepared by the program for concurrent clinical experiences in the community (e.g., working in the free clinic, shadowing physicians, etc.). For most students, knowledge of the importance of addressing these inter-relationships only emerged in instances of perceived misalignment that provoked issues of fairness. As one student explained:

> We are all bright students. If we are told to do something we are going to do it. No one actively tries to miss points. But sometimes you didn't know that you were supposed to ask in this way or that this was the purpose of this exam or that this was the physical you were supposed to do.

Administrators were much more aware of the importance of these inter-relationships, highlighting such connections as essential to the planning of the CSE system and to the use of evaluative information. Addressing these inter-relationships ensured the utility and feasibility of the program model and its evaluative tools at the student level. Administrators were also responsible for considering these inter-relationships in light of various "student profiles," that is, typical trajectories of students demonstrating relatively high, average, or low performance. In alignment with the minimal competency design, evaluative information was used differentially for each student profile. As one administrator offered:

> There are students we know that "pass" the course but only just barely and have a pattern of that level of performance. And over the years in clinical, those students that are consistently in the lower quartile, we have proactively reached out to them and said, "Hey, you know, yes, you got a passing score but we are not okay with that. We want to give you additional help. We want to talk through this a little bit more." We have been much more proactive in identifying these students in the lower quartile and diving into what is it that is impacting their ability to rise above that. We ask ourselves how we can potentially change that trajectory.

Beyond the student level, inter-relationships between the program and the institution were also identified as affecting the design of evaluative tools and information. External program stakeholders worked with internal stakeholders to ensure that evaluative evidence was generated toward accountability requirements. These efforts were facilitated through institution-level assessment processes. Collaboration across programs influenced the program model and its evaluative tools to ensure that the various information needs of external stakeholders were satisfied. Collaborations also served to align learning across programs (in terms of pacing and ordering) such that evaluative information gathered in one program might be useful in another (concurrent) program. As one administrator highlighted, learning and performance expectations across concurrent courses is informative for improvement purposes:

> We consider if the timing of topic was inappropriate across concurrent courses. Was content inappropriately placed within the context of the course interval? For example, was a clinical skill taught and assessed before the gross anatomy session that would have helped with that skill for clinical performance? If so, we might shift timing of evaluation to maximize learning through horizontal, as well as vertical, integration.

Additionally, since the program model was anchored on a developmental continuum, each evaluative experience contributed to a larger order understanding of student competency based on past evaluations and the expectation of future ones: the whole of the evaluative experience had a greater impact than the sum of its evaluative parts. This higher order reflection on longitudinal evaluation experiences and outcomes was unstructured within the program model yet pervasive in stakeholder interviews. Inter-relationships between the developmental continuum and evaluative tools manifested

as a need to customize the program model for each cohort of students since future evaluations might be impacted by differences in the intended and actual program model. Furthermore, the developmental continuum influenced the purposes of summative testing across the program, as each summative testing provided information that might be leveraged for growth and learning in upcoming courses of instruction.

Perspectives and boundaries. Inter-relationships imply connections between various elements. As previously stated, such connections are not neutral: perspectives of these connections result in differential interpretation of the nature of inter-relationships. Perspectives are affected by power dynamics and impact the generation of boundaries that determine perceived relevance. Perspectives and boundaries are considered for internal (i.e., students and program managers) and external (i.e., other program administrators) stakeholder groups.

Students. Students acknowledged a distinction between accountability and learning when describing the authenticity and utility of the CSE system. From the student perspective, learning seemed to be the primary target with accountability aspects serving as a mechanism for generating proof of learning. As a general rule, students seemed to expect integration of accountability and learning purposes within the CSE system, understanding that though evaluative evidence essential for academic training programs, learning was the purpose of the program. Most students acknowledged that the CSE system, being a simulation of patient encounters and thus distinctive from the true day-to-day practice of a physician, consisted of some elements that were only useful for evaluation purposes. A program administrator provided a clear example of such a

distinction: while physicians can use hand sanitizer in clinical settings, student physicians were required to wash their hands with soap and water since this is an expectation of the STEP 2 CS exam. A few students mentioned this example when they discussed how the CSE system was designed to be as authentic as possible while highlighting areas of discrepancy between varying sets of expectations. Indeed, two students used this example to express appreciation for the difference being identified and explained. Beyond this example, almost all students described differences between the CSE system and real patient encounters, specifically regarding some evaluative aspects of the CSE system, as offered by five student accounts. The first student acknowledged the importance of having simulation experiences since they would be subject to evaluation based on simulated experiences through the STEP 2 CS exam:

> We're working under the assumption that this will be part of a standardized exam we will take some day. You have to assume that regardless of how imperfect any simulation might be, assuming there is a STEP 2 CS exam down the road that will be a simulation, you have to learn how to perform in that simulation environment so you can do as well as you need to when that day comes.

A second student emphasized a primary separation between real and standardized patients in terms of how they communicate: standardized patients are trained not to reveal information unless it is explicitly asked while real patients may divulge more information than is pertinent:

> For me personally these scenarios are sometimes not real life and that's the other reason why it is hard to take it super seriously. I realize I need to change that type of thinking because we have to do STEP 2 CS exam. Just sometimes the scenarios or the way standardized patients respond is not like what it is like in [a real patient] room. It feels artificial, like they are not giving up information that I feel

> would have been given up in the office. Standardized patients are waiting for you to ask a certain question while in the hospital people are typically freaking out and telling you everything they can.

A third student delineated how students knew how to "go through the motions" and "game the system," a limitation through which the simulation environment affects evaluation:

> [In the patient encounter], there are several aspects of it that you can just fake your way through and no one will ever know. For that you really don't need to adequately prepare. Now whether or not you do because you want to be a good physician someday, that's another story. A personal decision . . . I've heard people say on a number of occasions, "You know what, I held this thing up and I looked through it and I said the parts that I knew were supposed to be in there. I just couldn't see anything today." From the camera in the corner of the room, it is going to look exactly the same. So when you are studying and preparing, you ask yourself if you really needed to master the technique to get a scope in an ear and around the corners while not making the patient scream in pain versus just wedging it in there a little and saying, "Oh yeah, I see it" and save yourself some time. You do that a few times and you really don't need to know as many things. Maybe you can figure those things out later.

A fourth student described how the evaluative criteria constrained learning when students did or said things solely for the sake of obtaining points:

> Sometimes I feel like the [evaluation tools] put you in this tiny little box so you aren't learning the skills as much as you need to. Instead, you're trying to get the points for saying that you're doing something. And maybe that is the purpose of [our learning] at this point. I feel like having the checklist for scoring for the exams purposes pushes you toward just saying things even if you don't know what you're' doing . . . I try to know what I'm doing, maybe more than maybe some people, but obviously if you are floundering during an exam, if you know what you are supposed to say you know you can get credit for it.

A fifth student explained this dynamic in greater detail by describing how students "triaged" of evaluation elements based on evaluative checklists and rubrics:

> Ten minutes before, people would say, "Okay I can skip this . . . I need to do that . . ." They were triaging point values and we're like, "That's not really how we planned for it but alright, you do you." Not to say they are slacking off, they were just focused on something else at the time.

Awareness of these perceptions and actions can be valuable for development of the program model and its evaluative tools.

In alignment with the aspirations for assessment in higher education to use student learning evidence for program improvement, students viewed evaluative information as serving two key purposes: establishing their current status along a developmental continuum and identifying areas for their continued improvement. They described identifying specific knowledge and skills to target their study time and personal development, using information to conceptualize how much they might possibly improve from additional efforts, and collaborating with peers who had stronger scores to get help before the end of course exam. The use of evaluative information was, in some ways, inherent within the program design, as three students suggested. First, the feedback was offered in alignment with the program resources (e.g., curricular resources, instructional sessions, etc.) to clarify evaluative expectations and to target feedback toward actionable areas for future development, as one student explained:

> Ultimately, from the standardized patient encounters throughout the block, you get rubrics that indicate your trajectory for your final exam, which matters. I think that's a good gauge. If we didn't have that, at least my cohort is very vocal,

> everyone would be up in arms like, "How are we supposed to know what to expect?"

Second, it was assumed that students and administrators were actively using evaluative information. It was also assumed that these uses were supported by clear and timely reporting systems and structures across stakeholder groups, including access to all evaluative information for self-directed analysis by individual students. One student described such self-directed use:

> [For an evaluation last year], I passed but it was the lowest score I ever received and it was on a part that I normally do really well on. I practice really hard and pride myself in that. I got my score and I was like, "Wow, that was close." I went back to find where I had faulted and I remember I reviewed each checklist . . . I saw what I hadn't done and remembered I hadn't done those things, and I was really nervous. At first it was hard for me to believe but I started adding it up and it made more sense. After reviewing the report, it made more sense and it wasn't as shocking . . . I said to myself, "I'm going to do what I need to so I don't have to cut it this close again," so I poured over the report and noted what I missed and why. I remember a few specific things that I missed and I never missed those again.

Third, the CSE system (and all of its elements) were perceived of as constructive and supportive, not punitive. As one student summarized: "Thinking of people in my class who struggled, they did seek help. [The CSE system] was constructive and the students cared enough to not have that mistake again."

Meta-analysis of one's performance seemed to increase as students progressed through the training program, suggesting that use of evaluative information became more customized as students consolidated findings from each evaluation over time into a larger picture of their own performance. One student described how changing an established

behavior was more challenging that remembering to include a forgotten element. Another student pointed out that strong performance, longitudinally, generated less useful feedback for improvement: “In my experience, [the evaluation] has mostly been more positive feedback than things I could improve on . . .” Students expressed a desire for more consolidated feedback regarding performance over time since evaluative information was shared following each evaluation with little aggregation of findings between or across evaluations at the student level. To this end, students suggested two key areas for development: understanding their longitudinal performance relative to the specific aspects of the training program and receiving aggregated reports of their performance over time. As one student described:

> I hope the scores would be used to compare you with yourself over time: Have you made fewer errors of omission? Have you been more consistent in getting through your exam in a complete, smooth fashion? If you have been dinged repeatedly on a single area, whether or not you agree with it, are you taking steps to identify ways in which that might be addressed? If there is something not coming across to someone reviewing you in these simulations, for whatever reason, are you taking steps to rectify that now while you have the time and ability? Are you getting through the exams in a clean and efficient, time sensitive way? Are you actually doing the exam or just going through the motions, saying you see the retina versus actually observing the anatomy? That would be my hope.

Such uses of evaluative information require consolidation of outcomes across individual evaluations to generate a greater understanding of student performance relative to the key areas of the program model.

Program administrators. Administrators viewed the authenticity and utility of evaluative information from a different viewpoint. Interestingly, most of the

administrators interviewed described the purpose of evaluative information by illustrating its utility to students. This seemed to indicate that administrators had a strong understanding of the student perspective, including the possible "blind spots" associated with student standpoints. Program administrators described a plethora of ways in which evaluative information influenced the experiences and outcomes of individual students, groups of students, and the program. Overall, administrators perceived evaluative information as useful for fulfilling the responsibilities of their roles, for ensuring fairness, for supporting student learning, and for collaborating with students for program improvement. A primary focus on learning was evident in administrators' depiction of their role as educators, as one administrator offered: "We have to use this information to help students because this is not a high stakes exam where tests can determine your whole future and career. Students are here to learn." Administrators described student performance as a manifestation of the interaction of the student with the program model. One administrator described this responsibility as one of intervening to improve outcomes for students who fail to demonstrate minimal competency or who request additional support:

> You have students at different levels. [The CSE system] isn't to 'get' students. It's to prepare students. It's in the terminology. If they fail, we fail, too. It isn't as much pass or fail—it is "This is what you need to work on and we're going to get you there." That changes the whole dynamic. We have a dialogue and explain "This is why this is the way it is." We've got the physicians explaining rationales to communicate the medical knowledge, and we bring students back in immediately to hone in on their improvement for growth.

All administrators responsible for facilitating the CSE system acknowledged the importance of understanding student profiles or trajectories for moving through the program and of using evaluation to clarify misunderstandings about expectations, both for present and future students.

Administrators described mechanisms for ensuring the fairness of evaluations in great detail. With regard to evaluative tools, administrators analyzed outcomes to identify reasons why students might have missed an item, to analyze item statistics and trends over time for an item across cohorts, and to evaluate rater reliability for physicians and standardized patients. These efforts include consideration of the reliability and validity of each evaluative tool, though further discussion of the nuances of these uses is beyond the scope of this study. Findings from such analyses suggested needs to modify scoring when rating was deemed unfair, to evaluate the format of teaching and testing, and to evaluate the "test in light of teaching and the teaching in light of the test." As one administrator summarized,

> I've always looked at grades as just a way of you gauging where you are, but it's also about where we are. If we, as educators, are not helping them succeed then somewhere we are not doing something to help them succeed. Sometimes we look more at a specific student instead of looking at how we are teaching. Not everyone learns the same way. You need lots of ways of teaching since there are so many types of learners with different needs. So it isn't the test scores—it's what they represent. Maybe we think it reflects what a student needs when maybe it is not. Prime example: we'll see a test score and if so many students miss it, we ask two questions: was it taught? And did we not explain it well enough for them to understand what we were looking for? In that case it's less about the students and more about us and what we did.

Administrators also highlighted the importance of recognizing the temporal dimension of evaluative information to ensure a fair interpretation of the relevance of evaluative information, as one administrated expounded:

> If [the student] gets an 80% in the first year and an 80% in the third year, there's been a lot of growth. They might wish they got a 90% [in the third year], but if they went back to the first-year exam they could probably get that 90% [on the first-year exam]. So there's growth, but you can't see that in the number. That said they may get to third year and get a 90% [on the third-year exam] and might go back to that first-year exam and get a lower score. We've never tested it. First year you know it's going to be a cardiac exam because that's the topic of the block. Third year you are in internal medicine and the assessment might be cardiac or pulmonary or abdominal—you don't know what you're getting. It is hard to make comparisons with just the number.

Since each cohort of students received instruction from preceptors in smaller groups, many efforts were made to ensure fairness through consideration of opportunity to learn by checking the alignment of curricular resources, instruction, and testing. Norming preceptors to standardized instructional and evaluative materials posed a significant challenge, one exacerbated by many students privileging physician instruction and feedback over that from standardized patients or educational administrators who were not physicians. Opportunity to learn considerations also involved analyzing the pacing and progression of teaching and learning within and across courses to ensure students had adequate practice and preparation time between instruction and testing, not only relative to the pacing of the course but also regarding the other courses in which students were concurrently enrolled. Decisions were made using multiple sources of evidence to give a good indication of students' ability to think critically and perform skills at the expected level. Scores initiating amelioration and remediation were graded by multiple raters to

ensure fairness, an effort also undertaken when a student had an outlier performance (relative to the student's longitudinal pattern of performance).

Administrators ascribed a primary importance of evaluative information as supporting student learning. Program administrators designed the model to ensure minimal competency through ongoing improvement, intertwining accountability requirements within the pre-established learning framework. Students were engaged in small group and whole class discussions using evaluative information, providing a mix of personalized and cohort-referenced feedback on trends and areas for development. These activities were spaced meaningfully throughout each course within the training program in effort to use the evaluative information to inform students' progression along the developmental continuum. The majority of interviewees described how evaluative information was used to initiate dialogue, including between students and faculty, between students and standardized patients, and between faculty and standardized patients. Students were also pulled aside for one-on-one consultations in which students and administrators could review evaluative information together, an activity that typically resulted from low scores, "red flag" comments or behaviors, or student requests for additional feedback. Students were referred to other professionals within the academic community (academic support services, assessment offices, etc.) for additional support, as appropriate. Furthermore, these mechanisms were pre-determined based on cut-off scores for aggregate and component-level outcomes that suggested the need for amelioration or remediation. Administrators analyzed evaluative information to look for

themes at the student and cohort levels that might explain why a student failed to demonstrate mastery.

These efforts did not end at the student level: administrators used student learning evidence as the basis for most program improvement efforts, including collaborating with students as informants about areas for development in the program model and its evaluative tools. Through one-on-one interviews, focus groups, and surveys, administrators collaborated with students to identify areas of the program that impeded learning and performance. Through these efforts, administrators solicited student feedback as well as student ideas for improvement, extension, different learning modalities, and/or creative ways to approach teaching the topic in future iterations. Students shared resources that could be integrated into future courses. Administrators also kept an ongoing log of their own ideas about areas for development, logs that could then be analyzed in tandem with student learning evidence to select assessment goals and create assessment plans.

Stakeholders external to the program. External program stakeholders were involved in higher order evaluations of the program, typically using student learning evidence to understand the program relative to other elements of the institution (e.g., other academic programs, academic or behavioral intervention systems, etc.). Administrators working in assessment, evaluation, institutional research, and institutional effectiveness indicated the possibility of using student learning evidence in their roles though these efforts were largely aspirational at the time of the case study. Interestingly, one administrator was viewed as being responsible for evaluative outcomes and for

relaying and contextualizing the meaning and relevance of the evaluative information to external program stakeholders. All administrators within the external program stakeholder group expressed trust and confidence in the program and its leadership, suggesting that direct ownership of the program may influence how evaluative information is viewed and used. This sense of ownership seemed to amplify use by providing program administrators with a sense of power of the program model and its evaluative tools. Evaluation use was thus integrated into the organizational culture in alignment with various accountability and learning purposes.

Invariably, internal program stakeholders combined student learning evidence with other sources of evidence (such as findings from perceptual surveys of students) to contextualize outcomes. In this sense, administrators were held accountable for the findings of their evaluative information through institution-specific systems and structures that required program-level administrators to present findings and action plans resulting from analysis of cohort-level testing and survey outcomes to peers within the organization. Interestingly, though these systems and structures align with the overarching goal of assessment-based educational evaluation, the institution did not consider them as part of the formal assessment process. These structures tended to focus on peer-peer accountability for administrators, on fostering collaboration across programs to address trends for improvement, and to identify institutional elements or individual students who might need additional resources or support to continue within the program. As a result, strong or acceptable performance elicited milder responses than lower-than-

normal performance or atypical findings. In these ways, the purposes for acquiring and using evidence of student learning were largely rooted in an accountability mindset.

At levels of the educational organization beyond the scope of the program (e.g., the department, institution), relevance boundaries seemed to focus mostly on accountability. This was possibly due to an assumption that learning purposes had been addressed within the program by its leadership. For external program stakeholders, evaluative evidence provided documentation of student performance to justify decisions made about the student. Discussion of the importance of documenting student outcomes centered on ensuring student physicians demonstrated a sufficient level of competency to be permitted to treat patients, connecting institutional accountability efforts with greater social accountability concerns. Inferences made based on evidence from each separate evaluation supported promotion of the student based on demonstrated competency for the specific construct of clinical knowledge and skills at each evaluation. As one administrator described, these inferences are limited by the content of the evaluation as well as by the developmental milestone:

> [From the outcomes] you know [the student] can pass this test at this time. You can make a claim they know how to do it, but you can't claim they will do it. That's a big difference, an important difference. You can claim that they can do it for various cases or situations and many of those things do apply across cases but it is not all-encompassing. So you can't make the claim that because they can do this they are ready to do it [in all clinical contexts]. This is why, to me, we have [medical residency training]. So the claim now is they can do it to a certain level. The reality is that the can do it to a certain level with supervision until they are ready to do it on their own.

At the program level, accountability requirements facilitated analysis of student learning evidence for groups and cohorts of students as an aggregate indication of the program. Administrators were required to complete student learning outcomes and administrative assessment plans and reports that were filed electronically for accreditation purposes. Interestingly, these assessment plans and processes were considered separately from other learning and improvement efforts despite the many instances in which student learning evidence was used for accountability and improvement at the student and program levels within the normal day-to-day functions of the program. When asked about assessment, all administrators referred specifically to formal documentation requirements, confirming a trend in the literature of a perceived distance between assessment and the ongoing everyday responsibilities of teaching and learning. As explained by one administrator, an explicit focus on compliance requirements caused the work to feel tedious and demanding despite the administrator feeling a strong commitment to its importance and utility:

> Assessment is probably my least favorite part of my job. While I do not enjoy it, I believe it is important because it holds the faculty and supervising administrators accountable to taking the time to identify where within the curriculum we need to focus on and improve. I appreciate the process of having to dissect the data and really look at where the breakdown potentially is: is it the faculty member, the teaching style, the content, the curricular materials, degree of exposure/mastery and time allocated? Is it within the way we are assessing or the portrayal of the case, or maybe even a matter of the entire curriculum? Assessment forces us to consistently look at where the student outcomes point to weaker performance and to apply a process by which we can identify or hypothesize why that may be . . . Assessment is an important part of medical education and these accreditation expectations need to be in place.

One administrator outlined the compliance expectations for assessment within the educational organization:

> In our system, we have student learning outcomes plans [based on student learning evidence] as a part of our requirements as well as administrative outcomes plans [to address improvement of the learning environment unrelated to learning evidence]. Together, this offers a larger structural balance. The two types of plans are different although they overlap a lot. It is easier for faculty to look more at the administrative outcomes plans. If this type of analysis wasn't mandated, we would miss opportunity to try to maximize individual student performance based on data.

The same administrator further described the value and relevance of assessment based on these two types of assessment efforts:

> Administrative outcomes goals are easier to address based on a hunch or observations, but student learning outcomes require the use of data to identify weaknesses. You really can't use a hunch about how students are doing, you have to look at every component, each student, overall cohort averages, and patterns over time because not every cohort is the same. Now, after many years of doing assessment work with a stable and consistent curriculum, we can see patterns in individual student performance that allow us to more confidently identify students in trouble or who might be an outlier in our program. Previously we couldn't do that—we could only identify challenges at the instructor or curricular level. Now we have enough data to see variation due to student and we are able to tease out what might be a learning style or learning disability versus a personality or professionalism issue.

As an exception to this general trend toward an accountability focus at higher levels of the educational organization, the aforementioned dialogue between peer administrators for collaboration and institutional improvement promotes learning and improvement at the institutional level. Additionally, the organizational structure includes committees and boards that address learning and improvement through strategic

planning. For this specific case, such functions involve different stakeholders than those interviewed and do not use evaluative evidence from the CSE system within their proceedings. Overall, while outcomes from each evaluation were treated and stored separately, and uses of student learning evidence existed in silos separated by role, purpose, and function, it is conceivable that this rich evaluative information could be databased and leveraged for more sophisticated analyses supporting learning and improvement.

Problematic 3: Attending to Power and Information Gaps

A lack of cogency across these layers of systemic complexity result in substantial information and power gaps. This is foremost evident in the compliance culture that permeates assessment practice wherein faculty, who are best positioned to use student learning evidence, are oftentimes distanced from assessment work (Ikenberry & Kuh, 2015). Prominently, the engagement of faculty throughout all phases of assessment to engender ownership and foster the use of assessment for improvement has been offered as a means to close this gap (Kinzie et al., 2015). Effective communication structures and practices are necessary to bring assessment efforts into alignment (Jankowski & Cain, 2015) and to bridge course, program, department, and organizational stakeholder groups, assessment priorities, and information needs. This is necessary to address the shifting needs of users of student learning evidence (Jonson et al., 2017) as well as to develop shared understanding of assessment processes and products (Jankowski & Cain, 2015). Though useful, these efforts will be unsuccessful if they fail to account for persistent factors of assessment contexts that evolve within and across academic periods. While

each specific educational organization can expect local variation of assessment processes and products based on the issues facing their campuses and the interests of their faculty, Ikenberry and Kuh (2015) have identified five broad evolutionary trends in assessment contexts across educational organizations: ever changing student characteristics and needs, never ceasing technological advances, growing competition for students, more challenging economic circumstance, and skepticism about higher education quality. Developing internal systems and structures that facilitate engagement and communication across multiple tiers of stakeholders (each of whom hold different positions of power within educational organizations) while addressing multiple purposes (accountability, learning) provides a formidable obstacle to actualizing the use of student learning evidence for educational improvement.

Overall, assessment presently faces a critical juncture as it seeks to engage educational improvement using student learning evidence, expanding the traditional scope of assessment work beyond that of accountability and compliance. As Ikenberry and Kuh (2015) assert, accountability is a necessary purpose for assessment work but without alignment with local learning and information needs, accountability-based assessment constitutes a missed opportunity and a waste. Kinzie and colleagues (2015) frame this transition in an exploratory light, stating there is still a great deal that needs to be learned about how to do this work well, including: How can practitioners transition from “doing assessment” to using assessment findings? What strategies effect use? What principles can be used to guide the field in its efforts to further the use of assessment evidence to improve student learning? To these ends, stakeholders responsible for

facilitating and assessing educational programs must develop rational program models with clear purposes and expectations as well as evaluative tools that generate information that directly facilitates improvement and learning. The program model and its evaluative tools/information must be held in alignment relative to at least three important dimensions, summarized here as quality frameworks and information use protocols, ultimate information needs of internal and external program stakeholders, and evaluative milestones within the program model.

To generate and maintain alignment across these dimensions, stakeholders must plan a rational program model and its implementation, create evaluative tools (including mapping the intended process and findings use of evaluative information generated by these tools), design responsive management systems and structures to adapt to complexity factors, and implement the program model and its evaluative tools in response to unexpected deviations (i.e., differences between the intended and actual programs). The last two steps are essential for facilitating the use of evaluative information across power and information gaps: more than descriptively, the program model must include plans that support responsiveness to emergent issues that alter the meaning of student learning evidence as well as to differential student-level outcomes, such as extension, amelioration, or remediation. This is necessary given the need to attend to the program model as it transpires for each cohort, section, or individual as a means of responsively managing the experience and outcomes of the program model and the evaluative tools that provide an indication of its quality.

Quality frameworks and information use protocols. In the modern era of standards-based education, educational programs are intended to provide equitable learning opportunities for students across many educational organizations offering a specific type of training. Standards-based programs are thus designed in alignment with internal and external quality frameworks that describe what students must know and able to do upon completion of the program (hereafter referred to as quality frameworks). This is directly aligned with approaches to evaluation that are objectives-based or goal-based. With the increasing focus on the use of student learning evidence in assessment, institutions of higher education are increasingly required to produce documentation that showcases how information is used within the educational organization. Thus, in addition to quality frameworks, program designers must consider the requirements of protocols that mandate the use of the information generated (hereafter referred to as information use protocols). Taken together, quality frameworks and information use protocols provide a means for communication from external program stakeholders regarding the nature and structure of the educational program. These materials are received by internal program stakeholders who must access them, understand them, crosswalk them with the existing program model and its evaluative tools, and adapt the program model to meet new requirements when these frameworks and protocols change over time (Spillane, 2002).

At the student level, the educational program must be able to communicate the extent to which each student has achieved each objective of each quality framework in alignment with the mandates of each information use protocol. These standards evolve over time, requiring the program model and its evaluative tools to be checked for

alignment and updated for each iteration of the program. Within the case study, eight distinct sets of quality standards and three information use protocols influence the CSE system. Taking an example from the most recently added set of standards, the program model and its evaluative tools were aligned with the *Core Entrustable Professional Activities for Entering Residencies* (EPAs) framework of the *American Association of Medical Colleges* (AAMC) within the last five years. These expectations were mapped across all programs related to clinical science with the educational organization, including multiple program across which standards had to be partitioned and aligned. Since the program was previously aligned with internal expectations for third year clerkships, the EPAs framework was already largely aligned with the program model and its evaluative tools; however, the new framework included some standards and required evidences for which information was not previously generated, introducing the need for new evaluative tools. A short paper assignment and rubric were integrated into one course of the training program.

With regard to this incremental revision and updating process, an administrator described the anchor for such efforts as a means of addressing the shifting target of post-graduation preparation for success:

> Scores give us confidence that we have created a system that says, "When a student hits an established threshold at a particular point in time, it is gauged at the right level and if the student is there then they will know that that they are at the appropriate developmental level" and that should give them the confidence to know that they are on the right track.

This sentiment was echoed by a student who reflected on the purpose of the training program, in metaphor:

> You have to know where to go. You have to know how to drive, but you don't have to know every turn. You have to be able to pull it out when you need it, even if you don't need to pull it out every time.

At the level of standards and objectives, the purpose of the training program is to equip students with a specified toolkit of knowledge, skills, and mindsets that set them up for successful use of these tools in future real-world settings. At the program level, alignment with the quality frameworks and information use protocols ensures that the model is viable and that the educational institution is following through on its mission, vision, and purpose of preparing physicians. These guiding visions clarify expectations for all stakeholders about content, and its relative importance, along with longer term needs and outcomes.

Since the program administrators have been stable presences in the program for almost a decade, administrators often synthesized their own reflections on the program, its evolution, and its needs for development without codifying this information formally. For example, one administrator observed the differential performance of students based on their pre-medical school education, relative to their age and previous professional experience, and concluded that these factors influenced student evaluations resulting from the CSE system. The administrator integrated these observations with those made based on evaluative information to intervene with students who may seem to follow similar trajectories as those of former students who had similar entry points or

background training. Through these efforts, seasoned program administrators can provide customized teaching, feedback, and mentorship. The familiarity of all administrators with the ebbs and flows, or hot and cool spots within the program model and its evaluative tools, seemed to support their ability to know, almost intuitively, how to tweak the model through time. As previously emphasized, these efforts were rooted in a focus on aligning the student experience with future expectations and needs.

Since the system was tweaked and refined slowly over time, it is now considered a pillar of the educational organization and a "well-oiled machine" that consistently performs and is received well. Given an ability to achieve intended outcomes with fidelity over many cohorts, the program model and its evaluative tools were maximized to meet the needs of individuals students through customized support based on evaluative information from the CSE system. This level of program development increased the capacity and bandwidth of the program to include new CSE system-related offerings (e.g., one-on-one trainings, practice sessions, and workshops) that furthered the quality of student learning experiences. More than simply satisfying checkmarks on rating tools to earn points, students were offered optional training sessions through which they could personalize their practice and receive evaluative feedback customized to their information needs. This was particularly useful in the context of the training program since there a few other means of gaining access to such evaluative feedback. Through the experience of the evaluation, students develop skills that can only come from the experience of the evaluation (a process use of the CSE system), such as timing, organization, flow, and efficiency.

Importantly, educators and students depended on evaluative information to suggest not only the degree to which students met the necessary standards of current performance, but also to indicate how students can expect to perform in the future based on having completed the program. This required an understanding of how quality frameworks and information use protocols support alignment across these sets of performance expectations at various points along the developmental continuum. Such alignment of expectations is particularly salient in the context of medical education, wherein student physicians will graduate medical school programs and matriculate into differential residency training programs to concentrate in one of many different medical specialties (i.e., family medicine, pediatrics, anesthesia, surgery, etc.). Program models and their evaluative tools must be rationally positioned to support student transitions through the curricular system, an effort that requires responsiveness to individual and cohort-level student needs, to evolving characteristics of educational contexts, and to a shifting understanding of science and medical practice. For most stakeholders, this expectation of alignment between current performance standards and future expectations manifested around the vision of success for a graduate of the program, as one administrator described:

> Mostly we want to influence change at the individual level . . . The ultimate goal is really to develop astute physicians that are knowledgeable, appropriate with patients, and passionate. As well as we can, we encourage that. I'm not sure that just having a score and telling them areas where they are weak means we are demonstrating that behavior but it may be at least something for them to think about where they can say, "Okay I'm not doing well on that. What do I need to do to improve? What is the expectation?"

Students described this vision similarly, positioning the training program as a setting in which to integrate basic science knowledge and clinical practice within the context of a patient encounter.

All students described how this construct shifted from the beginning to the end of the program with their initial efforts focused on engaging and communicating with patients and with a gradual, long-term shift in focus toward responsively customizing the encounter, generating differential diagnoses, and documenting the encounter in a medical record format. All students also discussed the importance of establishing a strong baseline of communication skills onto which all subsequent skills could be added, offering that skills like connecting with patients of a varying temperament, making patients comfortable when talking about intimate topics, directing patients to move into positions for the physical exam, avoiding seeming judgmental, and steering the conversation were much more challenging than performing physical exam maneuvers. In the later stages of the training program, students grappled with increasingly more advanced skills, such as conducting multiple encounters in sequence, thinking critically about possible diagnoses while simultaneously conducting the interview and physical exam, and reporting information through a timed electronic medical record-keeping system (which required students to increase their typing speed and efficacy, learn acceptable medical abbreviations, and avoid typing or shortcut key errors that might inadvertently delete patient information). Development of skills like these (e.g., those skills not driven by a specific content focus) was pervasive across units or courses of

instruction, a longitudinal dimension along which student performance was expected to improve through exposure and repetition.

Information needs. A rational and responsive educational program model must consider the various information needs of stakeholder groups associated with accountability and learning purposes. Bringing these purposes into alignment is a challenging endeavor given the well-acknowledged truth that evaluation cannot address every information need or answer every question (Weiss, 1988). As previously outlined, educational evaluation has historically focused more on accountability purposes than on leveraging information for learning and improvement. As a result, educational information has historically focused on grading students and making decisions about credit acquisition and promotion/retention. While it is implied that students can learn from analyzing graded work and overall performance reports, such learning is not assured unless it is included with the program model. If evaluation of student performance and program performance is to be predicated on learning and improvement, accountability purposes must be integrated with learning purposes in alignment with the program model and its evaluative tools. Stakeholder relevance boundaries and perspectives need to be pared in alignment with the program model as the primary intended stakeholder.

Rooting a program model and its evaluative tools in learning suggests a primary focus on objectives and standards, measurement of student performance on those objectives and standards, and using evaluative information to make decisions and take actions. In this case study, students expected to be taught knowledge and skills, to gain comfortability and confidence in that knowledge and skill, and to know how and when to

use that learning in future practice as a physician. In this sense, the program model can be viewed as a set of knowledge and skills for which clinical science proficiency must be demonstrated for students to earn credit and progress to the next course. In such a model, more power is shifted to the student stakeholder group: not only do students' needs and performance outcomes dictate future action, but student performance is externalized from the student, viewed as an intersection or interaction of the student and program model at a given moment in time. Student learning can be achieved by applying more resources and customized teaching and support to fill gaps and ensure students have met minimal competency. Rather than being a regimented protocol 'done to' students with fidelity, the program model was structured to identify areas for development and to provide mechanisms for closing identified gaps. Evaluative tools were designed for objectivity, standardization, and fairness, as two administrators explained. One administrator emphasized the importance of evaluation as accountability for students:

> I think [the CSE system] is extremely important. I think it is one of the most important things we do. It's human nature: unless we are held accountable for what we are assigned or expected to cover, no matter what level we are, not just medical students but also for practicing physicians. We have recertification exams for a reason. If we are not held accountable for the information that we are expected to learn then we don't cover and internalize it to the depth that we should.

The second administrator expanded on this accountability purpose by highlighting how evaluative feedback informs learning and a greater understanding of one's performance and abilities:

> We have to grade them at the end of the day, so we throw that into the experience. As educators, we need to make sure our students are also meeting certain standards. I think purpose of the CSE system is for them to understand if there are certain things they didn't do in an evaluation that would make a patient more comfortable. Students are supported to learn and grow from that and not make that same action in the future. They can apply learning and feedback to another patient scenario. The other parts that we are grading are myriad, things like how well they are able to write a medical record for example. While we do put down all the little pieces and checkmarks on [the evaluation tool] to see if the student included each part needed, I don't know if that is necessarily as important as being able to put some of this information together in a diagnosis.

All students who were interviewed felt the CSE system was appropriate, supportive, and useful, highlighting that any disagreement about evaluations tended to be rooted in small discrepancies. Evaluative tools were also designed in alignment with those expectations that are in place for the STEP 2 CS exam that students will take following their third year of medical school. Every effort was made in the design of the CSE system to ensure the experience and feedback aligned with STEP 2 CS exam expectations as closely as possible, ensuring that students received adequate preparation. This might imply an overemphasis on standardized testing; however, testing preparation considerations were balanced with other important standards and learning goals, ensuring that test preparation did not dominate the program model. By the end of the training program, as one student described, patient encounters in the CSE system permitted students to hold checklists and rubrics in the back of their minds while performing patient encounters.

For this CSE system, accountability requirements were integrated within a broader focus of student learning along the developmental continuum. Rather than producing individual evaluations and reports for each set of requirements (i.e., quality frameworks and information use protocols), all information needs were integrated into

the existing CSE system and the other mechanisms for generating student learning evidence which, taken together with perceptual surveys and occasional focus groups, constituted all the evaluative information for the program. Accountability requirements were not, however, integrated within a broader focus on program improvement using assessment systems and structures. As previously mentioned, ongoing improvement efforts were seen as "part of what we do" while assessment processes were seen as distinctive. Program administrators chose to integrate program improvement structures into the day-to-day processes of the program. Streamlining these direct and indirect sources of evidence by integrating them into the program model saved time and effort on behalf of a small program administration team. It also avoided redundancies and ensured a high degree of alignment between the program model and its evaluative tools since program administrators reviewed them within and between the courses that constituted the training program. Alignment in this (educational) context was quite formidable given the need to bring many diverse elements into symphony: first, within the program model, curricular resources, instructional sessions (across multiple preceptors), opportunities to learn, and evaluative tools had to be aligned; second, within the evaluative tools, items had to be aligned with the case, the case had to be aligned with developmental level, the patient portrayal had to be aligned with logistical constraints, and the scheduling of exams had to be considered relative to demands being made of students in other courses. Holding these elements in alignment required not only that alignment be planned, but also that alignment be responsive to emerging and oftentimes unanticipated factors (i.e., cancelling of class sessions due to inclement weather, scheduling issues, time constraints,

etc.). As a result, alignment was a malleable and ever-shifting notion to which administrators had to continually be responsive.

Evaluative milestones. Along the developmental continuum of the training program, evaluative milestones provided students with feedback on their performance relative to each expected performance landmark. These milestones are aligned with the overarching purposes, key learning outcomes, and big ideas for each course within the progression. As previously described, each course utilized at least one formative and one summative examination. Students were required to participate in amelioration or remediation if they did not demonstrate proficiency on each summative exam and were additionally permitted to opt-in to additional practice sessions offered throughout each course. The CSE system was thus built to be consistent across courses in overall format and structure with each individual summative evaluation distinctive in the particulars of its content, timing, and scope. Importantly, students were never tested on skills in a summative evaluation if they had not been allowed the opportunity to practice the skills in a formative evaluation earlier in the course. Evaluative milestones thus offered memorable patient cases that encompassed key learning goals for the course, that created opportunities for students to practice the skills associated with the key learning goals, and that provided individualized evaluative information for students on each of the skills associated with the key learning goals.

By anchoring these evaluative milestones in expected visions of student performance at each stage of the developmental continuum, program administrators could use the information to generate an understanding of outcomes at the student and program

levels. From course to course within the developmental continuum, or from cohort to cohort, comparisons of performance identified areas for improvement. Importantly, the alignment of these evaluative milestones with one another positioned evaluative information to be useful across courses: findings from a course in the progression might suggest areas for development in courses immediately preceding or following it. Also, since the training program included a recursive element (i.e., the program model first focused on healthy patient encounters, then patients with pathology, for each type of physical examination like musculoskeletal, cardiovascular, neurological, etc.), evaluative findings from a specific course within the first iteration program informed the second iteration of training program for a given cohort.

For students, this gradual and recursive program model allowed them to address their feelings of nervousness and intimidation in a simulated context. All students expressed appreciation for the opportunity to address these developmentally normal feelings in the context of a simulated patient encounter instead of in a clinical setting with real patients. In this way, the CSE system created a safe space in which students could gain experience "without hurting patients," an observation made by the majority of students. Students suggested that, with time, they grew comfortable with the CSE system and began to relax their expectations about obtaining flawless scores once they learned to expect that some encounters may not "be perfect." Such observations suggest the conceptual process use of evaluative information generated by the CSE system. As one student summarized,

> I feel [the CSE system] is a priceless tool—invaluable. I've had a great experience with it so far. Not just looking at it as a standardized test but as a way of introducing more of an artistic side of medicine than a bare bones basic science part.

Many students echoed this sentiment with a description of how the training program facilitated their personalization of interviewing and examination style within the guidelines and checklists offered. This was an area of tension for students, who seemed to struggle oftentimes to identify the elements of an approach that could be customized. Students also acknowledged that becoming too comfortable with the routine of an exam (i.e., the evaluative checklist) could be problematic: complacency might cause a student to "fall into a trap and miss something big." Taken altogether, as they transitioned into later courses in the training program, students seemed to focus more on developing their own practice and less on obtaining a specific grade. As one student offered,

> Initially it was more about feeling comfortable. I'd say throughout our first year, you were cognizant of the fact that you were being graded and that mattered, but I think the score felt more like a practice run so you could get comfortable in that situation. Moving into the second year, you become a little more aware of your score more as a benchmark of, "Am I doing this the way that I'm supposed to be doing this?"

This seemed to suggest that, early in the training program, evaluative information was used more mechanistically with many mentions of "checking off boxes" on the evaluative tools to earn points while later use of evaluative information bolstered students' internalization of basic skills. Many students described this use of the CSE system as one of "getting into a good habit," "building muscle memory," and "making it second nature"

so that they would be able to perform capably in unexpected situations, as is the nature of patient care. One student described:

> In year one, you're just learning all the questions you could possibly ask. When I was a first-year student, my second-year peer mentor explained it to me as Mr. Miyagi from *The Karate Kid* is getting Danny to wax on and wax off. Danny doesn't get the point until Mr. Miyagi starts punching him and he realizes the point of doing all the cleaning work was to apply it to karate, even though it didn't feel pertinent at the time. This is like for me when it didn't feel it was pertinent to do the entire neurological exam for a healthy person. But sometimes that turns out to be helpful, like in your internal medicine rotation [in third/fourth years] when you have a stroke patient—just knowing the different skills you could use to figure out the issue and knowing the different things you could test that you sometimes don't think about testing in a normal healthy person. With a stroke patient, you have to locate the neuro deficits. The first year going through a whole physical exam helps to tell us different ways to test out each specific cranial nerve.

The findings from evaluative milestones were used in many diverse and interesting ways, especially feedback from formative encounters. Program administrators reviewed feedback generated by raters to identify cohort-level trends and to identify areas for improvement, both for the program and for the students. For some courses, students conducted self-evaluations of their performance and/or students and instructors reviewed videos of patient encounters in small group discussions. These activities, though focused on learning, resulted in unexpected consequences for students. A few students expressed discomfort with having peers watch their experiences, explaining that this practice affected peer-peer perceptions and relationships. In self-evaluations, students created plans for improvement for which they were not held accountable. Depending on the degree of student investment, the plans might not be of high quality or followed through upon, weakening students' commitment to learning from these plans in light of the

necessary prioritization of time associated with being a medical student. In such cases, these reflections constituted a somewhat symbolic use of evaluative information.

Evaluative milestones are an essential element of the training program. The program has been built locally by the faculty and staff to structure student experiences in alignment with specific learning goals and milestones in other parts of medical education, specifically preparation for third-year clerkships, residency and fellowship programs, and a career as a physician. One administrator described the purpose of the program as "providing students with a toolkit for real-world practice in a sequential and supportive way." The administrator stressed the need for students to practice and refine skills iteratively since "longitudinally sequenced learning is really challenging for students." By breaking down the end goal into more approachable milestones, and by providing structured and systematic feedback at each of these milestones, program administrators could teach and evaluate "developmentally identified aspects of the interview, physical exam, and clinical reasoning for your given level that will be needed to prepare you for patient care in an instrumental way." Students developed these fundamental skills through didactic and heuristic sessions that focused primarily on exposing students to new content through multiple methods (i.e., reading textbooks, reviewing evaluative tools, viewing example videos) and facilitating guided practice in which students could imitate examples and receive formative guidance on their developing skills (i.e., workshops, small group discussions, reflections, optional practice sessions, etc.).

Importantly, all elements of the program model and its evaluative tools connected back to the developmental anchor and were aligned with one another with particular

attention paid to how students "move through" the system. The student experience was aligned across all courses that students take at a given time (horizontal alignment), across the training program (recursive alignment), and across all four years of medical school (vertical alignment) to ensure students meet minimal proficiency requirements regarding the knowledge, skills, and mindsets inherent in the program, department, school, regional, and national standards and/or objectives. To further "influence change at the student level," administrators identified the need to communicate evaluative information developmentally instead of relative to the testing instance (e.g., exam 1, exam 2, exam 3, etc.) for the student and the program. For the program, performance relative to objectives and standards has been in place for many years: student performance is parsed by objective, course, and instructor for analysis through assessment processes. By extending this approach to student-level data, evaluative information can be repackaged and reported relative to core knowledge and skills, facilitating the evaluation of longitudinal student performance relative to specific objectives and standards. Such an effort that might more effectively identify areas for improvement at the student and program level and extend the existing use of student learning evidence for learning and improvement.

Findings from Reflective Exercises

Findings from the reflective exercises are presented in this section in effort to draw a distinction between the reflections of program stakeholders (i.e., indirect reflections and document review findings presented in the previous sections), and the actual sensemaking processes in which I observed students engaging as they made sense of evaluative information directly following the release of score reports. When analyzing

actual score reports immediately following their release, students expressed their impressions, understandings, and anticipated uses of the evaluative information. For this particular evaluation, no students were identified for amelioration or remediation; all students received passing marks. Five students were available for interview as they explored their score reports for the first time.

All students reflected on their experience of the evaluation to make sense of their evaluative feedback. This process use of the evaluative experience impacted how students made sense of and used findings, as three students described. As one student explained:

> Overall, I tend to feel pretty good about the exam as long as there are not glaring deficiencies [in my performance]. Most of the things I know I forgot by the time I'm out. If there are discrepancies with that [in my scores] then that bothers me but I'm not seeing anything that is too much of a concern to me.

The second student emphasized how the experience of the evaluation (conceptual process use) frames a student's reception of feedback:

> Usually I'll get a pretty good feeling during the exam about if I blatantly missed something or if something was jumbled. I did not get that feeling this time around so I felt pretty confident about what my scores were going to be.

The third student positioned the feedback as provided a more fine-grained understanding of performance than the experience offered, which supported learning purposes:

> I feel like you leave knowing you have a good sense of how you did, but I think the more important feedback is the stuff you can actually improve on. Like asking a specific question differently or something like that.

Furthermore, students described how their perception of the experience led them to expect much lower outcomes than they received. In this way, students' process use of the evaluation experience caused them to experience feelings of dread and failure, as two students shared. For one student, this manifested as relief:

> It is very easy to take these scores and conflate them with completely good or completely bad and it's important not to do that. So [I am feeling] partially relieved and partially as expected, but mostly just relieved that it wasn't as bad as I thought it would be.

Another student's reflection showcases the importance of considering pacing and timing in the development of evaluative tools so that time management (a single problem or area for development) does not dominate results:

> I think [my evaluative information] is very reflective of how I did. I think it was very generous. I feel like I should have been able to time-manage and actually finish my whole exam, but I obviously didn't, so I thought it was very generous of a grade that I got.

One student went on further to describe having stressful dreams regarding the elements of the exam that the student neglected to perform.

All of the five students analyzed their reports by first looking at quantitative outcomes to target subsequent analysis. One student received a perfect score on all checklists and rubrics, prompting the student to end the analysis without reading qualitative feedback. Of the remaining four students, three students only accessed evaluative checklists and rubrics on which they "missed points." One of these three students then went through each checklist line by line to analyze feedback, while the

other two only reviewed feedback based on quantitative deficits. In this way, students' analysis stemmed from a depreciative lens, seeking out errors and areas for improvement without much consideration of areas of strength or positive commentary/feedback. The fourth student checked each checklist line by line regardless of sum score. None of the students acknowledged looking for feedback on a specific area they had previously identified as needing improvement based on past performance. Three of the five students identified pacing and efficiency as areas in which future improvement was needed.

Personal factors seemed to impact how students engaged with evaluative information. For example, students received evaluative feedback from a variety of raters, some of whom were standardized patients and some of whom where physicians. Three of the five students discussed prioritizing one type of feedback over another. Students seemed to focus more on physical examination skills (rated by physicians) or on interpersonal, communication, and interviewing skills (rated by standardized patients), offering rationales for their prioritization of one set of skills or the other. As one student offered,

> My standardized patient checklists are usually pretty high. They are the more important ones to me. The social and interpersonal skills are more important and everything else I feel will come more naturally the more I do it, so I always check those things first.

Furthermore, two of the five students described themselves as being "overly neurotic" about evaluative information, acknowledging how their personal identity affected their analyses.

Interestingly, four of the five students identified specific items on which they believed they were graded incorrectly but acknowledged that the discrepancies did not have a significant enough impact on their overall performance for the student to investigate further. One student divulged, “I feel like I got some points that they didn’t count, but that’s okay . . .” further explaining the discrepancies might result from differences in interpretation of subjective interpersonal elements of the evaluation. Another area of discrepancy in scores resulted from the constraints of the testing environment. As one student explained, it was challenging to verbalize mental processes so that the rater was aware of the student’s intent:

> I felt pretty confident with [this topic] and it looks like indeed the comments [about my performance] were pretty good. Says I only asked one question about the throat and mouth just saying, “Any sore throat?” Oh and I remember that was because, myself, I knew that [the patient] had just said that she was having post-nasal drip. So I put that down as having something in the throat. There are always some things where you know what’s going on and you put it down but you forget to say it out loud.

While a video recording of the patient encounter was available for review, only two students elected to access the video. Both students did so for the purpose of checking to see if they had completed tasks for which they did not receive points. Finally, two students described how software/technology constraints hindered their use of evaluative information, specifically regarding having to be on campus to access the evaluative information.

When reflecting on the meaning of the evaluative information, two students focused on how the information indicated how far they had come since the start of the

program, two students focused on how the information highlighted areas for future improvement, and one student engaged in reflection on both the past and the future. The evaluative experience itself provided students with an opportunity to reflect on improvement over time, as one student explained:

> You just feel so much more confident [over time] when you are doing this. You feel like you made improvements. I feel like the first time when I felt like I couldn't ask anything. It's only been a year but I feel like oh gosh I was so awkward back then . . . not getting as nervous about little simple things shows a lot of how you've developed.

One student explained the immediate relevance of the information as a measure of the student's ability to perform a complete, appropriate, and accurate exam:

> First off, [the evaluative information tells me] that I hit every point on the checklist that we had to memorize and internalize and work on. That's the first thing that I look at just to make sure that I hit the points. Secondly, techniques, which are hard to grade in this setting, but there are gross aspects of technique that are pretty obvious from the camera: how to swing a hammer, or use a tongue depressor, or the sides of the body you are supposed to look at. So primarily it was making sure I had everything memorized and went through the checklist and secondarily that my technique was sound.

The same student continued by positioning the relevance of the information along the student's progression along the developmental continuum:

> I don't think I take anything grand from it, like, "Oh yeah, I've got it down, don't need to worry about it anymore…" but at least it is comforting that, where I am now, I haven't misunderstood anything. I'm going down this path through medicine and I'm pretty confident that, at least at this stage, what I am supposed to know, I know and I haven't misunderstood or misconstrued how to do something or my technique being off at this point. I understand it is not in the past yet but at least I know that I'm still on the path that I'm supposed to be on, if that makes sense.

This sentiment was echoed by another student who reflected:

> This really just tells me overall that I'm on track and on point with the exception of time and I've got to increase my clinical pace because I know from experience from [shadowing] and other clinical settings that if you take more than five minutes for anything then you are in trouble. And we have 45 minutes for this [evaluation].

Finally, one student expressed a desire for raters (both standardized patients and physicians) to recognize longitudinal improvement across evaluations. As the student articulated: "It's always interesting to see what their feedback is. If they remember you they'll usually say, 'Oh, you did much better this time than a while ago,' so I think that's more developmentally helpful."

Overall, findings from the reflective exercises tended to corroborate and illustrate the findings from the 24 stakeholder interviews, primarily regarding the first problematic, facilitating sensemaking processes. Student reflections were predicated on a notion of a developmental continuum as students actively positioned their current performance relative to their past performance. Interestingly, students did not utilize an appreciative lens to identify personal growth (i.e., by carrying forward identified areas of deficit from previous evaluations to look for improvement in the current evaluation); instead, they connected past and current performance based on a more macro-level conceptualization of their typical performance, and this conceptual use of evaluative information manifested as a deficit analytic lens as students identified specific areas for future development. Personal factors were apparent in students' reflections regarding their experiences of the evaluation and how this compared with evaluative information

reported to them. Initial reactions upon receipt of reports and after review of all feedback suggested that emotional responses to quantitative indicators influence perception and receptivity to feedback while analyzing reports. The design of the study did not permit analysis of how students may have consolidated evaluative information from this report into longitudinal notions of performance across courses longitudinally. Students' reflections evidenced how they balanced competing sources of information, mostly between their experiences and their reports and between comments made by various groups of raters (i.e., standardized patients or physicians). The design of the study did not permit analysis of how students analyzed their performance relative to peers or how program administrators used information from this evaluation. Similarly, the second and third problematics (engaging systemic complexity and attending to power and information gaps) were not directly addressed by student reflections on evaluative information.

Summary

In this chapter, three problematics were used as an organizational framework for relating my findings about evaluation use and influence in assessment-based educational evaluation. The problematics were developed through the analysis of a single CSE system that produces student learning evidence relative to a variety of standards and objectives for use across multiple levels of the educational organization. Illustrations of the problematics were made using case-specific descriptions of how stakeholders perceived the use and influence of evaluation, descriptions that constitute the uniqueness of the case. However, at a higher order of interpretation, these problematics can be

considered as representative of some typical tensions that result from the use of evaluation for educational accountability and learning.

The first problematic centers on the need for values-oriented sensemaking processes to address the interplay of meaning and values within the analysis of student learning evidence. Taken as an absolute measure of ability, student learning evidence provides an indication of the degree of proficiency students have demonstrated relative to specific objectives or standards. However, for this information to be actionable, program stakeholders must position the meaning of student learning evidence relative to considerations that are not necessarily indicated within the evaluative information. For example, the meaning of a sum score of 80% might vary, with some stakeholders considering this degree of achievement to be strong and others viewing it is an area for targeted improvement. In such circumstances, how do stakeholders perceive and mobilize information for accountability and learning at the student and program levels? What factors influence how stakeholders make determinations about what or how to improve, balance competing information, and select mechanisms for improvement? While the influence of stakeholder orientations, values, and positions is oftentimes acknowledged after differential interpretations of evaluative information emerge, consideration of these influences on sensemaking is rarely addressed proactively, meaningfully, or directly within educational program models. Rather, it is oftentimes assumed that program stakeholders can interpret the meaning of evaluative information as a nomothetic representation of reality, a position that creates space for *misinterpretation* and

misunderstanding. In the educational context, this dynamic also fosters *missed opportunities* for student- and program-level learning.

The second problematic centers on the contextual complexity of teaching, learning, and evaluation as each relates to the use of student learning evidence. The intended program model exists at the intersection of many policies and a variety of stakeholder groups, all of which can exert influence on how evaluation is used. Inter-relationships between these elements produce a dynamic tension in which change to any one element may produce ripples throughout the system. Inter-relationships, and their relative importance, are perceived differently by various groups of program stakeholders, each of whom have unique stakes and perspectives and who draw relevance boundaries differentially. How can evaluators collaborate with educators to generate evaluative evidence that addresses a variety of stakeholder groups? What tensions and challenges emerge from attempting to use a consolidated source of evidence as a simultaneous indicator for many purposes and across many levels of the educational organization? While it is rarely possible to construct multi-site standardized measures that serve multiple purposes (Koch, 2013), it is possible to design site-specific sources of evaluative evidence that may be translated relative multiple purposes. The key to this effort is a focus on contextualizing evaluation through intentionally supporting sensemaking processes: ensuring tight alignment of purposes and indicators can reduce the likelihood of misuse of evaluation.

The third problematic is centered on the notion of alignment, first in the design of the program model and its evaluative tools to attend to power and information gaps and

second, in the responsive facilitation of the program model to engage the inherent complexity of education and educational evaluation *in vivo* (i.e., addressing discrepancies between the intended program model and the lived experience of the program model). Educational programming tends to be a rational, information-generating process in which stakeholders may or may not be aware of the model on which the program is predicated. Tools developed by evaluators like logic models and program theories can be useful for helping stakeholders understand and communicate the basic premises and goals of their programs. Additionally, stakeholders need help anticipating and responding to evaluative findings, an effort that requires consideration and management of emergent, in-the-moment factors that influence the meaning, relevance, and use of evaluative evidence. Stated another way, responsive facilitation of the program by its facilitators requires awareness of areas of how misalignment between the intended and actual program models influences the accuracy or appropriateness of evaluative tools and their results. To ensure the trustworthiness and credibility of evaluative information, stakeholders may need to tweak evaluative mechanisms while the program is underway. How are evaluative mechanisms designed and modified in response to "lived experience" of the program to generate evaluative evidence of student learning for use at the student and program levels? What are the inputs, activities, outputs, and outcomes (manifest in the standards of quality, curricular deliverables, rating schemes, and experiential milestones that substantiate the program model) with which student learning evidence must be aligned in its generation, interpretation, and use? How do differences in the intended and actual program model influence the meaning and relevance of student learning evidence?

While attending to information and power gaps in the design and implementation of the content of the program model is a responsibility that falls to educators, evaluators can call on a wide variety of evaluative tools (i.e., logic models, evaluation frameworks, and program theories) to integrate and improve the quality of the program model utilizing the evidence it generates. Addressing this practical problem facing program stakeholders in educational contexts poses a particularly fruitful opportunity for evaluators: educational program models necessarily produce student learning evidence in response to a plethora of malleable factors, including standards of quality, curricular frameworks, teaching and learning strategies, technology, and measurement methods.

CHAPTER V

DISCUSSION AND CONCLUSIONS

The purpose of this case study is to describe the uses and influences of evaluative information on educational accountability and learning at the student and program levels for an innovative clinical skills examination system for first- and second-year medical students. By applying research knowledge of evaluation constructs in a relatively new practical context for evaluation, I sought to characterize the types of use (process and findings use, further categorized as instrumental, conceptual, and/or symbolic/ legitimative), the purposes for use (accountability and learning), and the effects of considering the notion of evaluation use and influence on people and situations in assessment-based educational evaluation. As indicated in the literature, program stakeholders in the case study saw distinctions between accountability and improvement efforts that were self-selected by internal program stakeholders versus those that were imposed externally. While this perceived distinction was noted and tracked throughout the course of this analysis, the use of evaluation for learning and improvement is considered in this discussion regardless of whether it was attributed specifically to assessment structures.

The nature of evaluation use in assessment-based educational evaluation was organized around three problematics: facilitating sensemaking processes, engaging systemic complexity, and attending to power and information gaps. These problematics

suggest the nature of some formidable obstacles to the use of evaluative information for learning and improvement. In this chapter, I will discuss how evaluators working in assessment-based educational evaluation contexts, or educational professionals responsible for assessment, may begin to address these problematics in the program models and evaluative tools they craft. These efforts are rooted in a need to analyze the use and influence of evaluation on two levels: within the program model and on the program model as it is refined and developed through time. This chapter begins by analyzing each of the three problematics regarding the notions of evaluation use and influence. This chapter ends with some conclusions about evaluation use and influence in assessment-based educational evaluation based on the research questions.

Three Problematics for Assessment-Based Educational Evaluation

In the field of evaluation, much research has been conducted to probe the nature of evaluation use. Use can be decomposed into process or findings use, with the latter suggesting the results of an evaluation inform decisions and the former suggesting the evaluation process itself exerts influence on people and situations. Regarding process and findings use, instrumental use occurs when direct use of the evaluation or its products exerts influence, conceptual use occurs when the evaluation influences how stakeholders think, and symbolic (process) and legitimative (findings) use occurs when evaluation fails to exert influence or is only used to support past endeavors. Beyond these theoretical frameworks for conceptualizing use and influence, much study has been focused on understanding the notions in applied domains and practical contexts of evaluation (e.g., educational or health care settings, organizational contexts for internal evaluation use).

The remainder of this section briefly connects the three problematics facing assessment-based educational evaluation with this broader literature.

Problematic 1: Facilitating sensemaking processes. Unsurprisingly, in this case study, revelatory processes (typically characterized as 'interpretation of results' in educational settings and treated as 'sensemaking' in the context of evaluation) varied by stakeholder in myriad ways that were challenging to address in a meaningful way. These factors are described as "personal factors" and warrant additional study. Two key findings regarding the factors affecting evaluative information are offered: trust and credibility of the program leadership and integration of specific mechanisms for supporting sensemaking of evaluative information within the program model. Interestingly, students and administrators alike seemed to express a need for confidence in the quality of the program and for trust in the program leadership as factors influencing their reception of evaluative information. Students suggested that they were more receptive to "jumping through hoops" because they perceived the program as a strength of the medical school and because they trusted the program leadership to hold them accountable while supporting their learning. Students furthermore suggested that they felt positive about being held accountable for strong performance and that they trusted the program to provide them with the best quality of training available. This allowed students to overlook perceived deficits of the program and to assume the best of program stakeholders, even when expressing frustration or dissatisfaction. The extent to which the program was organized and accessible to the students, including their ability to understand the major learning goals and expected outcomes for each course and each

evaluation, seemed to bolster student satisfaction with the model, suggesting that a rational, communicable program model is useful for garnering student buy-in. Similar belief in the quality and credibility of the program and its leaders was expressed by program administrators, who communicated being impressed with the thoughtful and artistic integration of accountability expectations within a learning framework. This suggests that, to some extent, personal factors (such as perceptions of and relationships with leadership) influence the extent to which evaluative information is received and used. In short, how program leadership frames and positions evaluative information (for accountability, for learning) seems to influence how stakeholders perceive its utility and relevance.

Evaluation can be described as a professional field focused on the "assisted sensemaking" (Julnes, 2012) of information based on contextual considerations. By approaching assessment from the same perspective, the meaning of assessment outcomes can more easily and directly be negotiated with the value and practical implications of outcomes. The need for considering value-oriented dimensions of evaluation (e.g., organizational, political, etc.), is well-documented and existing at multiple grain sizes of the organizational context. In considering individual measures, or indicators, Dahler-Larsen (2014) advocates for a shift in focus away from understanding 'unintended consequences' toward 'constitutive effects' (i.e., effects resulting from the use of indicators in organizational procedures that provide a shared language and through which dis/incentives are attached) that are affected by contextual factors and stakeholder responses. This shift places more emphasis on the actual mechanisms producing

evaluative information than on the intended or theoretical indicators (and the desired outcomes). An important implication of focusing on constitutive effects of indicators is to realize that the indicators themselves define and constitute the social reality they are meant to measure (i.e., indicators define interpretive frames and world views, content, time frames, social relations and identities, and their own meaning based on their practical use/implications; p. 977).

At a more macro-level, Hammersley (2013) emphasizes that practice informed by evidence "necessarily involves the exercise of interpretation and judgment, rather than simply the 'application' of research findings" (p. 8), an activity that he characterizes as involving reasoning as well as consideration of utility, values, emotions, and political dynamics. These dimensions are held in tension in light of "multiple values, tacit judgment, local knowledge, and skills" (p. 20), tensions that force stakeholders to make judgments by deferring either to instinct (i.e., based on one's experience and professional judgment) or the evidence (i.e., assuming research is "always sounder than other sources," p. 22). At the organizational level, Weick (2001) delineates six themes that suggest people and situations involved in organizational contexts cause the organization to function less rationally than is oftentimes assumed, including the need to see organizations as an amalgam of smaller factional, inter-related units of stability across which connections may be differentially strong. At the most macro level, Bogenschneider and Corbett (2010) describe the tenuous relationship between policy and evidence as being mediated by variations in types of policy questions, knowledge consumers,

knowledge producers, research strategies, results, and interpretations, interactions between knowledge producers and consumers, and complexity.

The need for sensemaking in educational organizations is not a new notion. Spillane and Miele (2007) position sensemaking as a means for recognizing how revelatory processes (i.e., attention, interpretation, schemas, mental models, accommodation or assimilation, expertise, and bias) are affected by situation and context. These elements are codified within local practices as organizational routines and tools that mediate evidence use, as described by Spillane and Miele:

> But to understand evidence use, we must attend to practice, which necessitates attention to interactions among people, as well as to how these interactions are mediated by aspects of the situation (such as organizational routines and tools). A simple first step for school leaders might involve taking stock of the formal and informal organizational routines in their schools. A next and somewhat more complex step might involve asking some tough design questions about these organizational routines—what purpose do they serve? How should they work in order to achieve this purpose? How should they not work? A third and substantially more complex step involves analyzing the interactions in the performance of these routines— attending not only to the people involved, but also to how aspects of the situation frame and focus their interactions. What tools are in use, and how do they frame the interactions? How might these tools be redesigned, or new tools developed, to frame and focus the interactions in new ways? (p. 68)

This study followed a similar process of taking stock of how people and situations are affected by evaluative information in alignment with somewhat disparate purposes. As program administrators involved in assessment build connections between the program and its evaluative tools, they may leverage organizational systems and structures to facilitate sensemaking processes in order to drive the use of evaluation for learning purposes. Dahler-Larsen (2012) offers multiple organizational models for evaluation: the

rational model in which evaluation involves instrumental use and summative decisions, the learning model in which evaluation centers on organizational learning, and the institutional model in which evaluation is ritualized and mediated by organizational and societal factors. Interestingly, student learning evidence may ideally function on different models at different levels of the organization: a rational model might be most appropriate for student-level use of evaluative information to ensure results accurately and appropriately inform development relative to performance standards while a learning or institutional model may be more appropriate to target program-level improvement. Negotiation of meaning and values across these levels is necessary to leverage evaluative information for "offensive, experimenting, and future-oriented views of quality" (Dahler-Larsen, 2012, p. 229).

Problematic 2: Engaging systemic complexity. The need to address contextual complexity is a pervasive theme with the education and evaluation literatures. Findings from this case study serve to probe the nature of that complexity and to use complexity-oriented evaluation approaches and tools to discern overarching themes and salient aspects of these intricate educational contexts. Paying particular attention to primary intended use by primary intended users (Patton, 2005) is cited as a guiding principle for reducing complexity, both in evaluation (e.g., utilization-focused evaluation) and in assessment (Ikenberry & Kuh, 2015; Jonson et al., 2017). This focus, however, varies based on the 'commissioners' of an evaluative effort and can be distinctive from a focus on practical utility, balancing the interests of multiple stakeholder groups, and/or addressing the larger 'social good.' When evaluation is put in place by external

stakeholder groups, a tension between accountability and learning purposes can emerge. In the case of education, accrediting bodies have leveraged this position (i.e., as the primary audience of assessment outcomes) to encourage assessment use for institutional learning. As assessment systems and structures continue to shift toward improvement, evaluation use and influence continues to focus more squarely on program beneficiaries (i.e., students). As Bogenschneider and Corbett (2010) warn, "there are many hands involved in determining what shapes the experience of the consumer . . . to exclude any of these actors is to fail to recognize the inherent complexity of the policymaking process" (p. 18). When considering policy and implementation, these degrees of complexity are further amplified by the varying evaluation purposes (i.e., accountability and learning) across stakeholder groups.

Much research on evaluation is engaged in increasing the quality of the information it produces by emphasizing its credibility and utility. A recent volume on the nature of evaluative evidence calls for an expanded notion of credibility (e.g., beyond methodological considerations) as well as for a focus on actionability (Donaldson et al., 2015). In effort to emphasize the importance of utility and feasibility, Chen and Garbe (2011) have suggested the term "viable validity" to consider the extent to which a program or intervention is "practical, affordable, suitable, evaluable, and helpful" (p. 100). In bringing attention to this concept, the authors elevate the importance of attending to those implementation issues that affect how the program achieves its primary purpose of serving program beneficiaries' needs. Similarly, Williams and van't Hof (2016) approach the reduction of stakeholder complexity as one of 'making victims,' a

description that acknowledges some stakeholders' interests, questions, and positions may be affected negatively or marginalized by a program. Victimization is necessary given that an "evaluation study cannot cover all aspects of the program, and it can never be the only basis on which decisions are made" (Weiss, 1988, p. 17). By definition, however, victims must be external to a program, meaning that program beneficiaries or facilitators (e.g., students, teachers, or program administrators) should never be victimized. If a program is taken as the primary stakeholder, program beneficiaries' needs and uses of information should be prioritized since they are an integral aspect of the program. Systems-oriented approaches to evaluation use inter-relationships, perspectives, and relevance boundaries to reduce complexity, tools that appear well-discussed in the organizational and management/business literature, offering the opportunity to cross-pollinate these ideas within the assessment context. Such efforts could simplify and communicate prioritization decisions in ways that might support the streamlining and codification of more salient elements of educational programs as well as the identification of areas in which adaptability and flexibility are needed to respond to the shifting nature of education and educational evaluation.

Problematic 3: Attending to power and information gaps. Broadly speaking, evaluation efforts tend to focus on creating alignment between program goals and measures so that evaluators can capably judge the merit, worth, or significance of an intervention toward its priorities and anticipated outcomes. Many tools and resources exist to support evaluators in eliciting this information from program stakeholders (e.g., logic models, evaluation frameworks, program theories). Evaluation is distinctive from

research in this regard since it is a process steeped in the consideration of such issues as culture, context, values, communication, and/or power in establishing the foreground and background (i.e., what is signal and what is noise) in the development of evaluation questions, outcomes, indicators, and so forth. In its process and findings, an evaluation is indubitably a reflection of how the evaluator chooses to respond to these factors, and how the evaluator approaches each of these factors is mediated by power dynamics that mediate asymmetric relationships between stakeholder groups (Baur, Van Elteren, Nierse, & Abma, 2010). The issue of power is particularly relevant as evaluation continues toward a trend of 'new public management' in which transparency is essential for sharing with a wider audience (e.g., program stakeholders, politicians, the general public) the indicators being employed to judge a program and areas in which it might improve (Hammersley, 2013). For assessment, this means that evaluators need to communicate with a broad audience (students, parents, internal and external program stakeholders, accrediting bodies, politicians, etc.) not only regarding findings and their relevance, but also in terms of how indicators were designed or selected and how other 'peripheral factors' (Lin Miller, 2015) were addressed in these indicators of quality. Another issue of power, then, is the tension between judgments based on professional expertise and that of evidence (Hammersley, 2013), or some combination of expertise and evidence, since evaluative information, like all measurement, is an imperfect reflection of social reality and since sensemaking is an idiographic process that may result in a diversity of opinions regarding future action.

Information gaps exist when unidirectional or symmetric relationships are assumed. As Spillane (2002) emphasizes, those responsible for upholding mandates "must decipher what a policy means to decide whether and how to ignore, adapt, or adopt policy proposals into local policies and practices" (p. 378). While it is oftentimes assumed that program stakeholders need merely to understand and implement mandated standards, in reality, program stakeholders must make sense of standards, determine how these mandates fit within the local, pre-existing program context, and choose how to respond to each mandate individually. This sentiment is echoed by Bogenschneider and Corbett (2010) when they argue that "discretion at the operational level is a powerful tool for doing a lot more than merely carrying out what others have determined" (p. 18). Norris and Kushner (2007) suggest three conditions, or gaps, in information toward which evaluation might be useful to bridge degrees of separation. First, since information communicated from external stakeholder groups (i.e., standards and objectives shared by accrediting bodies, professional organizations, or government entities) cannot dictate every essential element of a program needed to ensure goals are met nor can it anticipate every possible outcome of the frameworks mandated to shape the program, evaluation can support program stakeholders in moving beyond compliance toward meaningful integration and ownership over mandated elements of programs. Second, since sites in which standards are taken up are distinctive regarding their potential and actual program effectiveness relative to those standards, evaluation can serve to "increase 'intelligence' and 'even up' the distribution of information" (p. 8, emphasis in original). Third, since

"evaluation is thought to increase or promote trust," (p. 8) evaluation can serve to connect and hold accountable both parties involved in standards-based interventions.

Connections to the assessment context. In the concluding chapter of a volume on using student learning evidence composed by many contemporary leaders of the assessment field (Kuh et al., 2015a), it is suggested that professionals engaged in assessment-based educational evaluation need to consider evidence use in light of three main concerns: transitioning from a culture of compliance to a culture of assessment, focusing on primary stakeholders' information needs and uses of evidence, and using evidence to improve teaching, learning, "and the coherence of the student experience" (p. 221). Acknowledging these themes, the authors suggest five key principles for assessment practitioners: focus on the relevance, meaningfulness, and utility of evidence; employ a variety of formats to communicate with multiple audiences in a way that clearly suggests future action; embed assessment in typical teaching and learning processes; involve key stakeholders in assessment efforts; and prioritize campus-based information needs within larger frameworks of information needs. To engage these principles, they offer seven strategies: embrace accountability, think of end users at the outset, organize assessment work to respond to high priority questions, share widely and transparently, lead rather than manage, look behind demands, and focus but adapt (through methodological pluralism).

The problematics that emerged from this case study resonate with these themes, principles, and strategies while concurrently deepening their consideration through anecdotal illustration. The case study provides one example of how an educational

program has leveraged student learning evidence by incorporating accountability into its pre-existing framework of learning, sharing evaluative information with all internal and external stakeholder groups, aligning assessment mechanisms with prioritized standards and objectives, owning the use of evaluative information, and utilizing multiple evaluation methods to contextualize outcomes. The case study shows that, given the iterative and evolving nature of education and assessment, these efforts may never be considered satisfied, but must be addressed anew with each cohort of learners and ever-changing expectations. To these efforts, evaluative tools (i.e., logic models, program theories, evaluation frameworks, etc.) and evaluative thinking can be useful; within these efforts, evaluation researchers can delve deeper into the factors affecting the uses and influences of evaluation.

Conclusions

Inquiry into the uses and influences of this evaluative information was guided by the following research questions:

Research Question 1. Based on a specific case of an innovative clinical skills examination system (CSE system) in a medical school, what is the nature of evaluation use and influence in assessment-based educational evaluation?

1.1 What types of use (instrumental, conceptual, symbolic, process, etc.) are made of student learning evidence within the educational organization?

1.2 To what extent are uses oriented toward accountability and learning?

1.3 How are these findings affected by expanding the notion of evaluation use to consider dimensions of evaluation influence?

In the remainder of this section, each of the sub-questions (1.1-1.3) are addressed, followed by a discussion of the overarching research question.

Research Question 1.1: Types of evaluation use. Research on evaluation has suggested that use is decomposed into process or findings use, with the latter suggesting the results of an evaluation inform decisions and the former suggesting the evaluation process itself exerts influence on people and situations. For each of these types, use may be considered instrumental, conceptual, or symbolic/legitimative: instrumental use occurs when direct use of the evaluation or its products exerts influence; conceptual use occurs when the evaluation influences how stakeholders think, and symbolic (process) and legitimative (findings) use occurs when evaluation fails to exert influence or is only used to support past endeavors. Each

Findings use. Educational evaluation through course-embedded tests, projects, portfolios, presentations, etc., is predicated on the use of evaluative information. Standards-based teaching and learning necessitates the generation, interpretation, and use of evaluative information for accountability and learning purposes at the student and program levels. From this vantage point of education, acquisition of a degree implies not only that the student has completed an appropriate course of study, but that the student has done so by demonstrating minimal competency on all key standards and objectives. Likewise, accreditation of educational programs implies not only that program inputs created situations in which students could learn, but also that the program is leveraging evaluative information to make programmatic improvements (i.e., to the program model and its evaluative tools) that will better support future students. Shifting to a learning-

centered paradigm (Barr & Tagg, 1995) positions the use of evaluative information as instrumental: when anchored on knowledge, skills, or mindsets, evaluative information suggests not only that performance might be unsatisfactory, but the specific elements relative to which performance needs to be improved.

Process use. A key finding from the study of evaluative information in educational contexts suggests that the process of taking a test, creating a portfolio, making a presentation, etc., influences how stakeholders interact with findings. Almost all stakeholders interviewed suggested that misalignment between expected and actual findings served as a "red flag" for possible failures in teaching or misevaluation. These discrepancies, resulting from misalignment between process-based expectations and findings-based implications, create challenges for educators and frustrate or confuse students, endorsing these areas for development within the program model and its evaluative tools through assessment-based educational evaluation. They may also suggest areas for development in the longitudinal program model in ensuring that courses are aligned well enough with one another in terms of their standards and objectives. To this end, the developmental continuum anchor provides a useful tool, or scaffold, onto which these connections and alignments may be organized, mapped, and reconfigured. Furthermore, the length of time between the evaluation experience and the reporting of its evaluative information influenced students' perceptions and emotions, suggesting that systems and structures in which evaluation is embedded need to be planned and elaborated meaningfully to encourage reception of evaluative information.

Instrumental use. A common theme in the evaluation literature is a concern for the lack of instrumental use of evaluative information (Peck & Gorzalski, 2009; Preskill, 2008; Stitt-Bergh, 2016). In educational contexts, by contrast, much instrumental use is assumed of evaluative information, specifically for grading and promotion/retention decisions. Evaluative information is typically quantified using points and percentiles, then combined into grades at the course level to provide an overall indication of student performance. These course-level grades are used to substantiate decisions at the student, course, and program levels. Accumulated success in passing courses indicates that students have adequately learned the program of study; high or low rates of student passing indicate the general success or failure of a given course; and program improvement is targeted toward the development of courses with low passing rates or perceptual survey outcomes. In this way, the use of evaluative information can be considered instrumental: evaluative outcomes determine decisions that should be made at the student level and aggregate evaluative outcomes suggest areas for future improvement at the course and program levels. However, it may also be argued that such use is merely legitimative if the evaluative information does not have a meaningful substantive anchor: there is a substantial difference in the use of evaluative information to provide evidence of how student performance is graded when compared with the use of that same information for learning, as elaborated in the following paragraph.

Conceptual use. Findings from this case study seemed to support a marked use of evaluation by members of all stakeholder groups that was conceptual in nature. Program administrators and students alike described how evaluations changed their ways of

thinking about teaching and learning as well as about individual-, group-, and cohort-level performance. Findings from this case study suggest that conceptual use of evaluative information is a key area for development in training programs. The way that students are introduced to the evaluation process and trained to use its information is oftentimes overlooked in the development of educational programs, either because the information is viewed as being straightforward, objective, and nomothetic or because the information is not seen as useful for learning, instead constituting evidence for grading decisions that are absolute and final without time for improvement. Stakeholders thus make sense of the substantive meaning of evaluative information differentially based on their varying perspectives and relevance boundaries, and these differential interpretations are rarely unearthed or acknowledged. Issues of power affected conceptual process and findings use, with administrators exerting considerable effort to create and maintain alignment within a consolidated program model and its evaluative tools in response to shifting expectations imposed on the program, and with students sometimes artificially constraining their learning based on value judgments regarding the utility of the training for their current and/or long-term careers. This poses a challenge to learning since students are not well-positioned to understand what knowledge and skills might be relevant for their future use of their training.

Symbolic use. Both students and administrators described instances in which evaluative processes and findings failed to achieve their intended purposes of learning or improvement. Students described developing personal action plans without intention to follow through on defined next steps. They also shared instances in which the evaluation

experience or findings seemed to be a process of "checking boxes," "jumping through hoops," or performing actions in an artificial, simulated environment that could not always mimic reality. Program administrators described having to give grades based on student learning evidence as well as having to use evaluative information as evidence for decisions regarding promotion or retention. Interestingly, stakeholder perspectives and how they constructed relevance boundaries seemed to affect use: while some students perceived the generation of an improvement plan as symbolic, others utilized this opportunity to use evaluative information either instrumentally or conceptually by implementing a personal improvement plan. This suggest that stakeholder buy-in affects evaluation use. Overall, it is unlikely that symbolic uses can be avoided entirely in educational evaluation, an issue that is discussed further in the next section regarding accountability and learning purposes for educational evaluation.

Research Question 1.2: Purposes for evaluation use. Regarding the purposes for evaluation, the case provides an example of a program model that successful integrated accountability purposes within a learning framework. While accountability was positioned as secondary to learning, accountability was perceived as valuable and important and was, as such, integrated meaningfully within the program model. Ownership over accountability purposes seemed an inherently important element as did the degree of integration and internalization of the expectations: elements of the program model that were perceived as only useful for the sake of accountability were identified as such and seemed to be somewhat less appreciated by students and administrators alike. While it is impossible to guess how a lack of accountability requirements might affect

program stakeholders, it seems that the presence of these expectations calms questions about the appropriateness or sufficiency of the program model that might otherwise affect stakeholders' confidence in the model. Accordingly, communicating the origination of the various elements of the program model and explaining their relevance to students seems a key aspect of developing a strong baseline of relevance for evaluative tools used within a program. Indeed, students seemed to perceive the elements of the CSE system that were seemingly disconnected from the expectations of a physician's daily practice the most problematic or unfair.

One unexpected finding from this case study was an emphasis on the need for accountability at the student level. Prior to the start of the study, I envisioned evaluation use without clearly addressing the separation of accountability and learning purposes at the student level. While accountability is oftentimes associated with a sort of imposed use (Weiss et al., 2005) from external stakeholders or accreditation agencies, program administrators described the need to hold students accountable for their learning at regular intervals throughout the training program. Interviews with both administrators and students emphasized the need to separate these purposes when considering the student experience while suggesting that the tension between accountability and learning in how information is used to make decisions about individual students is an issue of high importance (Chouinard & Cousins, 2015; Chouinard & Milley, 2015). Through the course of the study, I began to conceptualize the dimensions of evaluation use from the perspective of internal program stakeholders around two key particularities of the assessment context: an explicit focus on learning (Rickards & Stitt-Bergh, 2016b) and an

increasing focus on leveraging evidence from course-embedded assessments for multiple purposes (Kuh et al., 2015b). From this viewpoint, use occurs at a minimum of two distinctive levels within the educational organization: the student level and the program level. At each level, the driving purpose for evaluation can be considered as informing accountability or learning purposes, resulting in a quadrant of purposes for student learning evidence, as shown in Table 3 with typical targets for the use of student learning evidence at each purpose-level intersection.

Table 3

A Quadrant of Purpose and Use

	Accountability	Learning
Program	Documenting Goals/Outcomes Reporting Findings 'Up & Out'	Elaboration of Evaluative Tools Evolution of Program Model
Student	Evidence of Competency Grading/Promotion/Retention	Analysis of Learning Mastery Extension/Amelioration/Remediation
	"What"	"So What"

By drawing out such distinctions, this depiction of the dimensions of purpose is particularly relevant for evaluators to consider while collaborating with program stakeholders. By understanding how stakeholders interact with student learning evidence within these four dimensions, evaluators are better equipped to close the perceived gap between the use of evaluative information for "us vs. them." According to this framework, mechanisms of accountability are subsumed within larger learning goals: that

is, the mechanistic processes and outcomes undertaken to satisfy accountability purposes, while distinct, are natural steps toward facilitating learning purposes.

Finally, and perhaps expectantly, the use of evaluation for learning was differential across formative and summative evaluations. Formative evaluation is predicated on the notion of learning and improvement, while summative evaluation determines the degree to which learning goals were met without aspirations of learning or improvement. Students and program administrators described formative and summative evaluations in alignment with these descriptions. In addition to emphasizing the need for alignment between formative and summative evaluations (e.g., in evaluative tools, the evaluative setting, etc.), program stakeholders highlighted how summative evaluations could be used for learning purposes for students who performed below expectations. Through remediation and amelioration, students were provided with multiple opportunities to demonstrate proficiency even after a summative evaluation. This was achieved by integrating responsiveness to evaluative information into the program model: after summative evaluation, students who demonstrated unacceptable levels of proficiency received additional instruction that targeted the measured areas of weakness and participated in a new, yet aligned, evaluation to retest their proficiency following remediation. Students with borderline proficiency (i.e., passing according to a quantitative measure, but with significant areas for development) also received additional instruction and were re-evaluated. In this way, applying the learning-centered paradigm (Barr & Tagg, 1995) positioned summative evaluation to satisfy both an accountability

and a learning purpose. In this way, the degree of proficiency mediated the purpose of the summative evaluation.

Research Question 1.3: Evaluation influence. In her seminal work on evaluation influence, Kirkhart (2000) suggests expanding the notion of evaluation use to one of influence by analyzing use relative to three dimensions: source (process, findings), intention (un/intentional), and time (immediate, end-of-cycle, long-term). Kirkhart asserts that consideration of use types along these additional dimensions overcomes issues of construct underrepresentation, offering a more nuanced understanding of how people and situations interact with information. Alkin and Taut (2003) suggest that the intention dimension might be analyzed relative to awareness, resulting in three types of intention/awareness: intended and aware, unintended and aware, and unintended and unaware. Positioning types of use along these aspects partitions evaluation influence into 18 distinct aspects, some of which can be attributed to use (i.e., all findings or process uses that occur in the immediate or end-of-cycle and are either intended/aware or unintended/aware) or influence (i.e., all findings or process uses that occur in the long-term or are unintended/unaware). Consideration of these aspects was integrated into the current study: process and findings use were key targets in the analysis; the interview sample was large (30 interviewees) and diverse (internal and external administrators, current and former students of the training program) in effort to research intention and awareness; and interviews prompted discussions of temporality as much as possible.

Since the research study was conducted within a short time period, it is not possible to fully address the immediate, end-of-cycle, or long-term effects of a single

evaluation; however, respondents discussed evaluative information in terms of immediate, end-of-cycle, and long-term influence. From these findings, it seems that program models must include defined methods for soliciting feedback on evaluation use and influence in order to become aware of unintentional and possibly undesirable consequences. It also appears that long-term influence of evaluative information constitutes an area that has not yet been addressed within the program model. Administrators and students alike identified a need for "un-siloed" analysis of trends over time, efforts that might be facilitated by developing an evaluative framework focused on the key knowledge, skills, and mindsets that the program seeks to develop within each of its students. Liberation of evaluative information from "one-off" reports to longitudinal consideration of information across evaluation instances is not novel in assessment-based educational evaluation, rather constituting an area that has been identified for development within internal evaluation units in many professional sectors (Loud & Mayne, 2014). This may be achieved through the consolidation of evaluative information into a central database, through the creation of a developmental scale for reporting student performance, or through the use of dashboards for reporting student performance on key skills or topical areas across evaluations.

Influence focuses largely on unintended outcomes of evaluation that are unknown to those responsible for a program and/or that occur in the long-term. Accordingly, a focus on evaluation influence calls for efforts to identify unintended/unaware and long-term effects. As was already taking place in the case study, those stakeholders responsible for assessment (i.e., evaluators or educational practitioners responsible for

evaluation) can utilize interviews, focus groups, perceptual surveys, and other methods of data collection to unearth unintended/unaware effects. These professionals are also well-positioned to utilize surveys or interviews to learn from program graduates, who have graduated from medical training programs and joined the physician workforce, about the long-term effects of their training program (including its evaluative experiences). While these mechanisms are useful for further exploring the influence of evaluation within the program, doing so increases the quantity of information being utilized to refine and develop the program model and its evaluative tools.

Some educational contexts suffer from what is known as DRIP syndrome, or being "data rich, information poor" (Goodwin, 1996). DRIP syndrome occurs from the use of many sources of evaluative information without meaningful consolidation of its utility to the overarching purposes for evaluation: accountability and learning. When attempting to satisfy a variety of evaluation requirements, programs may stretch resources thin to generate all of the mandated evidence while failing to ensure the evaluative information is useful toward the primary purposes of the program. To this end, evaluators working in assessment-based educational evaluation have many tools and resources to offer. Thoughtful integration of evidence gathering mechanisms can reduce the burden of assessment on educators and extend the scope of their efforts through prioritization and the subsequent maximization of assessment toward the most important and meaningful directions for program improvement. Ultimately, the degree to which these processes are internalized within the normal operations of the program seems of great importance. As seen in the case study, much evidence-based program improvement

was undertaken as part of the normal operation of the program, none of which was "counted" toward the efforts mandated within accountability systems and structures within the educational organization. While it seems intuitive that continuous improvement efforts need only be codified and recorded to satisfy accreditation requirements, this might not be the case: impromptu improvement work tends to take place during low resource, high urgency times within the academic semester, prohibiting the ability of educators to commit time to designing and implementing thoughtfully planned improvement efforts. The degree of separation between immediate improvement efforts and more substantial interventions is interesting and merits further investigation.

The use of rational, well-organized program models within which evaluative tools have been integrated is not necessarily a new idea in educational evaluation. Within the practitioner community, the need to "unpack" and "backwards plan" mechanisms of gathering student learning evidence are frequently the focus of educator-oriented professional development. These efforts are typically limited in their scope, however, by the relevance boundaries established by the educators designing educational programs. Establishing relevance boundaries involves understanding curricular and educational standards (quality frameworks and information use protocols) at various tiers of educational organizations, an understanding that can only be fostered through increased communication and collaboration across the various tiers of stakeholders and their information needs. To this end, evaluative tools and frameworks may be quite helpful. Treating external standard sets as inputs, cross walking various sets of standards, and consolidating value-based decisions in a single logic model or program theory could

provide a powerful tool for communication and collaboration. The creation of these tools is, however, not the primary target; rather, the use of these tools throughout the life of a program—in its design, implementation, and improvement—can serve as a key lever for learning and improvement. A rational and responsive program model may be capable of engendering more instrumental use of evaluative information toward learning and improvement if educators and evaluators can work together to build programs and evaluative tools that identify what information will be generated, what relevance it has for learning and improvement, and what specific follow-up actions should result from various types of performance. Logic models and program theories may help educators address complexity and navigate "DRIP" experiences by supporting decision making based on a pre-determined framework of relevance and importance, reducing the mental drain that results from reactive decision making. In this way, relevance boundaries may be drawn based on the program model as a means of reducing complexity and negotiating a diversity of expectations and information needs.

Research Question 1: Evaluation use and influence in assessment. This study provided the opportunity to consider educational assessment and testing from the perspective of evaluation, specifically with regard to the potential impact that professional evaluators may have in attending to evaluation use and influence in this context. Attention to the ways in which evaluation affects people and situations positions evaluators to engage with a deeper understanding of the nature of educational evaluation. This more nuanced understanding contributes to ongoing discussions in the field of evaluation regarding types of use, the tensions between use for accountability or learning,

and the value of considering use more holistically by analyzing the dimensionality of its influence. To these ends, I offer five key considerations regarding evaluation use and influence in assessment-based educational evaluation.

First, through the course of analyzing this rich dataset, I began to recognize a theme implicit within the interviews and reflective analyses: stakeholders expressed a strong degree of trust and confidence in the leadership of the program. This trust and confidence seemed to stem from stakeholders' perceptions of the key program administrator as exceedingly competent, not only in the content area (i.e., the professional practice of medicine) but as a responsive and cogent educator. Students remarked that the administrator "knows what she's doing" and that the administrator "had very high expectations of us, and that was good." An administrator attributed the success of the program to its leadership, describing the training program as a "longitudinal program that is so beautifully crafted and she's always adding to it. It's not only developmentally sound, it is right on target." I believe this sense of confidence in the program leadership is impacted by the administrator's sense of absolute ownership over the program. As standards and objectives are mandated and suggestions are made about program changes or developments by external stakeholders, program administrators work these changes into the operational program model so that these elements are not externalized or seen as "extra." While this may seem a small nuance, it is possible that this sense of ownership for the program exerts a meaningful influence on how stakeholders perceive of the training program and its improvement using student learning evidence and other evaluative information over time. Stakeholders' feelings about evaluative experiences (a

conceptual process use) seemed to impact stakeholders' degree of instrumental findings use. Stated differently, stakeholders seemed to develop opinions about evaluations based on their experiences, and these perceptions seemed to impact the extent to which evaluative information was later utilized, suggesting that stakeholders' degree of receptivity and trust in the quality and utility of information may have affected its use. For evaluators, this implies a need to ensure that evaluative tools are well-received and "owned" by stakeholders responsible for programs, a notion strongly advanced in the assessment literature (Kuh et al., 2015b). It may also suggest the need for evaluative evidence about the evaluation environment to supplement student learning evidence, efforts to which evaluators can contribute meaningfully.

Second, standards-based education is one context of objectives-based evaluation. Though the field of evaluation is replete with objectives-oriented approaches to evaluation, using these approaches in assessment-based educational evaluation requires recognition of a few key features of the educational context. To begin with, standards and objectives for a given class, course, program, or degree are positioned relative to a larger learning framework: present learning connects past performance with future expectations. Since past performance and future expectations affect the efficacy of the program model and its evaluative tools, evaluators can profit from understanding this relative positioning of objectives within a larger spectrum of learning. Evaluative findings from past learning experiences may be useful sources of evidence, and alignment of current evaluative efforts with future evaluative expectations might be useful areas for development. In addition to needing to understand current learning relative to past and future expectations,

as is widely accepted in culturally responsive approaches to evaluation, no such thing as a singular "program" exists for multi-site or iterative programming, (Conner, 1985). Program theories are affected by contextual and cultural influences that evolve with time and vary based on the identities and characteristics of the program's beneficiaries. As such, though it is often perceived as a singular entity, a program is not a nomothetic object; rather, it varies malleably in response to iteration and program participant. Furthermore, while it is essential to plan evaluations and plan for the use of evaluative information prior to the onset of an educational course or program, it may not be possible to use proactively planned measures with fidelity. Instead, evaluative tools must be modified in response to the reality of the program, how it actually played out in time. Evaluators should encourage consideration of the accuracy and appropriateness of evaluative tools by prompting educators to identify areas of discrepancy between intended and actual program models, making the necessary adjustments to evaluative tools to avoid misevaluation. For evaluation practice, this implies a need for responsiveness in evaluation and in the reporting of evaluative information. Finally, to ensure fairness in accountability-based decision making, standards and objectives that were excluded for the actual program theory should be identified and responded to separately. This can help ensure that low performance can be appropriately attributed, either to the student or the program, without confusing these measures as indicators of mastery if the objectives were not taught sufficiently.

Third, evaluative tools can help bring educational programs, and their evaluative information, into focus. While educators utilize curriculum mapping to identify

appropriate sequencing of learning goals for a particular program of study, evaluative tools can supplement and enrich these processes. Mapping educational programming according to the intended program theory clarifies for program stakeholders what the inputs, activities, outputs, and outcomes for the program should be. This tool can be useful for planning program resources, communicating across stakeholder groups, and identifying areas for program improvement. It can also be useful for making decisions when stakeholder groups disagree since the program theory should provide a mechanism for prioritizing information needs in alignment with primary directives of the program. Tracking the actual implementation of the program relative to the intended program provides a way to clarify areas of discrepancy for program stakeholders. This information can help with the revision of evaluative tools to ensure the accuracy and appropriateness of evaluative information and to ensure the fairness of decisions made based on this information. As a reflection point, a program theory may also create a mechanism and opportunity to identify and address effects of the program that might otherwise remain unrecognized. While program theory is widely recognized in evaluation as a tool for organizing and evaluating a program, educational programming may also benefit from generating a program theory that organizes and conceptualizes the use and influence of evaluative information. Such a theory would provide a consolidated representation, graphically and textually, of the evaluative thinking employed to make evaluative judgments within the program. This opportunity is discussed in further detail in Chapter VI.

Fourth, in educational contexts, evaluation use and influence exist simultaneously at two important levels of the educational program: the student and the program. Educational programming is predicated on learning by program beneficiaries, necessitating that evaluative information be generated, interpreted, and used at the level of the individual program beneficiary, or student. This is somewhat unique in the greater field of evaluation, where the focus tends to be placed on the program itself with some consideration of beneficiaries as a key stakeholder group. In education, evaluative data is already available: it is gathered through course-embedded tests, projects, portfolios, presentations, etc., that are designed by educators and used for assigning grades and making promotion/retention decisions. With an increasing focus on using this evaluative information at higher levels of the educational infrastructure, evaluators are uniquely positioned to help close the gap between use at the student and program levels. Leveraging evaluation skills and tools, evaluators can collaborate with educators to design grading mechanisms (e.g., assignments, rubrics) such that they simultaneously generate information relative to programmatic purposes. This intentionality in design is necessary to reduce the likelihood of misuse and misevaluation. When structuring evaluation within the typical course of teaching and learning can reduce the burden of evaluation on program stakeholders, timing is key. When a program is underway, the day-to-day experience of facilitating the program leaves little time for extensive analysis of evaluative information. This suggests that decisions rules for making sense of evaluative information need to be planned proactively (e.g., prior to the start of a specific course of program component) in order to ease the burden on sensemaking in real time.

This also creates an opportunity to leverage a more nuanced understanding of expected evaluative outcomes in order to support other program stakeholders in their sensemaking processes. It will still remain necessary to revisit evaluative tools prior to their dissemination in order to ensure the information they produce will be accurate, appropriate, and useful based on how the program played out. Evaluators' skills in data collection and management can be an asset to these efforts, both at the program and student levels. When undertaken at the student level, intentional design of responsiveness to evaluative information can facilitate the customization of learning through amelioration, remediation, and extension.

Finally, evaluative information is intended to be used robustly in educational contexts. This notion may be unique among contexts of evaluation practice, and may be somewhat taken for granted. Educators are expected to assign grades and make decisions based explicitly on evaluative information. As such, the connection between information and use is clear and direct in educational programs: the purpose of student-level evaluation is to generate evidence of learning and to make accountability-based decisions. At the program level, evidence of student learning is used to identify areas for growth and development. Used only in these ways, evaluation constitutes a missed opportunity: for learning and improvement at the student level as well as for prioritized development of the program model and its evaluative tools at the program level. The field of evaluation faces a similar tension between accountability and learning purposes for evaluation. Like assessment professionals, evaluators are also interested in how one can best leverage evaluative information for learning and improvement. Accordingly, the

educational context provides a unique opportunity for research on the purpose-use relationship since evaluative information is used readily, in a variety of ways, and in relatively quick iterations. Evaluators can leverage their experiences in assessment contexts to explore how to make evaluative information more useful, how to promote learning uses with various types of stakeholders, how to leverage evaluation for organizational learning, and how to promote evaluation use and influence. This dynamic is described in greater detail in Chapter VI.

Summary

This chapter discusses the problematics identified in Chapter IV and addresses the research questions for this inquiry. First, the three problematics are positioned within the evaluation literature on sensemaking, complexity, and attending to gaps in power and information. Though these dilemmas were identified in a case study within an assessment context, the evaluation literature suggests that these issues exist across many applied domains and practical contexts of evaluation. The research questions guiding this inquiry focused on characterizing evaluation use and influence in the assessment context. Findings from this single case study suggested an important role for process and conceptual use, in addition to perceptions of robust instrumental and symbolic use, within this specific context. By subsuming accountability purposes within learning ones, program administrators demonstrated a strong sense of ownership over the program model and its evaluative tools to address accountability and learning at the student level; at the program level, however, program administrators saw distinctions between accountability and learning purposes based on the origin of accountability structures (i.e.,

internal “for us” or external “imposed”). Analysis of evaluation influence suggested the need for additional data collection around long-term and unintended/unanticipated consequences, some mechanisms of which had already been undertaken within the case study. Finally, five considerations were offered to characterize the nature of evaluation use and influence in assessment-based educational evaluation.

CHAPTER VI

IMPLICATIONS AND FUTURE DIRECTIONS

> . . . not everything that can be counted counts, and not everything that counts can be counted. William Cameron (1963, p. 13).

This study began as an effort to understand the ways in which evaluative information from a single CSE system was used and influenced within a specific educational program. In-depth case study of this narrowly defined context could serve to provide a rich understanding of the ways in which testing information influences decisions and actions at the student and program level while simultaneously exploring the factors influencing and purposes driving use. While the previous chapter focused on discussions and tentative conclusions, in this final chapter, I will offer future research directions for evaluation use and influence in assessment-based educational evaluation after specifying the contributions and limitations of the study.

Contributions

This study makes three contributions to evaluation and evaluators working in assessment. The use of a case study approach to characterize evaluation use and influence within emerging educational evaluation contexts advances not only our understanding of these evaluation constructs, but also of how to operationalize notions of use and influence in assessment work. This study brings together two field that are currently considered as distinct and demonstrates how the knowledge and practices of each can enrich the other.

Rich, in-depth analysis of the ways in which evaluative information affects people and situations in educational environments can inform educational program design and evaluation using contemporary evaluation tools and approaches. Research on evaluation concepts as they manifest in a learning-centered, iterative context should support evaluators working in educational contexts as they undertake efforts to increase assessment use and the use of student learning evidence by integrating evaluative practices in traditional educational systems and structures.

Second, this study serves as a preliminary effort to build connections between the literatures of assessment and evaluation by making connections between concepts across theoretical and applied boundaries. While undertaking this study, I identified several publications informing the applied work of assessment practitioners. Analysis of evaluation concepts found within these volumes, as well as analysis of assessment-based educational evaluation within evaluation publications, can further this effort. Connecting these literatures can be challenging given the specificity and nuance of the terminology employed in each area. As such, a systematic review of the literature is warranted to more fully conceptualize the relationship between evaluation and assessment to elaborate ways in which the longstanding research based of the field of evaluation may be leveraged in assessment. To this end, my study offers a preliminary characterization of assessment as a niche within evaluation that fits within the contemporary tension in evaluation between accountability and learning while being distinguished by its focus on measuring competency, its iterative and evolving nature, and its use of learning as a currency.

Finally, this study extends previous research on knowledge and evaluation use by applying these concepts to educational contexts that are predicated on learning and the measurement of learner competency through standards and objectives. Analysis of evaluation use and influence in assessment-based educational evaluation facilitates the application of these notions to contemporary concerns and challenges in education, extending their impact beyond the traditional boundaries of the field.

Limitations

The nature of the study as research on a specific case serves as a significant limitation on the findings. Per Simons (2009), there are four main limitations of case study. First, the case study research itself impacts the research context as uncontrolled involvement in the program of study. Second, the case study generates a snapshot in time of the program in its report, holding still a context and processes that have since continued to develop and change through time. Third, the subjectivity of the researcher is an inevitable influence on the findings and may result in distortion of the program. Fourth, though generalization from the case study is not always as straightforward as with other research methods, it permits us to draw connections between the current case and other similar cases that can be expected to be numerous in the context of assessment. As Simons (2009, 2015) explains, five specific forms of generalization from case study are possible: cross-case, naturalistic, concept, process, and situated. The findings from this study represent a moment frozen in time that, while characterizing that moment in rich detail, represent a single, fleeting manifestation of evaluation use and influence. Findings about this singular moment may be usefully generalized (naturalistic, process) to other

assessment settings as well as conceptually to other situations that use evaluative information at multiple levels and/or require integration of accountability and learning purpose.

With regard to specific research methods utilized in the case study, though open-ended interviewing methods are useful for building rapport and collecting rich, in-depth data, close relationships between interviewers and respondents, a lack of anonymity in the research process, and maintaining relationships with the research site can pose a challenge to evidence quality (O'Leary, 2014). Conducting interviews is time consuming and requires that the interviewer responds to unexpected variation and problems (Creswell, 1998). Findings from qualitative research methods are indelibly influenced by the researcher's lens, requiring not that these influences are eradicated, but rather that they are acknowledged and considered thoroughly when making sense of findings (Maxwell, 2005; Simons, 2015). Document review involves secondary data that is reinterpreted to address the research questions and may be misunderstood, taken out of context, or "tainted by subjectivities" (O'Leary, 2014, p. 247) that are challenging for the researcher to understand. Since the CSE system produces evaluative information and distributed to information users, it was not possible to utilize observation methods (though efforts were made to observe these processes through interviews in which a reflective exercise was conducted). While this case study is focused on a collective entity (i.e., a CSE system), the data collected was gathered at the individual (i.e., student, administrator) level, limiting the applicability and usability of these findings (Yin, 2009), specifically regarding precise extrapolation to specific populations (Maxwell, 2005).

Furthermore, the case selected for this study was within a medical school, a professional training program with a narrow scope and an exceptionally structure course of study. While some of the findings from this study speak to the nature of assessment and educational evaluation broadly, others (e.g., the developmental continuum anchor, sophisticated measurement tools and performance assessments, program administrators solely responsible for assessment and measurement) might not generalize to educational programs that have fewer sophisticated tools and resources or more global and diverse learning goals. Similarly, undertaking assessment in larger programs (e.g., general education) or departments may vary in nature and practice from that of a small, specialized unit embedded within an educational organization. Broadly, these findings may be useful for assessment contexts centered on leveraging student learning evidence from course-embedded evidence gathering within programs that consist of multiple courses whose standards and expected outcomes have been mapped along a longitudinal, developmental continuum anchor (i.e., curriculum mapping).

Implications

Findings from this study suggests that student learning evidence has a profound impact on educational programming, both for individual students as well as for the program (e.g., the program model and its evaluative tools). In educational programs, student learning evidence can be used simultaneously to evaluate student performance and to drive and animate the program model. In this case study, evaluative evidence generated by a longitudinal CSE system was used to evaluate the training program as well as to evaluate individual students simultaneously (i.e., the CSE system was designed

to ensure minimal competency for student participants while concurrently identifying areas for improvement within the program model, its implementation, and/or its evaluation). The CSE system produced evaluative information that was used for course-, program-, and institutional-level goals. Interestingly, the program model included provisions for the use of student-level evaluative information, meaning that responses to specific ranges of outcomes (i.e., low, average, or high outcomes) were determined prior to the onset of a specific component of the training program (e.g., a specific workshop or course). This innovative training model is premised on a standards-based (e.g., criterion-referenced) approach to teaching and learning in which evaluative information played a key role.

Broadly, evaluative information generated by the CSE system was used by many stakeholder groups throughout and between specific courses within the training program. Instrumental, conceptual, and symbolic/legitimative uses were detected, both in evaluative processes and evaluative findings. Interestingly, process and findings uses were ample and their inter-relationships were apparent. A variety of factors seemed to influence how meaning was made of evaluative information. In evaluative tools, fair and meaningful integration of accountability and learning expectations was expected, both in terms of opportunity to learn prior to evaluation and in terms of current performance aligning with future expectations of performance without egregious gaps or "leaps in skill level." Considerations of use dimensions served to further identify areas for development for the program model based on evaluative findings. Finally, the need for alignment across systemically complex elements of the program and its evaluative tools was readily

apparent, with most areas for program improvement indicated by evaluative information related to a need for alignment.

Ultimately, designing mechanisms for eliciting information about student learning that can be useful for higher order (programmatic purposes) requires evaluation knowledge and skills. Importantly, it is unlikely that evaluative information gathered as student learning evidence will directly align with programmatic information needs unless evaluative tools are specifically designed to generate information that will be useful for these higher order purposes. To ensure appropriate and accurate use, as well as to increase use, evaluators must ensure that sources of evaluative information directly align with tiered information needs. This should serve to reduce the likelihood of misuse and misevaluation through collaboration of educators and evaluators. Evaluators can support educators in analyzing outcomes to target areas for improvement, but this does not need to occur only after the evidence has been gathered; rather, evaluators can help educators plan for the use of evaluative information within their course and program plans so that these activities can receive adequate time and attention in the educator's regular course of activities like planning, instruction, and grading. Proactive understanding of the relevance of evaluative information and pre-determined mechanisms for responsiveness with course and program design are essential for ensuring the utility and feasibility of evaluative information. In contexts of adult learning, and possibly also with younger learners, use of information is not limited to internal and external administrators but is also typically shared with learners (and the parents and guardians of younger learners). As compared with other contexts of practice for evaluators, in educational contexts, evaluative

information can be expected to be accessed and used by a wide variety of stakeholders. Accessibility of information, relative to technology and software as well as to comprehensibility and "actionability" (Donaldson et al., 2015), is particularly important in the parsing and aggregation of evaluative findings. Information needs to be packaged and reported with different stakeholder groups in mind and with specific attention paid to the utility and clarity of meaning of the evaluative information for the group to which it is being reported. Evaluators should also endeavor to reduce the likelihood of misinterpretation by considering alternative interpretations, whether accurate or inaccurate, that may be made by those accessing evaluative information.

At the program level, evaluators can help educators understand how individual students move through a training program and assist educators in developing evaluative measures that might inform movement through key transitions or identify areas for improvement at the student, course, or program levels. By identifying the information and intervals at which information might be useful, evaluators can increase the utility of evaluative information for internal program purposes, closing the distance between educator's perceptions of accountability and learning purposes. Assessment-based educational evaluation calls for the use of student learning evidence for program learning and improvement, an effort accomplished by stakeholders within this case study by aggregating student outcomes across standards and objectives to identify areas for program improvement. A similar approach is possible at the student level by aggregating outcomes across evaluative instances to consider students' longitudinal performance on various standards and objectives. If evidence is meant to inform learning and

improvement, the timing of evaluation within a course or program is paramount: learning purposes may be better served by a more bell curved approach to learning in which new material and practice is the focus early on in a course or program with later sessions focused primarily on refining previous learning based on evaluative information.

A hallmark of this case study was the integration of courses in a single training program along a developmental progression anchor onto which all knowledge, skills, and mindsets that constituted learning targets had been mapped. Ordering and alignment of skills followed a rationally devised program model in which performance in each phase was supported by performance in a previous phase and suggested the student was prepared for future learning. The program model was thus predicated on a gradual release of responsibility in which students became increasingly autonomous in their learning through teaching and modeling structured on increasingly challenging knowledge and skills. In what is known as the "Dreyfus model" (Flyvbjerg, 2001), human learning can be characterized as a progressing along five levels: novice, advanced beginner, competent performer, proficient performer, and expert. As Flyvbjerg explains, novices follow "rules" of performance legalistically before learning to accommodate situational elements based on experience as they become advanced beginners. As competent performers, goals and plans are used to shape and capture information that is or is not influenced by context until the learner is capable of intuitively anticipating problems and analyzing options prior to action as a proficient performer. Expert status is obtained when learners' behavior in response to a situation is "unhindered by analytical deliberations" (p. 21) thus "releas[ing] a picture of problem, goal, plan, decision, and action in one instant and with

no division into phases" (p. 21). Flyvbjerg explains that, while the first three phases are dominated by learner's application of logic, the final two phases are based predominantly on learners' experiences. Within each course and along the progression of the entire training program, these levels of learning seemed to be addressed within the program model and evaluative tools of the case study. Key targets for development model were understood relative to the anticipated degree of proficiency such that some targets were meant to be demonstrated at the novice level and other targets were meant to be demonstrated at the advanced beginner or competent performer levels. Interestingly, though students received a great deal of experience through practice as a result of the program, they were not expected to perform at the proficient performer or expert levels for any of the knowledge, skills, or mindsets/attitudes that they were taught. As an introductory training program, learning was focused on the application of rules, accommodations of situational elements, and the responsiveness to contextual factors in analytical decisions. As expected, alignment of the training program with future stages of student training (clerkship training, residency and fellowship programs, and career physician expectations) allowed for experience-oriented levels of learning to be addressed father along the developmental progression continuum.

Understanding teaching and learning in this way has implications for evaluators and for the use of evaluative information in assessment-based educational evaluation. Not only does it shape the way in which stakeholders ground and understand evaluation processes and findings, but it also positions evaluative information along a longitudinal continuum of learning and proficiency development. Alignment is key, both between the

program model and its evaluative tools as well as across evaluative instances, horizontally and vertically. As discussed in the previous chapter, this creates the need for evaluators to attend to the balancing of competing sources of information and to synthesize the meaning of evaluative information across evaluations. In this case, the robustness of the evaluative information, derived from many raters and multiple sources of feedback, facilitated the use of evaluative information to pinpoint levels of learning for a specific topic or course that could be improved to strengthen the program and the outcomes it generated for current and future students. In lieu of focusing on satisfactory performance at the course level, this approach places a premium on satisfactory performance at the standard or objective levels, effectively utilizing demonstrated learning as currency for progression through the training program.

Within the field of evaluation, more attention is being paid to the distinctions between what has been called "single loop" versus "double loop" learning (Patton, 2015b). As Reynolds (2014) explains, single loop learning focuses on progress toward established goals while double loop learning endeavors to consider whether goals are appropriate: "The two can be summarised by questions raised through each loop: first, Are we doing things right? (single loop) and second, Are we doing the right things? (double loop)" (p. 1382). Single loop learning aligns well with uses of evaluative information at the student level while double loop learning characterizes learning and improvement at the program level. Student learning evidence can be useful for double loop learning, but only to the extent that evaluative tools are designed with these purposes in mind. This is the challenge facing evaluation practitioners who are working

in assessment contexts. Since educational programs exist within a vast educational infrastructure, those stakeholders most responsible for implementing programs are not always permitted to choose the goals to which their programs are held accountable, forcing the focus of educational evaluation toward single loop learning by addressing such questions as: have students demonstrated proficiency relative to the goals defined by quality frameworks? and, does evaluative information provide adequate evidence of students' degree of proficiency? If, through assessment-based educational evaluation, educators are meant to use student learning evidence for program improvement, additional considerations are warranted. I believe a key element in this transition is the internalization and customization of program goals—the ownership of program goals—by those educators most directly responsible for the planning and implementation of the program. Evaluators can be key players in this process by helping educators to navigate the tensions between various types of use and influence, including how to subsume accountability purposes (imposed uses) within learning purposes (instrumental and conceptual uses).

The use of evaluation practices and tools, specifically logic models and program theories, can help stakeholders to organize, balance, and understand the various purposes and elements of their program models. Assessment practitioners have suggested that utilization focused evaluation, with its explicit focus on "intended use by intended users" (Patton, 2005), is a useful approach to assessment. Furthermore, Jankowski (2015) has called for the use of theory of change, an element of program theory, to leverage evidence in telling the story of one's educational program. Toward these ends, evaluators

are uniquely positioned to have a substantive and meaningful impact on assessment by leveraging knowledge of evaluative information use and influence in this context of practice. As Rickards and Stitt-Bergh (2016b) suggest, evaluators can do so by leveraging interpersonal skills to collaborate with educational stakeholders, and by utilizing culturally responsive approaches to evaluation. Furthermore, evaluative tools can also be useful to organize, balance, and create a shared understanding of the use of student learning evidence for decision making and action taking. By applying logic models and program theories not only to program models but also to how information is generated, interpreted, and used, evaluators can curate information and, in doing so, promote accurate and appropriate evaluation use in educational programs. In many ways, this is but one manifestation of a more global effort in evaluation, one focused on evaluative thinking. Indeed assessment-based educational evaluation creates a productive context in which research on evaluative thinking may be conducted.

A relatively new construct in evaluation (Vo, 2013), evaluative thinking addresses the application of critical, responsive thought to the design and organization of an evaluation *in situ*. Since typical approaches to evaluation follow a somewhat standard evaluation logic (establishing criteria, generating standards, measuring performance relative to those standards, evaluating findings, and making recommendations; Preskill & Russ-Eft, 2016)), evaluation logic has historically been considered as a manifestation of method selection (Vo, 2013), failing to address the broader aspects of evaluation practice that guide its process and findings. Such broader consideration is comprised in the notion of evaluative thinking. While evaluative thinking may seem a necessary, central element

of the "doing evaluation," it is not an inherent element of evaluation practice (Archibald, Sharrock, Buckley, & Cook, 2016; Buckley, Archibald, Hargraves, & Trochim, 2015). As Buckley and colleagues (2015) suggest, evaluation without evaluative thinking undercuts motivation, fosters resistance to change, overlooks important connections within the findings, and facilitates decision making based on incomplete understandings. Thus, evaluative thinking constitutes more than simply mapping out the methods and providing the reasoning for their selection and use. Rather, it is an approach through which evaluation design is customized, and those customizations are justified, relative to contextual and cultural constraints as well as other peripheral factors.

While there is not a broad literature base for research on evaluative thinking, two recent efforts have sought to define and clarify its theoretical and practical nature: a dissertation study that surveyed evaluation experts to generate a preliminary definition of evaluative thinking (Vo, 2013) and a theoretical article that defines evaluative thinking using practical findings while offering strategies for implementing evaluative thinking in evaluation capacity building (ECB; Buckley et al., 2015). Buckley and colleagues determined that most articles in which the term was used (89%) failed to define its meaning. Of the articles in which the nature of evaluative thinking was addressed, they found it was oftentimes connected to process use (Patton, 1997), that it is not constrained merely to evaluation activities but permeates all forms of disciplined systematic inquiry and reflective practice across all of an organization's processes, and that it seems especially related to evaluation capacity building. Like Buckley and colleagues, Vo positions evaluative thinking as a form of critical thinking that influences evaluation

processes and findings as well as a means of explicating the additional considerations influencing evaluation design and use. Vo asserts that how evaluators treat context affects the nature of the information produced by the evaluation, identifying other dimensions (i.e., social betterment and organizational learning) that require evaluators to record how contextual constraints and values have shaped the character of the findings (2013). As such, research on evaluative thinking in assessment-based education may provide a worthwhile means for increasing the use of assessment and student learning evidence while simultaneously advancing our understanding of evaluation use, evaluation influence, and evaluative thinking.

Future Directions

Exploring the use and influence of evaluation in practical contexts is of enduring interest to the field and its practitioners. As evaluators continue to find professional homes in assessment-based educational evaluation, positioning use and influence relative to accountability and learning purposes becomes a more pressing issue. Rather than utilizing educational evaluation as an accountability-based intervention to prove the worth of a program, evaluators can leverage an understanding of the unique contribution evaluation can make to improving learning, not only for program stakeholders (educators, educational administrators) but also for program beneficiaries, whose stakes in educational programs could not be greater. Rickards and Stitt-Bergh (2016b) describe the need for evaluators in assessment who are familiar with educational contexts, who have strong interpersonal skills, and who have capacity in culturally responsive approaches to evaluation. To this list, I add the need for an understanding of learning as a

developmental progression and the backbone on which educational programs must be built. Close collaboration with education professionals can permit evaluators working in educational contexts to make significant and important contributions to student and organizational learning.

Given the vested interest of accrediting bodies in increasing assessment use and the use of student learning evidence for improvement, further case-based research into evaluation use and influence in educational contexts governed by assessment and accreditation processes is warranted. Additional in-depth case study can permit researchers to more fully explore how evaluative information exerts influence on the people and situations across a variety of educational contexts, including other adult professional training programs (in healthcare, dentistry, law, social work, business, architecture, etc.), institutions of higher education (vocational, trade, community, undergraduate- and graduate-level training programs, etc.), and even PK-12 educational contexts. Evaluation use and influence can be expected to vary across these types of educational context, permitting in-depth analysis of the factors impacting use and influence by researching use and influence across sites and organizational types. Research in this area creates the opportunity to address more macro-level planning, tracking, evaluation, and improvement of training programs by integrating higher-order goals and information uses within the program model and its evaluative tools. Such research could possibly lead to the improvement of training programs, the inter-relationships of courses embedded within the program, and the use of evaluative information for learning and improvement, targets that will continue to be relevant given

the ever-evolving nature of educational contexts regarding student demographics, content, skills, technology, etc.

Likewise, study of the processes and findings of the interventions implemented within assessment-based educational evaluation is a fruitful area of inquiry. Evaluators' use of logic models, program theories, and other evaluative frameworks to organize and operationalize assessment can only serve to facilitate assessment efforts using student learning evidence and course-embedded assessment. Shared language and vision for program learning and improvement can serve as a launch pad for collaboration and communication across diverse stakeholder groups and their varying interests, laying plain how evaluative information informs questions and suggests future action. Given the iterative and learning-focused nature of assessment efforts, learning contexts provide an interesting opportunity for evaluators seeking to refine tools and strategies for evaluation capacity building, organizational learning systems and structures, and other such evaluation tools. Learning contexts typically exist within power structures that might permit research into the relationship between information and power with accountability and learning purposes, efforts that might be helpful to educators and evaluators alike. Continued research on evaluation influence within assessment-based educational evaluation might consider the specific processes and outcomes (Henry & Mark, 2003) underlying educational programs like Gildemyn's (2014) effort to characterize the influence of governmental monitoring and evaluation in the health sector. The utility of Jonson and colleagues' integrated model of influence in assessment is an additional direction for future research.

Considering the plethora of approaches to evaluation available, research may be warranted into the most appropriate or useful approach to assessment-based educational evaluation. Rickards and Stitt-Bergh (2016b) have suggested that culturally responsive approaches to evaluation and appreciative inquiry are particularly meaningful in educational evaluation. Professionals engaged in assessment who have become aware of utilization focused evaluation have acknowledged its utility in assessment contexts. They may also find developmental evaluation, which facilitates evaluation in emergent, dynamic realities that result from complex or uncertain environments (Patton, 2016), an instructive approach for addressing the evaluation of new or innovative educational programs. Evaluators interested in such approaches might relish the opportunity to research evaluation approaches within an iterative and ongoing process of continual development. Finally, and perhaps most interestingly for evaluators, research on evaluative thinking in assessment-based education may provide a worthwhile means for increasing the use of assessment and student learning evidence while simultaneously advancing our understanding of evaluation use, evaluation influence, and evaluative thinking.

Concluding Statement

The problematics generated through the course of this study suggest the nature of some formidable obstacles to the use of evaluative information for learning. To support stakeholders in making sense of evaluative information, evaluative processes and findings should apprise progression along a developmental continuum anchor; anticipate and address the influence of personal factors on sensemaking; support stakeholders'

ability to balance competing information sources; and be designed in alignment with multi-level learning and improvement efforts. To overcome systemic complexity, an excessive number of inter-relationships must be identified, prioritized, and reconciled across a diversity of stakeholder perspectives and relevance boundaries. To address the inherent separations within and between stakeholder groups and many varying elements of the program, a rational and responsive program model that addresses quality frameworks and information use protocols, that balances information needs of various stakeholder groups, and that leverages evaluative milestones to operationalize evaluative information is warranted. While these notions are not specific to the assessment context, findings from this in-depth case study provide a sense of how evaluative information affects people and situations within educational organizations, and how evaluative information interacts with the program model to address accountability and learning at the student and program levels. These findings further suggest that educational evaluation, through its processes and findings, exerts a substantial, meaningful, and ongoing influence on educational programming, not as a separate or distinctive entity that might be distinguished from the program, but as an integral element of program design and implementation through time.

REFERENCES

Alkin, M. C. (2011). *Evaluation essentials from A to Z*. New York, NY: Guilford Press.

Alkin, M. C. (2013). *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed.). Los Angeles, CA: Sage.

Alkin, M. C., Daillak, R., & White, P. (1979). *Using evaluations: Does evaluation make a difference?* Beverly Hills, CA: Sage.

Alkin, M. C., & King, J. A. (2016). The historical development of evaluation use. *American Journal of Evaluation, 37*(4), 568–579.

Alkin, M. C., Patton, M. Q., Weiss, C. H., National Institute of Education (U.S.), University of California, L. A., & Center for the Study of Evaluation (Eds.). (1990). *Debates on evaluation*. Newbury Park, CA: Sage.

Alkin, M. C., & Taut, S. M. (2003). Unbundling evaluation use. *Studies in Educational Evaluation, 29*(1), 1–12.

Archibald, T., Sharrock, G., Buckley, J., & Cook, N. (2016). Assumptions, conjectures, and other miracles: The application of evaluative thinking to theory of change models in community development. *Evaluation and Program Planning, 59,* 119–127.

Association of American Medical Colleges. (2008). *Recommendations for pre-clerkship clinical skills education for undergraduate medical education.* Task Force on the Clinical Skills Education of Medical Students. Washington, DC: Author.

Banta, T. W., & Palomba, C. A. (2014). *Assessment essentials: Planning, implementing, and improving assessment in higher education* (2nd ed.). San Francisco, CA: Jossey-Bass.

Barr, R. B., & Tagg, J. (1995). From teaching to learning: A new paradigm for undergraduate education. *Change: The Magazine of Higher Learning, 27*(6), 12–26. doi:10.1080/00091383.1995.10544672

Baur, V. E., Van Elteren, A. H. G., Nierse, C. J., & Abma, T. A. (2010). Dealing with distrust and power dynamics: Asymmetric relations among stakeholders in responsive evaluation. *Evaluation, 16*(3), 233–248. doi:10.1177/1356389010370251

Blaich, C., & Wise, K. (2011). From gathering to using assessment results: Lessons from the Wabash National Study (NILOA Occasional Paper No. 8). *National Institute for Learning Outcomes Assessment*.

Bogenschneider, K., & Corbett, T. J. (2010). Exploring the disconnect between research and policy. In *Evidence-based policymaking: Insights from policy-minded researchers and research-minded policymakers*. New York, NY: Routledge.

Brandon, P. R., & Singh, J. M. (2009). The strength of the methodological warrants for the findings of research on program evaluation use. *American Journal of Evaluation, 30*(2), 123–157.

Buckley, J., Archibald, T., Hargraves, M., & Trochim, W. M. (2015). Defining and teaching evaluative thinking: Insights From research on critical thinking. *American Journal of Evaluation, 36*(3), 375–388.

Cameron, W. B. (1963). *Informal sociology: A casual introduction to sociological thinking.* New York, NY: Random House.

Chen, H. T., & Garbe, P. (2011). Assessing program outcomes from the bottom-up approach: An innovative perspective to outcome evaluation. *New Directions for Evaluation, 2011*(130), 93–106. doi:10.1002/ev.368

Chouinard, J. A., & Cousins, J. B. (2015). The journey from rhetoric to reality: participatory evaluation in a development context. *Educational Assessment, Evaluation and Accountability, 27*(1), 5–39.

Chouinard, J. A., & Milley, P. (2015). From new public management to new political governance: Implications for evaluation. *The Canadian Journal of Program Evaluation, 30*(1), 1–22.

Christie, C. A. (2015). Setting the stage for understanding evaluation use and decision making. In C. A. Christie & A. T. Vo (Eds.), *Evaluation use and decision making in society: A tribute to Marvin C. Alkin* (pp. 1–15). Charlotte, NC: Information Age Pub.

Christie, C. A., & Vo, A. T. (2015). *Evaluation use and decision making in society: A tribute to Marvin C. Alkin*. Charlotte, NC: Information Age Pub.

Conner, R. F. (1985). International and domestic evaluation: Comparisons and insights. *New Directions for Evaluation, 25,* 19–28.

Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56*(3), 331–364.

Cousins, J. B., & Shulha, L. M. (2006). A comparative analysis of evaluation utilization and its cognate fields of inquiry: Current issues and trends. In I. Shaw, J. Greene, & M. Mark, *The SAGE Handbook of Evaluation* (pp. 267–291). London: SAGE Publications Ltd.

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Cunliffe, A. L. (2004). On becoming a critically reflexive practitioner. *Journal of Management Education, 28*(4), 407–426.

Dahler-Larsen, P. (2012). *The evaluation society*. Stanford, CA: Stanford Business Books, an imprint of Stanford University Press.

Dahler-Larsen, P. (2014). Constitutive effects of performance indicators: Getting beyond unintended consequences. *Public Management Review, 16*(7), 969–986.

Donaldson, S. I., Christie, C. A., & Mark, M. M. (Eds.). (2009). *What counts as credible evidence in applied research and evaluation practice?* Los Angeles, CA: Sage.

Donaldson, S. I., Christie, C. A., & Mark, M. M. (2015). *Credible and actionable evidence: The foundation for rigorous and influential evaluations.* Thousand Oaks, CA: Sage.

Dweck, C. S., Chiu, C., & Hong, Y. (1995). Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological Inquiry, 6*(4), 267–285.

Fleischer, D. N., & Christie, C. A. (2009). Evaluation use: Results from a survey of U.S. American evaluation association members. *American Journal of Evaluation, 30*(2), 158–175.

Flyvbjerg, B. (2001). *Making social science matter: Why social inquiry fails and how it can succeed again*. Cambridge, UK: Cambridge University Press.

Gargani, J. (2013). What can practitioners learn from theorists' logic models? *Evaluation and Program Planning, 38,* 81–88.

Gildemyn, M. (2014). Understanding the influence of independent civil society monitoring and evaluation at the district level: A case study of Ghana. *American Journal of Evaluation, 35*(4), 507–524.

Goodwin, S. (1996). Data rich, information poor (DRIP) syndrome: Is there a treatment? *Radiology Management, 18*(3), 45–49.

Hammersley, M. (2013). *The myth of research-based policy & practice*. Thousand Oaks, CA: Sage.

Henry, G. T., & Mark, M. M. (2003). Toward an agenda for research on evaluation. *New Directions for Evaluation, 2003*(97), 69–80.

Herbert, J. L. (2014). Researching evaluation influence: A review of the literature. *Evaluation Review, 38*(5), 388–419.

Hofstetter, C. H., & Alkin, M. C. (2003). Evaluation use revisited. In T. Kellaghan, D. L. Stufflebeam, & L. A. Wingate (Eds.), *International handbook of educational evaluation* (pp. 197–222). Dordrecht, The Netherlands: Kluwer Academic.

Hutchings, P., Kinzie, J., & Kuh, G. D. (2015). Evidence of student learning: What counts and what matters for improvement. In G. D. Kuh, S. O. Ikenberry, N. A. Jankowski, T. R. Cain, P. T. Ewell, P. Hutchings, & J. Kinzie, *Using evidence of student learning to improve higher education* (pp. 51–72). Indianapolis, IN: Jossey-Bass.

Ikenberry, S. O., & Kuh, G. D. (2015). From compliance to ownership: Why and how colleges and universities assess student learning. In G. D. Kuh, S. O. Ikenberry, N. A. Jankowski, T. R. Cain, P. T. Ewell, P. Hutchings, & J. Kinzie, *Using evidence of student learning to improve higher education* (pp. 1–23). Indianapolis, IN: Jossey-Bass.

Jankowski, N. A. (2015). *Evidence-based storytelling: Sharing our narratives*. PowerPoint Presentation presented at the Drexel University Annual Conference on Teaching & Learning Assessment, National Institute for Learning Outcomes Assessment. Retrieved from http://slideplayer.com/slide/10591972/

Jankowski, N. A., & Cain, T. R. (2015). From compliance reporting to effective communication: Assessment and transparency. In G. D. Kuh, S. O. Ikenberry, N. A. Jankowski, T. R. Cain, P. T. Ewell, P. Hutchings, & J. Kinzie, *Using evidence of student learning to improve higher education* (pp. 201–219). Indianapolis, IN: Jossey-Bass.

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation, 30*(3), 377–410.

Jonson, J. L., Guetterman, T., & Thompson Jr., R. J. (2014). An integrated model of influence: Use of assessment data in higher education. *Research & Practice in Assessment, 9,* 18–30.

Jonson, J. L., Thompson, R. J., Guetterman, T. C., & Mitchell, N. (2017). The effect of informational characteristics and faculty knowledge and beliefs on the use of assessment. *Innovative Higher Education, 42*(1), 33–47.

Julnes, G. (2012). Managing valuation. *New Directions for Evaluation, 2012*(133), 3–15.

Kinzie, J., Hutchings, P., & Jankowski, N. A. (2015). Fostering greater use of assessment results: Principles for effective practice. In G. D. Kuh, S. O. Ikenberry, N. A. Jankowski, T. R. Cain, P. T. Ewell, P. Hutchings, & J. Kinzie, *Using evidence of student learning to improve higher education* (pp. 51–72). Indianapolis, IN: Jossey-Bass.

Kirkhart, K. E. (2000). Reconceptualizing evaluation use: An integrated theory of influence. *New Directions for Evaluation, 2000*(88), 5–23. doi:10.1002/ev.1188

Koch, M. J. (2013). The multiple-use of accountability assessments: Implications for the process of validation. *Educational Measurement, Issues and Practice, 32*(4), 2–15.

Kuh, G. D., Ikenberry, S. O., Jankowski, N. A., Cain, T. R., Ewell, P. T., Hutchings, P., & Kinzie, J. (2015a). Making assessment matter. In G. D. Kuh, S. O. Ikenberry, T. R. Cain, P. T. Ewell, P. Hutchings, & J. Kinzie, *Using evidence of student learning to improve higher education* (pp. 51–72). Indianapolis, IN: Jossey-Bass.

Kuh, G. D., Ikenberry, S. O., Jankowski, N. A., Cain, T. R., Ewell, P. T., Hutchings, P., & Kinzie, J. (2015b). *Using evidence of student learning to improve higher education*. Indianapolis, IN: Jossey-Bass.

Lin Miller, R. (2015). How people judge the credibility of information: Lessons for evaluation from cognitive and information sciences. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *Credible and actionable evidence: The foundation for rigorous and influential evaluations* (2nd ed., pp. 39–62). Thousand Oaks, CA: Sage.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage Publications.

Loud, M., & Mayne, J. (2014). *Enhancing evaluation use: Insights from internal evaluation units*. Thousand Oaks, CA: Sage.

Mark, M. M. (2011). Toward better research on—and thinking about—evaluation influence, especially in multisite evaluations. *New Directions for Evaluation, 2011*(129), 107–119.

Mark, M. M., & Henry, G. T. (2004). The mechanisms and outcomes of evaluation influence. *Evaluation, 10*(1), 35–57.

Mathison, S. (Ed.). (2005). *Encyclopedia of evaluation*. Thousand Oaks, CA: Sage.

Maxwell, J. (2005). *Qualitative research design: An interactive approach* (2nd ed.). Thousand Oaks, CA: Sage.

Mayne, J. (2009). Building an evaluative culture: The key to effective evaluation and results management. *The Canadian Journal of Program Evaluation; Toronto, 24*(2), 1–30.

Mayne, J. (2014). Issues in enhancing evaluation use. In M. L. Loud & J. Mayne (Eds.), *Enhancing evaluation use: Insights from internal evaluation units* (pp. 1–14). Thousand Oaks, CA: Sage.

McDavid, J. C., Huse, I., Hawthorn, L. R. L., & McDavid, J. C. (2013). *Program evaluation and performance measurement: An introduction to practice*. Thousand Oaks, CA: Sage.

Merriam, S. B. (Ed.). (2002). *Qualitative research in practice: Examples for discussion and analysis* (1st ed.). San Francisco, CA: Jossey-Bass.

Mertens, D. M., & Wilson, A. T. (2012). *Program evaluation theory and practice: A comprehensive guide*. New York, NY: Guilford Press.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Miller, M. D., & Linn, R. L. (2013). *Measurement and assessment in teaching* (7th ed.). Boston, MA: Pearson.

Norris, N., & Kushner, S. (2007). The new public management and evaluation. In S. Kushner (Series Ed.), *Advances in program evaluation* (Vol. 10, pp. 1–16). Bingley: Emerald (MCB UP). http://www.emeraldinsight.com/10.1016/S1474-7863(07)10001-6

Nutley, S. M., Walter, I., & Davies, H. T. O. (2007). *Using evidence: How research can inform public services*. Bristol, UK: Policy Press.

O'Leary, Z. (2014). *The essential guide to doing your research project*. Los Angeles, CA: Sage.

Palomba, C. A., & Banta, T. W. (1999). *Assessment essentials: Planning, implementing, and improving assessment in higher education*. San Francisco, CA: Jossey-Bass.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text*. Thousand Oaks, CA: Sage.

Patton, M. Q. (2005). Utilization-focused evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 430–433). Thousand Oaks, CA: Sage.

Patton, M. Q. (2008). *Utilization-focused evaluation*. Thousand Oaks, CA: Sage.

Patton, M. Q. (2012). Contextual pragmatics of valuing. *New Directions for Evaluation*, *2012*(133), 97–108.

Patton, M. Q. (2015a). Misuse: The shadow side of use. In C. A. Christie & A. T. Vo, *Evaluation use and decision making in society: A tribute to Marvin C. Alkin* (pp. 131–148). Charlotte, NC: Information Age Pub.

Patton, M. Q. (2015b). *Qualitative research & evaluation methods: Integrating theory and practice* (4th ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2016). What is essential in developmental evaluation? On integrity, fidelity, adultery, abstinence, impotence, long-term commitment, integrity, and sensitivity in implementing evaluation models. *American Journal of Evaluation*, *37*(2), 250–265. doi:10.1177/1098214015626295

Peck, L. R., & Gorzalski, L. M. (2009). An evaluation use framework and empirical assessment. *Journal of MultiDisciplinary Evaluation, 6*(12), 139–156. https://eric.ed.gov/?id=EJ853621

Peshkin, A. (1988). In search of subjectivity—One's own. *Educational Researcher, 17*(7), 17–21. doi:10.3102/0013189X017007017

Porteous, N. L., & Montague, S. (2014). From discrete evaluations to enhance evaluations to a more holistic organizational approach: The case of the Public Health Agency of Canada. In M. L. Loud & J. Mayne (Eds.), *Enhancing evaluation use: Insights from internal evaluation units* (pp. 113–146). Thousand Oaks, CA: Sage.

Preskill, H. (2008). Evaluation's second act: A spotlight on learning. *American Journal of Evaluation, 29*(2), 127–138. https://eric.ed.gov/?id=EJ794317

Preskill, H., & Caracelli, V. (1997). Current and developing conceptions of use: Evaluation use TIG survey results. *Evaluation Practice, 18*(3), 209–225.

Preskill, H., & Russ-Eft, D. F. (2016). *Building evaluation capacity: Activities for teaching and training* (2nd ed.). Thousand Oaks, CA: Sage.

Prewitt, K., Schwandt, T. A., Straf, M. L., National Research Council (U.S.), & Committee on the Use of Social Science Knowledge in Public Policy. (2012). *Using science as evidence in public policy*. Washington, DC: National Academies Press.

Reynolds, M. (2014). Equity-focused developmental evaluation using critical systems thinking. *Evaluation, 20*(1), 75–95.

Reynolds, M., Gates, E., Hummelbrunner, R., Marra, M., & Williams, B. (2016). Towards systemic evaluation. *Systems Research and Behavioral Science, 33*(5), 662–673.

Rickards, W. H., & Stitt-Bergh, M. (2016a). Editors' notes. *New Directions for Evaluation, 2016*(151), 7–9.

Rickards, W. H., & Stitt-Bergh, M. (2016b). Higher education evaluation, assessment, and faculty engagement: Higher education evaluation, assessment, and faculty engagement. *New Directions for Evaluation, 2016*(151), 11–20.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Schneider, C. G., & Shoenberg, R. (1999). Habits hard to break: How persistent features of campus life frustrate curricular reform. *Change: The Magazine of Higher Learning, 31*(2), 30–35. doi:10.1080/00091389909602677

Shulha, L. M., & Cousins, J. B. (1997). Evaluation use: Theory, research, and practice since 1986. *American Journal of Evaluation, 18*(3), 195–208.

Simons, H. (2009). *Case study research in practice*. Los Angeles, CA: Sage.

Simons, H. (2015). Interpret in context: Generalizing from the single case in evaluation. *Evaluation, 21*(2), 173–188. doi:10.1177/1356389015577512

Smith, D. E. (1987). *The everyday world as problematic: A feminist sociology*. Boston, MA: Northeastern University Press.

Spillane, J. P. (2002). Local theories of teacher change: The pedagogy of district policies and programs. *Teachers College Record, 104*(3), 377–420. doi:10.1111/1467-9620.00167

Spillane, J. P., & Miele, D. B. (2007). Evidence in practice: A framing of the terrain. *Yearbook of the National Society for the Study of Education, 106*(1), 46–73.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of Educational Research, 72*(3), 387–431. doi:10.3102/00346543072003387

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.

Stitt-Bergh, M. (2016). Assessment capacity building at a research university: Assessment capacity building at a research university. *New Directions for Evaluation, 2016*(151), 69–83. doi:10.1002/ev.20196

Stitt-Bergh, M., Rickards, W. H., & Jones, T. B. (2016). Beyond the rhetoric: Evaluation practices in higher education. *New Directions for Evaluation, 2016*(151), 123–132.

Suskie, L. A. (2004). *Assessing student learning: A common sense guide*. Bolton, MA: Anker.

Vo, A. T. (2013). *Toward a definition of evaluative thinking.* Retrieved from http://escholarship.org/uc/item/26t5x5f6.pdf

Weick, K. E. (2001). Sources of order in underorganized systems: Themes in recent organizational theory. In K. E. Weick, *Making sense of the organization* (pp. 32–56). Malden, MA: Blackwell Publishers.

Weiss, C. H. (1988). If program decisions hinged only on information: A response to Patton. *Evaluation Practice*, *9*(3), 15–28.

Weiss, C. H., Murphy-Graham, E., & Birkeland, S. (2005). An alternate route to policy influence: How evaluations affect D.A.R.E. *American Journal of Evaluation*, *26*(1), 12–30.

Williams, B., & Hummelbrunner, R. (2011). *Systems concepts in action: A practitioner's toolkit*. Stanford (CA): Stanford Business Books.

Williams, B., & van't Hof, S. (2016). *Wicked solutions: A systems approach to complex problems* (2nd ed.). Lulu.com

Yin, R. K. (2009). *Case study research: Design and methods* (5th ed.). Los Angeles, CA: Sage.

APPENDIX A

INTERVIEW PROTOCOL FOR EDUCATIONAL ADMINISTRATORS

Note: information that may identify the institution has been removed from the protocol

Purpose

1. What is(are) the purpose(s) of the CSE system?
2. Equally as important, there may be ways in which scores should not be used—that is, purposes for which the scores are either not informative or not validated. Can you think of any purposes for which scores from this CSE system should not be used?
3. What claims can be made about students based on their performance on a test?
4. Are there any specific claims about students that should not be made—because the scores are either not informative or not validated?

Administrators and Educators

5. What scores result from the CSE system?
6. How are each of these scores used within your program? What decisions or actions are taken based on the scores? What other information informs these decisions and actions? Outside of your program?
7. Are there any other ways in which test scores are used? Perhaps any unintended ways that administrators, educators or students use scores?
8. Do interventions developed in response to scores differ for students based on demonstrated proficiency levels (relatively strong, weak, or average students)?
9. Do you have any other aspirations for score use that we haven't discussed already?

Students

10. Is reason for testing provided (how?) Is the use of results explained to students (when)?
11. How are students trained to interpret scores? To question results?
12. How do students use the scores? What decisions or actions are taken based on the scores? What other information informs these decisions and actions?
13. Students may sometimes re-interpret scores when discussing their outcomes. Would you say that students do this? If so, what is the understanding reflected in the message about the message?
14. How do students tend to interpret and respond to criterion-referenced scores (standards-based) versus norm-referenced scores (peer-referenced)? Is there any difference?

Administrative Role

15. How many hours per block are devoted to test selection, administration, and interpretation of scores from this specific CSE system?

16. Do scores generally seem to make sense for each student? How do you respond to scores that do not seem normal for a student? For a group of students?
17. How do you engage with other information—multiple sources of information—to support or refute suggestions made by scores?
18. What are those other sources? Do you have an example of how they work together to tell a student's story?
19. How do various score users engage with information about the standardized test? (Previous experience, bulletin, talking to students after test)
20. How often does information about the standardized test change? How are you made aware of these changes? How are students made aware of these changes?
21. Overall, how confident are you in your ability to use test results? Statistics in score reports? (Mean, standard deviation, correlations, error, etc.)
22. How does your department consider score reliability and validity?
23. How important are testing and evaluation to your work?

APPENDIX B

INTERVIEW PROTOCOL FOR STUDENTS

Note: information that may identify the institution has been removed from the protocol

1. What is(are) the purpose(s) of the CSE system? How do you know this?
2. Can you walk me through what you do when you get score reports in a sort of step-by-step explanation?
3. Why do you think you interpret scores in that way? Is it based on the way students are trained to interpret scores, on you (your personality and past experiences), or possibly something else?
4. What other information do you consider when you are thinking about your scores and next steps?
5. What would you do if you disagreed with a score report?
6. Have you ever heard students misinterpret scores or use them in ways that you think weren't intended?
7. How do your professors and other staff members use these scores?
8. How would you describe your typical performance on these assessments? Can you tell me about areas in which you tend to perform better or worse?
9. How important are these evaluations to your development?
10. Overall, do you feel your scores seem to make sense for each experience? Do you think they represent your performance accurately?
11. Do you have any aspirations for evaluation of your performance that we haven't discussed already? Is there anything you wish the CSE system could do?

APPENDIX C

GUIDELINES FOR STUDENT INTERVIEW AS A REFLECTIVE EXERCISE

When students' grades are released (per the normal schedule), students will be scheduled into 10-minute time slots to engage in a think-aloud. I will request that they share with me the story that they make of the data by voicing, step-by-step, what they look at and what it means to them.

While this may seem like an unstructured interview, the purpose of this method is to have the students (who are adults) voice an internal process that they naturally go through, one that is impossible to observe without prompting the think-aloud. There will be no role for the interviewer as a member of this experience; rather, the interviewer will frame the request and then merely observe the student's process.

Following this self-directed think-aloud, the interviewer will ask follow-up questions to clarify understanding (member checking) and to clarify what takeaways the student has from the score report.

The interview will be audio-recorded and the same precautions will be used that follow for the interviewing protocol and that are reflected in the recruitment information sheet. A copy of the recruitment information sheet will be provided to each participant. As with the interviewing protocol, participants will be told that they are not obligated to participate and can end the experience at any time, that their voices will be recorded so absolute confidentiality cannot be ensured, and that the audio files will be selectively transcribed using pseudonyms so that no responses can be individually identified.