USING CURRICULUM-BASED MEASURES TO ASSESS SPECIAL EDUCATION TEACHER CANDIDATES' APPLICATION OF DEFINING AND MEASURING BEHAVIOR SKILLS

A Dissertation in
Special Education
by
Katie E. Hildebrand

© 2008 Katie E. Hildebrand

Submitted in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

August 2008
The dissertation of Katie E. Hildebrand was reviewed and approved* by the following:

David L. Lee
Associate Professor of Special Education
Dissertation Advisor
Chair of Committee

Frank R. Rusch
Professor of Special Education

Daniel Weiss
Assistant Professor of Psychology

Kathy L. Ruhl
Professor of Special Education
Head of the Department of Educational and School Psychology
    and Special Education

*Signatures are on file in the Graduate School
ABSTRACT

Determining teacher preparation program effectiveness is critical as students are served best by teachers who possess highly developed skill sets. Yet, research suggesting viable methods to evaluate preparation programs is sparse. This limited research base suggests that effectiveness is determined by both external reviewers (i.e., accrediting agencies) and internal evaluators (i.e., faculty within a preparation program). Both external and internal reviewers rely on much the same data, commonly using methods of surveys, interviews, examinations, portfolios, and to a lesser extent, observations individually or in combination to determine program effectiveness. As there are benefits to using each method, there are a multitude of limitations. For instance, all measures are summative in nature indicating that they provide an overall picture of the program and its teacher candidates at a given point in time, often at the program’s end. Additionally, measures such as examinations focus primarily on content knowledge and do not assess application of pedagogical skills. Given these limitations, teacher educators may find it difficult to identify gaps in the program or teacher candidates who are having difficulty developing necessary skills.

To address these limitations, an assessment system based on principles of curriculum-based measurement was created to assess teacher candidates’ application skills. In curriculum-based measurement, alternate forms of probes (i.e., quizzes) are administered repeatedly to monitor the students’ progress over time. In this study, 20 undergraduate special education teacher candidates enrolled in a large eastern university were evaluated on one strand of the teacher preparation curriculum, defining and measuring behavior, using a series of 4 application probes plus a post-test measure.
Results indicated that probes were internally consistent, and alternate forms of the probes were equivalent. In terms of validity, probes were not able to predict overall course grades, but were able to predict performance on the post-test. Additionally, when using the post-test data as the outcome measure, probes were able to discriminate among high, medium, and low performers. Implications, limitations, and future research directions are discussed.
# TABLE OF CONTENTS

LIST OF TABLES ............................................................................................................................ vii

LIST OF FIGURES ........................................................................................................................... viii

ACKNOWLEDGEMENTS ................................................................................................................ ix

CHAPTER 1: Introduction ............................................................................................................... 1

CHAPTER 2: Review of Related Literature ...................................................................................... 8

Evaluating Teacher Preparation Program Effectiveness ............................................................... 9

Internal Reviews .......................................................................................................................... 10

External Reviews .......................................................................................................................... 11

The National Council for Accreditation of Teacher Education ..... 11

Teacher Education Accreditation Council .......................................................... 13

Research on the Methods Used to Collect Data Program Efficacy Data.... 14

Consumer Opinion ....................................................................................................................... 14

Examinations................................................................................................................................. 18

Portfolios........................................................................................................................................ 20

Observations.................................................................................................................................... 22

Summary of Evaluation Procedures and Recommendations .................. 22

Using CBM to Address Limitations ......................................................................................... 23

CHAPTER 3: Methodology ........................................................................................................... 27

Participants and Setting ............................................................................................................... 27

Development of Predictor Variables ......................................................................................... 28

Validation of Predictor Variables ............................................................................................... 31

Procedures..................................................................................................................................... 33
Inter-rater Reliability ................................................................. 35
Procedural Integrity ................................................................. 35
Criterion Measures ................................................................. 35
CHAPTER 4: Results ..................................................................... 37
Data Analyses .............................................................................. 37
Were application probes internally consistent? ......................... 37
Were alternate-forms of application probes equivalent? .......... 38
Are application probes predictors of criterion measures? .......... 38
Can probes discriminate among performance levels? ............ 41
How early can application probes predict performance? ........ 43
CHAPTER 5: Discussion .............................................................. 50
Technical Adequacy .................................................................... 51
  Reliability .................................................................................. 51
  Validity ....................................................................................... 52
Implications .................................................................................. 53
Limitations ................................................................................... 55
Future Research Directions ....................................................... 56
Summary ....................................................................................... 57
REFERENCES ............................................................................... 59
APPENDIX A. Recruitment Script ............................................. 68
APPENDIX B. Informed Consent ............................................... 69
APPENDIX C. Application Probes ............................................. 71
APPENDIX D. Procedural Integrity Checklist ............................ 99
LIST OF TABLES

TABLE 1. Synthesis of Standards for Beginning Teachers from NCATE and TEAC ................................................................. 12

TABLE 2. Skill Areas Prior to Validation Study ......................................................... 31

TABLE 3. Skill Areas After the Validation Study ....................................................... 34

TABLE 4. Internal Consistency .................................................................................... 39

TABLE 5. Results from Linear Regression Analyses .................................................. 40

TABLE 6. Means and Standard Deviations of Probes ............................................... 42

TABLE 7. Differences Among Groups Across Probes ............................................... 49
LIST OF FIGURES

FIGURE 1. Matrix ........................................................................................................30
FIGURE 2. Averages Across Probes...........................................................................43
FIGURE 3. Data Disaggregated by Overall Course Average......................................45
FIGURE 4. Data Disaggregated in Classroom Management ......................................46
FIGURE 5. Data Disaggregated in Observing in Exceptional Settings......................47
FIGURE 6. Data Disaggregated by Post-test...............................................................48
ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, mentor, and now friend, Dr. David Lee whose love of research, writing, and teaching has inspired me more than he will ever know. Without Dr. Lee’s encouragement, kindness, and humor, this journey would not have been possible. I am forever grateful to him for his guidance and support throughout the completion of this dissertation and the program. Dr. Kathy Ruhl, whose wisdom, grace, and continuous support have tremendously impacted my life, deserves very special thanks. I only hope to affect as many teachers and students as she. I would like to thank Dr. Frank Rusch for teaching me the value of quality research and cheering me on throughout the program. Many thanks go to my outside member, Dr. Dan Weiss, who provided much insight during the dissertation process. Lastly, I must thank my very first mentor, Dr. Ellen Long, who encouraged me to fulfill this dream. All of these individuals are models I hope to one day emulate.

I am forever appreciative for two of my fellow graduate students. Youjia has the Wittiest and most innocent sense of humor, and Mandy has the kindest heart. Their endearing qualities helped me to persevere through the good times and the bad; their friendships are invaluable, and I know that our special bond will be everlasting.

Finally, I must thank my loving parents, Kevin and Kathlyn, who I cherish so dearly. Their unwavering support has been monumental. I would also like to thank my sister, Kari, who I admire greatly, and Sean and Liam who keep me smiling. Lastly, many thanks to my future husband, Sean Hoffman, who is incredibly patient, compassionate, and encouraging. His unconditional love has helped me through this journey.
CHAPTER 1

Introduction

It is the responsibility of special education teachers to foster growth and improve performance for students with special learning needs (Greenwood & Maheady, 1997). Working with students with special needs is a difficult job, as special educators must accommodate all learners, provide appropriate instruction, monitor student progress, and manage behaviors within the classroom. To that end, educators with high quality training are better prepared to handle the everyday classrooms tasks than educators with little or poor quality training, such as those with emergency certification (Nougarat, Scruggs, & Mastropieri, 2005). In an effort to promote positive academic and social outcomes for students, teacher educators must prepare teacher candidates to handle the challenges they will face on a daily basis. Teacher educators provide the key link between training and k-12 student achievement. Thus, examinations of teacher preparation in terms of both content and pedagogy seem critical.

In selecting content, teacher preparation programs often rely on national/organizational standards. For instance, in a recent survey examining special education teacher preparation programs across the nation, Conderman, Katsiyannis, and Franks (2001) determined many programs align their content with standards developed by The Council for Exceptional Children (CEC), a special education professional organization whose stated purpose is to improve educational outcomes for students with special needs (2003). Much of this content is drawn from research on teaching and learning across the fields of special education, general education, psychology, and educational psychology. However, presenting content alone is not sufficient to ensure
that teacher candidates consistently learn these important skills. Carefully constructed assessment tools that yield quantifiable outcomes must be implemented to guarantee the proficiency of programs and their graduates (Burkhart, 1996; NCATE, 2006).

Galluzzo and Craig (1990) suggest several reasons for examining the effectiveness of teacher preparation programs. First, systematic evaluations conducted by external agencies or the program itself can help identify strengths and needs to inform program modifications. For instance, if a cohort (i.e., group of teacher candidates) lacks critical foundational knowledge or vital skills, teacher educators can reexamine the program and provide additional opportunities to practice or re-teach skills as needed.

Second, conducting evaluations helps demonstrate program accountability to key stakeholders in the community. Given the results of evaluations, educational stakeholders may better understand teacher preparation, the experiences of teacher candidates, and the language used in the field. The better educational stakeholders understand the function of teacher preparation programs, the more likely they are to make informed decisions regarding funding and changes to programs (e.g., changes in state certification requirements). Finally, the results from and methodologies used in evaluations can add to the knowledge base on teacher preparation. As the findings of individual evaluations may not generalize from institution to institution, methods used to determine effectiveness are highly similar and may be applicable across a range of programs (Cochran-Smith, 2003).

Teacher preparation programs evaluate their effectiveness using a variety of methods. These assessment techniques are employed by both those inside (i.e., internal evaluators) and outside the program (i.e., external evaluators) (Conderman et al., 2001). Program faculty/administrators generally conduct internal reviews on a yearly basis in
order to determine program strengths and needs and make modifications as necessary (Conderman et al.). The methodologies utilized by internal evaluators typically take the form of surveys (e.g., Conderman et al.), licensure exams (e.g., Wineburg, 2006), portfolios (e.g., Conderman, 2001), and observations of teacher candidates in the field (e.g., Darling-Hammond, 2006). In addition to internal reviews, independent agencies such as the National Council for Accreditation of Teacher Education (NCATE) and Teacher Education Accreditation Council (TEAC) carry out external reviews of programs. External evaluators often examine the same data collected by internal evaluators in providing their unbiased reviews of programs. Conducted in 5- to 7-year cycles, teacher preparation programs often use results of external reviews to align programs with national standards (NCATE, 2008). Since both types of reviews use very similar, if not the same data, it is paramount that data used in any review process give an accurate picture of a given program.

Development of assessment methods that yield valid results with high levels of reliability would seem a critical first step in the evaluation of teacher preparation programs. However, while many methods used in both internal and external reviews have potential strengths, these methods are not without limitations. For example, surveys and interviews provide insight from graduates of programs and their current employers, but they generate descriptive data based primarily on opinion and perception, which may or may not be accurate (Darling-Hammond, 2006; Isaac & Michael, 1995). In addition, the psychometric characteristics of these instruments are relatively unknown. Many programs also require candidates to take licensure exams, which focus primarily on background and foundational knowledge (Selwyn, 2007). While efficient to administer, there are two
shortcomings of licensure examinations. First, the summative nature of examinations, which are often taken at the very end of a training program, does not allow for curricular modifications until teacher candidates have completed their training. Therefore, gaps in critical content and skills may go unnoticed until teacher candidates (or a cohort) have graduated. Second, the criterion scores on such exams are often set very low, thus not requiring teacher candidates to demonstrate mastery of program content, but rather demonstrate only a shallow level of understanding of content prior to certification. Portfolios contain examples of teacher candidates’ work over time, but are difficult to score in a reliable manner (Conderman, 2001). Similarly, observing teacher candidates while they work with children provides a valid means to assess generalization of learning from college classroom to k-12 settings, but is time consuming and often occurs intermittently in teacher preparation programs. Given the infrequent, summative nature of most of these assessments, teacher educators may fail to identify areas of difficulty at both the program and individual level. If deficits go undetected, teacher candidates may exit their program with gaps in their knowledge base and be unable to apply critical teaching skills (e.g., instructional strategies, behavior management, assessment) in school settings, which may impact their future students who need effective teachers.

In order to more effectively identify potential gaps during preparation, a formative measure that addresses some of the shortcomings of other methods (e.g., surveys, portfolios, examinations) is needed. Incorporating formative measures that examine foundational knowledge and pedagogical skills within a teacher preparation program directly aligns with suggestions presented by the Federal Government (e.g., The U.S. Department of Education, 2005) and accrediting agencies (e.g., NCATE, 2008; TEAC,
Given the need to evaluate overall program effectiveness and to monitor teacher candidates’ performance as they progress through a teacher preparation curriculum (NCATE, 2006), perhaps the research base on curriculum-based measurement (CBM) may provide insight into a new application of a valuable method of evaluation that addresses shortcomings of current methods.

There is a large body of research supporting CBM to monitor academic performance (Deno, 1987; 2003). Curriculum-based measurement relies on a series of standardized probes, typically administered once or twice a week, to monitor students’ movement through a curriculum (Stecker, Fuchs, & Fuchs, 2005). The high level of technical adequacy of CBM is one of its defining features, as it is a reliable and valid assessment across a variety of academic areas (Deno, 2003; Hosp, Hosp, & Howell, 2007). Commonly used in k-12 settings to document progress toward instructional goals, CBM technology may help teacher educators examine program effectiveness and monitor the progress of teacher candidates collectively, as well as individually.

Early research on expanding CBM to settings beyond k-12 is promising. For example, Hildebrand and Lee (2007) found that CBM was a reliable and valid predictor of outcomes in an introduction to special education course. These researchers suggested that teacher educators can use CBM data in several ways. First, when examining data collectively, teacher educators can determine if candidates have mastered critical material presented in class. Second, when disaggregating participant data by course outcomes, teacher educators can identify and modify instruction for low performers to account for the needs of all learners. Teacher educators could possibly use this information to ensure that all teacher candidates, collectively and individually, leave a given preparation
program well-trained and ready to meet the many demands of working with children with special needs. However, one potential problem in this prior work was the focus on recall of facts as opposed to application of knowledge. Teacher preparation programs need a system that ensures mastery of critical content knowledge and also allows teacher candidates to apply that knowledge to address classroom-based situations.

Building on prior work, the purpose of the current study was to further examine the use of CBM as an assessment in a teacher preparation program. This measure may help verify that programs are graduating teacher candidates who can apply the necessary content knowledge and pedagogical skills required in k-12 settings. To address gaps in prior work, this current study aimed to assess teacher candidates’ application of skills related to one strand of a teacher preparation program – defining and measuring behavior. Research questions parallel previous studies which examined the technical adequacy of CBM in other areas (e.g., Espin, Busch, & Shinn, 2005; Espin, Busch, Shinn, & Kruschwitz, 2001; Hildebrand & Lee, 2007). Specifically, this study sought to address the following five questions:

(1) Are probes examining the application of one content strand of a special education teacher preparation program, defining and measuring behavior, internally consistent?

(2) Are alternate-forms of probes examining application of defining and measuring behavior implemented in a teacher preparation program equivalent?

(3) Are probes examining application of defining and measuring behavior implemented in a teacher preparation program accurate predictors of criterion measures (i.e., performance)?
(4) Can probes examining application of defining and measuring behavior discriminate among performance levels?

(5) How early can probes examining application of defining and measuring behavior predict performance in a teacher preparation program?
CHAPTER 2

Review of Related Literature

Currently, there is much concern surrounding teacher preparation programs at both the national and state level. Members of society sometimes question the competence and quality of America’s teachers (Cochran-Smith, 2003), while holding teacher preparation programs accountable (Boyd, Goldhaber, Lankford, & Wyckoff, 2007; Galluzzo & Craig, 1990). Researchers (e.g., Darling-Hammond, Holzman, Gaitlin, & Heilig, 2005) and educational stakeholders (e.g., U.S. Department of Education, 2005) suggest that teacher preparation programs play a significant role in training teachers. As there are differences (e.g., philosophies, courses, practica experience) among the 1300 teacher preparation programs across the country, the primary function of a traditional program remains constant (Cochran-Smith & Zeichner, 2005). Teacher educators devote their time and energy to produce competent teacher candidates who are well-equipped to face the many challenges of the classroom (Korthagen, 2004; Thomas & Loadman, 2001).

Well-prepared teachers must demonstrate mastery of critical content, pedagogical skills, and dispositions (Da Ros-Voseles & Moss, 2007; Samuelowicz & Bain, 1992), which will help future students succeed academically and socially (Darling-Hammond, 1996). Therefore, teacher educators have a difficult job as they must dedicate sufficient time to teach critical content (i.e., background and foundational knowledge within a subject area) and develop pedagogical skills (i.e., art and science of teaching) of teacher candidates (Schulman, 1986; 1987). As both content and pedagogy are important, it is the blending of these two that creates effective teachers (Segall, 2004). Thus, to ensure that teacher candidates positively impact future students, teacher educators must build a
program with a strong conceptual framework grounded in theory and research. In addition, the program must be continuously evaluated and modified based on data (Cochran-Smith, 2003).

Increasing external demands, such as a cap on the number of credits allowed in many majors (i.e., degrees must be completed in four years) place preparation programs in a difficult position. Although there is a clear need for both training in content and pedagogy (Schulman, 1986; 1987), legislation such as No Child Left Behind (NCLB) effectively limits the time teacher educators can devote to pedagogical skills and dispositions of teacher candidates in favor of more content (Porter-Magee, 2004). As such, teacher preparation programs must be efficient in their use of time when developing program components where these vital skills are taught. One way teacher preparation programs can maximize their efficiency is to continually evaluate program content and outcomes. It is through this iterative process that content and pedagogy are refined, thus ensuring that candidates exit programs strong in key skill areas (Galluzzo & Craig, 1990). Unfortunately, there is little empirical work examining the evaluation of teacher preparation programs (Cochran-Smith, 2003). In this chapter I first identify and discuss the benefits and limitations of current evaluation methods and then propose a new method with the potential to address identified weaknesses.

_Evaluating Teacher Preparation Program Effectiveness_

Program evaluation systems generally fall into one of two categories – internal and external. Internal evaluators (i.e., faculty or employees within the program) use a variety of methods (e.g., surveys, licensure examinations, portfolios) to evaluate their programs (Darling-Hammond, 2006; Wineburg, 2006). In external reviews, a third party
conducts a non-biased evaluation of programs using much of the same types of data collected and compiled by internal evaluators during internal reviews. External reviewers are generally well-respected professionals who have the knowledge and ability to make comparisons across programs while potentially making recommendations based on collected data (Gall, Gall, & Borg, 2003). Both internal and external reviews can guide changes within the program ultimately impacting teacher candidates; however, the process by which internal and external evaluators review programs differs.

**Internal Reviews**

Internal evaluators (e.g., teacher educators, instructors, administrators) typically conduct internal reviews on a yearly basis (Conderman et al., 2001). In order to examine program effectiveness, internal evaluators collect quantitative and qualitative data, which can be used to inform both programmatic decisions (e.g., change requirements for program entry, include additional practica experiences, alter the sequence of courses) and curricular changes (e.g., change textbooks, modify courses, provide greater breadth or depth of content) (Conderman et al.). Additionally, internal program evaluations can validate teaching methods and assessments while verifying the current status (i.e., knowledge of content and development of skills) of their teacher candidates.

Internal reviews encompass a variety of methods that can help ensure program quality. Internal evaluators often use surveys (e.g., Conderman et al., 2001), licensure exams (e.g., Wineburg, 2006), portfolios (e.g., Conderman, 2001), and observations (e.g., Darling-Hammond, 2006) to gain information about program content and outcomes. Most programs use a combination of the methods noted above to determine effectiveness. These data are often summarized and faculty/administrators make necessary changes.
**External Reviews**

External reviews rely on much of the same data (e.g., examinations, observations, and to a lesser extent portfolios) collected during internal reviews. However, preparation programs and other stakeholders (e.g., legislators, administration, advocates) give more credence to external reviews because they offer an independent evaluation of a given program and in some cases colleges and institutions. Currently, there are two nationwide organizations, both recognized by the U.S. Department of Education and the Council for Higher Education Accreditation as accrediting agencies that are responsible for evaluating the overall quality of teacher preparation programs in the country (Murray, 2005). For some time, The National Council for Accreditation of Teacher Education (NCATE) was the exclusive accrediting agency examining teacher preparation programs. However, more recently in 1997, a second accrediting agency, The Teacher Education Accreditation Council (TEAC), emerged as NCATE’s competitor (Basinger, 1998). Both NCATE and TEAC dedicate themselves to ensuring that teacher preparation programs produce the most competent, qualified professionals in the field of education (Murray; Wise, 2005) (see Table 1 for a synthesis skills teacher candidates should possess by the end of their program).

*The National Council for Accreditation of Teacher Education (NCATE).* More than 50 years ago several teacher associations convened to create NCATE. In 1954, NCATE replaced The American Association of Colleges for Teacher Education (AACTE), which had previously been responsible for accrediting teacher preparation programs (Wise, 2005). Currently, NCATE partners with institutions in 48 states (plus
Table 1

*Synthesis of Standards for Beginning Teachers from NCATE and TEAC*

---

**Beginning teachers should:**

- have adequate content knowledge, pedagogical skills, and dispositions.
- employ appropriate instructional strategies.
- create an environment conducive for learning.
- assess students.
- communicate and collaborate with others.
- develop professionally.
- work with diverse individuals across settings.

---

Washington D. C. and Puerto Rico) and at present, accredits 632 colleges of education (NCATE, 2008).

The National Council for Accreditation of Teacher Education’s review is a systematic, multi-step process. After NCATE creates a board of reviewers, a two-step accreditation review process begins. First, reviewers evaluate the unit (i.e., school, college, or department) against NCATE standards. Next, the second half of the review focuses upon particular programs within an institution (college). Each program at the institution seeking accreditation submits the results from a series of 6 – 8 assessments that evaluated the skills outlined in Table 1. These quantitative data demonstrate how a given program evaluated mastery of specific standards as put forth by professional
associations in a given area of emphasis within education. For instance, when NCATE assesses a mathematics education program, NCATE examines the assessments put forth by the program to determine if they align with The National Council for the Teaching of Mathematics’ (NCTM) standards. For NCATE to accredit a special education program, NCATE examines documentation against standards promulgated by The Council for Exceptional Children (CEC) (NCATE, 2008).

Teacher Education Accreditation Council (TEAC). The Teacher Education Accreditation Council (TEAC) is a newer organization, founded in 1997 by The Council of Independent Colleges consisting of presidents from smaller, private colleges (Basinger, 1998). In 2003, the U.S. Department of Education recognized TEAC as a second educational accrediting organization (Murray, 2005). Founders of TEAC explained that this organization provides an alternative to NCATE as critics not only question NCATE’s standards but assert that their accreditation process is costly and inefficient (Basinger). Currently, TEAC accredits 49 programs from 43 institutions from 30 states (TEAC, 2008).

As their mission, very similar to NCATE, is to prepare quality educators, the manner in which TEAC accredits institutions is quite different from NCATE. Rather than comparing an institution and program to a series of professional standards, TEAC, through an academic audit, examines claims made by each program and determines if the program provides sufficient evidence to support these claims. For example, if a program states that their teacher candidates are knowledgeable in a particular subject area, then the program must support this statement with qualitative or quantitative evidence to verify
the claim (e.g., number of credit hours, licensure examinations, pre-test and post-test data).

**Research on the Methods Used to Collect Data Program Efficacy Data**

There is a great deal of overlap among methods used by both internal and external entities to evaluate program effectiveness. These methods generally fall into four categories: consumer opinion (e.g., Darling-Hammond, 2006), examinations (e.g., Conderman et al., 2001; Majsterek, Prigge, & Fennerty, 1994), portfolios (e.g., Conderman, 2001), and observations (e.g., Wineburg, 2006). Each method is described and advantages and limitations discussed below.

**Consumer Opinion.** Surveys and interviews are often used in research to depict the status of a current sample of individuals (Charles, 1995). When examining the research on program effectiveness, surveys and interviews surfaced as the most popular vehicle to collect data (e.g., Conderman et al., 2001; Darling-Hammond, 2006). Programs sought responses primarily from teacher education graduates, employers, and supervisors. Although surveys and interviews relied on self-reported data, they provided insight into programs based on graduates’, employers’, and supervisors’ perspectives. This type of information allows teacher educators to make modifications based on thoughts and opinions of others connected to the program.

Several studies examined the use of surveys and interviews in teacher preparation programs. Surveys and interviews were often conducted with teacher education graduates from focal institutions (Andrew & Schwab, 1993; Simpson & Sandidge, 1994; Whitney, Golez, Nagel, & Nieto, 2002). However, Whitney et al. also questioned teachers from additional institutions to determine how their graduates compared to other teachers in the
field. Most of those surveyed were novice teachers, only in the field for a maximum of three years. The number of graduates surveyed varied. Researchers reported surveying between 194 graduates (Simpson & Sandidge) to as many 1,380 graduates (Andrew & Schwab). After sending reminders to graduates, the return rate of the surveys ranged from 48% (Andrew & Schwab) to 70% (Simpson & Sandidge).

Survey and interview content differed across studies; however, similarities emerged. Background information on the teacher education graduates was frequently reported (Andrew & Schwab, 1993; Darling-Hammond, 2006; McEneaney & Sheridan, 1993; Thomas & Loadman, 2001). Graduates were asked to respond to questions about demographics, employment status, and current position. Teacher education graduates also reported their level of preparedness at the conclusion of the preparation program (Andrew & Schwab; Conderman, 2001; Conderman et al., 2001; Darling-Hammond; McGuire, 2001; Pellow & Kuhns 1992; Thomas & Loadman; Whitney et al., 2002). Questions primarily pertained to identifying what they felt their strengths and needs were when they started their position as a novice teacher. More specific questions were also included such as the adequacy of graduates’ knowledge of subject matter (McEneaney & Sheridan; Simpson & Sandidge, 1994; Thomas & Loadman). Surveyors also assessed the candidates’ use of instructional approaches (Andrew & Schwab; McEneaney & Sheridan; McGuire, Simpson & Sandidge; Thomas & Loadman), assessment techniques (Andrew & Schwab; Schumacher & Cauley, 1990; Thomas & Loadman), and behavior management skills (Simpson & Sandidge; Thomas & Loadman). McGuire as well as Thomas and Loadman inquired about the candidates’ abilities to work with diverse populations. Lastly, communication and collaboration with parents, colleagues, and
administrators also emerged as popular questions on surveys (McGuire; Simpson & Sandidge; Thomas & Loadman).

In order to further assess the perceived impact of graduates, researchers also sent inquiries to teachers’ current employers, namely principals (Andrew & Schwab, 1993; Darling-Hammond, 2006) and superintendents (Darling-Hammond). West and Freeman (1988) also sent surveys to school supervisors while Simpson and Sandidge (1994) questioned their teacher education graduates’ internship committees (i.e., principals, cooperating teachers, and teacher educators during the student teaching semester). The number of employers responding among articles ranged from 62 (West & Freeman) to 481 employers (Andrew & Schwab). The return rate of the surveys from employers ranged from about 50% (West & Freeman) to 70% (Andrew & Schwab).

Employers were asked to rate graduates on their level of perceived preparedness (Andrew & Schwab, 1993; Darling-Hammond, 2006) and to indicate how their graduates compared to teachers from different institutions (Andrew & Schwab). Darling-Hammond also inquired if employers would likely hire graduates of their program in the future. More specifically, employers rated graduates on their knowledge of subject matter (West & Freeman, 1988; Simpson & Sandidge, 1994) as well as their ability to deliver instruction, maintain interpersonal relationships, and demonstrate professionalism (Andrew & Schwab; Simpson & Sandidge).

There are certainly benefits of using surveys and interviews to evaluate programs. They are a systematic way to gather descriptive information from large groups of individuals in order to answer proposed questions regarding specific preparation programs (Gall et al., 2003; Isaac & Michael, 1995). They also, used in this manner,
provide feedback from a variety of respondents who potentially offer differing viewpoints. In addition, when and if the right questions are posed, surveys, more so than interviews, can be a time-efficient and a cost-effective way to gather important data on a particular population (Charles, 1995). Teacher educators can use survey data collected from a variety of sources in the form of feedback in order to make programmatic and curricular changes (Conderman et al., 2001). For example, upon analyzing survey data from various individuals, Simpson and Sandidge (1994) noted that their candidates needed additional knowledge and expertise in classroom management and organization. To address these issues, they created a supplementary course that focused on these topics. Whitney et al. (2002) collected survey data, and after evaluating the data, they formed focus groups. The focus groups provided more detail from those who were dissatisfied with pieces of the program.

As consumer opinions are undoubtedly useful, surveys and interviews are not without limitations. First surveys and interviews require respondents to self-report data, which can be problematic (Andrew & Schwab, 1993; Darling-Hammond, 2006). Data collected via surveys and interviews may yield some factual information, but they rely primarily on opinion, which may be accurate or inaccurate. In addition, information that is collected depends on the questions posed and how they are posed, potentially limiting the scope of information collected (Charles, 1995). Furthermore, respondents of surveys may not report information that is entirely accurate as they may tell the interviewer what they think the interviewer wants to hear. These inaccuracies may be unintentional, as novice teachers may not realize their own limitations.
Given the potential problems with surveys and interviews, it is surprising that technical data was often missing in studies. Andrew and Schwab (1993) were the only researchers to state that their survey was reliable and valid; however, specific evidence to support this claim was not provided. Without any reliability and validity data, it is not possible to determine the technical adequacy of the instruments used in studies. Reliability and validity of an instrument are critical to examine and report as they both provide evidence to suggest that the instrument measures the construct it was designed to measure and measures that construct consistently (Gall et al., 2003). In sum, surveys emerged as a useful summative tool to obtain descriptive data. However, much of the information gathered using this format was based on the opinion and perception of respondents.

Examinations. Much like surveys, collecting examination data as indicated in the literature, was a common method to determine program effectiveness. Unlike surveys, exams contain questions related to a specific topic and yield a score to help to determine how well individuals mastered key material (Salvia, Ysseldyke, & Bolt, 2007). Assessing content knowledge (i.e., background and foundational knowledge) to ensure mastery of material is not a new concept for teacher preparation programs (Ramsey & Algozinne, 1991). In 1990, Schumacher and Cauley described using the National Teachers Exam (NTE), now replaced by the Praxis Exam, which is a requirement across many programs to measure content knowledge (Conderman et al., 2001; Majsterek et al., 1994; Wineburg, 2006). In short, the Praxis I and II Exams are a series of educational tests created by the Educational Testing Services (ETS) and used to assess background knowledge in basic skills and foundational knowledge related to the field of study. Most institutions use a
licensing exam to ensure that teacher candidates demonstrate mastery of background and foundational knowledge. This exam is often linked to state certification (Conderman et al.; Majsterek et al.; Pellow & Kuhns, 1992; Wineburg).

Some programs go beyond a national exam and create their own summative assessments to evaluate their program. Rather than using a licensure exam, Conderman (2001) reported that the faculty at the University of Wisconsin at Eau Clair developed their own departmental test to assess content knowledge. The faculty at this institution developed a 55 question multiple-choice test and administered it to freshmen or undergraduates entering the program as a pre-test. The post-test was then given immediately prior to student teaching practica. This departmental test was given anonymously because the purpose of the test is to determine the effectiveness of the program. Reliability and validity details were not reported.

Exams are an efficient way to measure mastery of foundational and background knowledge, and they provide a source of objective, performance data. In addition, the reliability and validity of licensure exams (e.g., Praxis) has been well-established (ETS, 2008). Furthermore, because exams are summative in nature, they can be helpful when reporting the outcomes for teacher candidates and the program overall. For instance, if a program is undergoing an external review, the program can submit compiled exam data to an accrediting agency such as NCATE or TEAC. This allows for comparisons to be made across programs nationally or within the same state.

As exams generally provide an overall picture of how teacher candidates perform (Selwyn, 2007), their summative nature can also be viewed as a limitation. Because these types of assessments are given once and typically at the end of a program, there is no
time to intervene if teacher candidates are missing key content. Also passing the exam does not guarantee mastery of content as criterion scores are often set very low in some states (U.S. Department of Education, 2005), which results in little separation among candidates. This lack of separation (i.e., everyone passes) makes it very difficult to identify the strengths and weaknesses of a given preparation program, as well as teacher candidates who might be struggling. A second limitation of exams is that they are not curriculum-based. Exams (e.g., Praxis) are created to cover a broad scope and not intended to address unique curricular strands. This may make it difficult to assess programs with unique areas of emphasis (e.g., urban education, rural education). Thus, given the wide-range of content contained on exams, gaps in specific content knowledge or deficits in skills may go undetected.

*Portfolios.* The number of programs requiring their candidates to create portfolios is increasing (Conderman et al., 2001; Majsterek et al., 1994; Wineburg, 2006). In a statewide conducted survey in Pennsylvania, Pellow and Kuhns (1992) found that 38% of elementary education programs and 43% of secondary education programs utilized portfolios as an assessment tool. A portfolio is a summative measure and defined as a collection of work samples (Salvia et al., 2007). Portfolios can be used in a teacher preparation to document programmatic goals and objectives while simultaneously showcasing candidates’ work (Conderman, 2001; Hopfer, 1999). Elements contained in portfolios vary across institutions. Some include lesson and unit plans (Hopfer), while others require samples of student work and written reflections responding to time in the classroom (Darling-Hammond, 2006; Hus & Bergeron, 1997).
Just as there are benefits to conducting surveys, interviews, and licensure exams, there are also purported advantages of using portfolios as a measure of effectiveness. Overall, researchers agreed that portfolios are an authentic measure of performance as they provide opportunities for candidates to present evidence of their abilities to apply skills to real-life situations (Conderman et al., 2001). In addition, portfolios encourage reflective thinking as once teacher candidates receive feedback, they can revise their work while examining strengths and weaknesses of their product (Salvia et al., 2007). Lastly, a portfolio can be used as a “picture book” to showcase work at job interviews (Hus & Bergeron, 1997).

These reported advantages aside, there is a lack of empirical research and efficacy data supporting portfolios as a method of assessment (Russell & Butcher, 1999; Salvia et al., 2007). To that end, there are numerous limitations of using portfolios to assess individual teachers or programs. First, scoring portfolios can be a subjective endeavor making appropriate levels of inter-rater reliability difficult to achieve (Sparapani, Abel, Easton, Edwards, & Herbster, 1997; Tillema & Smith, 2007). For instance, two evaluators may assess the same portfolio, yet produce two very different scores. Whereas one assessor may believe that the candidate has met the requirements of a given program, because of the subjectivity in the scoring process, the second assessor may disagree. Although rubrics are often developed in order to aid in scoring portfolios (Conderman et al., 2001), these rubrics remain insufficient, as reliability is still unknown. Until experts establish better methods and evaluators have additional training, scoring portfolios remains a subjective task (Salvia et al.).
Observations. Observations during practica or student teaching provide direct evidence related to teacher candidates’ effect on students. There are a variety of ways to observe teacher candidates in the classroom (e.g., direct observations, videotaping, teaching mini lessons to peers) (Wineburg, 2006). Through the use of these different types of observations, teacher educators and cooperating teachers can determine if teacher candidates apply information they learned through their coursework in classroom settings. Furthermore, observations are both authentic and performance-based assessments in that the skills assessed are often required on a daily basis when teacher candidates have their own classroom.

As with other methods, observations also have limitations. Teacher educators typically conduct formal observations intermittently during practica and also throughout the student teaching semester (i.e., final semester in the program). Similar to examinations, observations may occur infrequently, and observers may be inconsistent on the content of feedback. The nature of observations also makes delivering feedback a difficult process. Feedback is often delayed and non-specific. Because the scoring of observations is often subjective, two observers, viewing the same lesson, may focus upon different skills or dispositions consequently resulting in a disparity between scores (Andrews & Barnes, 1990).

Summary of Evaluation Procedures and Recommendations

Through this review of the teacher preparation program evaluation literature, several important trends emerged. First, external reviews conducted by accreditation agencies, NCATE or TEAC, in combination with internal reviews carried out by internal evaluators both surfaced as popular approaches to improve teacher preparation programs.
A variety of methods (e.g., surveys, interviews, licensure exams, portfolios) are used in both reviews to demonstrate accountability and also to identify strengths and needs in order to make the necessary modifications.

However, as there are advantages to the teacher preparation program assessments when used collectively, there are also limitations. First, all measures including surveys, interviews, licensure exams, portfolios, and observations are summative measures that provide only an overall picture of a program at a given point in time, generally after a cohort of teacher candidates has graduated. As changes could be made to affect future cohorts, there is no time to intervene with the current cohort because they will have moved on and likely into a teaching position. Second, the technical adequacy of surveys, interviews, portfolios, and observations has not been established. Thus, given the lack of reliability and validity data supporting these measures and their summative nature, teacher preparation programs are in need of more technically sound evaluation methods.

Using CBM to Address Limitations

Given the problems with current assessments noted above, what is needed is a psychometrically sound evaluation system that (1) provides objective data linked to important content and skills, (2) yields data that can be used to make instructional/programmatic changes over time, and (3) can track individual teacher candidates throughout a program. One type of formative evaluation or progress monitoring system that has been used with a variety of students is curriculum-based measurement (CBM). Curriculum-based measurement is designed to systematically assess the effects of instruction and movement through a curriculum over time (Deno, 1987). Research suggests that CBM is a reliable and valid indicator of performance for k-
12 students across several content areas (Deno, 2003). Researchers have established the technical adequacy of CBM in reading (e.g., Madelaine & Wheldall, 1998), mathematics (e.g., Thurber, Shinn, & Smolkowski, 2002), and writing (e.g., Espin, Shin, Deno, Skare, Robinson, & Benner, 2000), and most recently in content areas (e.g., Espin et al., 2005; Espin et al., 2001; Espin & Foegen, 1996). This type of assessment, when used with teacher education candidates, allows for systematic, frequent evaluation to monitor progress and make program changes as necessary.

The advantages of CBM address the limitations of current program evaluation measures in several important ways. First, current measures (surveys, interviews, exams, portfolios, observations) are all static summative measures, which yield an overall picture of the preparation program and its teacher candidates typically at the end of a year or at the completion of a program. However, CBM is a dynamic, formative measure; multiple forms of probes are created and administered repeatedly over time, which potentially allows teacher educators to make modifications, based on data. Second, as many of the measures mentioned above lack reliability and/or validity (e.g., surveys, interviews, portfolios) the technical adequacy of CBM has been widely established (Deno, 2003; Hosp et al., 2007). Curriculum-based measurement has over 25 years of empirical evidence supporting its use in k-12 settings. More recently, the utility of CBM has been evaluated for use in post-secondary settings (e.g., Allinder, 1995; Bean & Lane, 1990). Larson and Ward (2006) recently conducted a study with 69 university students to examine the relationship between weekly CBM probes and students’ overall grades in a psychology course. These researchers suggested that CBM could predict positive course outcomes. However, key pieces of information (e.g., graphs, reliability information) were
missing. In an effort to fill this gap in the literature, Hildebrand and Lee (2007) evaluated the technical adequacy of CBM probes conducted in a university setting. These researchers found CBM to be both a reliable and valid method to monitor the acquisition of vocabulary terminology in an introductory special education course. Moreover, CBM discriminated among high, medium, and low performing groups. Teacher educators could use this information to ensure that all candidates exiting a given preparation program are well-trained and ready to meet the many demands of working with children with special needs.

As in prior research (e.g., Espin et al., 2001; Larson & Ward, 2006), Hildebrand and Lee (2007) used a vocabulary matching task to evaluate student performance. However, most matching tasks are limited to recall of factual information. While important, this measure may fall short of allowing teacher candidates to convert this knowledge into application. An assessment of application of skills over time is a key component to determining the effectiveness of a program.

Given the empirical evidence supporting the use of CBM, and the need for a formative assessment that will inform programmatic decisions and changes, the purpose of the study was to further determine the technical adequacy of CBM in a collegiate setting. As a first step, probes that required candidates to apply their knowledge within one strand of a special education program, defining and measuring behavior, were created and technical adequacy examined. The following five research questions were used to guide this study:
(1) Are probes examining the application of one content strand of a special education teacher preparation program, defining and measuring behavior, internally consistent?

(2) Are alternate-forms of probes examining application of defining and measuring behavior implemented in a teacher preparation program reliable?

(3) Are probes examining application of defining and measuring behavior implemented in a teacher preparation program accurate predictors of criterion measures (i.e., performance)?

(4) Can probes examining application of defining and measuring behavior discriminate among performance levels?

(5) How early can probes examining application of defining and measuring behavior predict performance in a teacher preparation program?
CHAPTER 3

Methodology

The main goal of this study was to better establish the technical adequacy (i.e., reliability and validity) of probes used to monitor the progress of teacher candidates through a special education teacher preparation program. As an initial step in this process, I selected a narrow band of information common to many teacher preparation programs, defining and measuring behavior. Researchers (Alvarez, 2007; Hertzog, 2002; Johnson, Rice, Edgington, & Williams, 2005) have argued that managing both the classroom and student behaviors is an important strand within a teacher preparation program, and it is often noted as a difficult skill for novice teachers. Defining and measuring behavior is considered a key aspect of the behavior management process (Van Houten & Hall, 2001). Targeting a strand that is vital within all teacher preparation programs helps ensure that information gained through this study will generalize across other teacher preparation programs.

Participants and Setting

Recruitment took place in a course common to all first-year, special education majors (i.e., Classroom Management) at a large eastern university. All participants were in the first semester of their junior year and at the beginning of a four-semester (i.e., two year) NCATE-accredited teacher preparation program in special education. Potential participants (n = 23) were informed that the purpose of the study was to examine ways of assessing teacher knowledge and skills (see Appendix A for recruitment script). A total of 20 undergraduate teacher candidates (17 female and 3 male) provided consent (see Appendix B for informed consent). Two individuals elected not to participate in the study.
One participant was excluded because he was enrolled as a part-time student in the program. Upon completion of the study, participants received $25.

Development of Predictor Variables

Predictor variables were a series of five probes designed to assess student application of knowledge of defining and measuring behavior. A multi-step process was used to determine the specific content of the probes. First, the standards promulgated by The Council for Exceptional Children (CEC entitled What Every Special Educator Must Know: Ethics, Standards, and Guidelines for Special Educators, 2003) were consulted. More specifically, Standard 5, referred to as the Learning Environment and Social Interactions, included information related to this topic. Next, an analysis of the curriculum of the program was conducted. Syllabi and course descriptions were examined and content related to defining and measuring behavior documented. Finally, popular behavior management textbooks were reviewed (e.g., Alberto & Troutman, 2006; Cooper, Heron, & Heward, 1987; Salvia et al., 2007; Sprick & Howard, 1995; Van Houten & Hall, 2001). In all, six skill areas related to defining and measuring behavior were identified and included: (1) dimensions of behavior (e.g., duration, frequency, latency), (2) objectively defining behavior, (3) observational recording systems (e.g., checklists, duration recording, event recording), (4) writing goals and objectives, (5) interpreting data using a graph, and (6) calculating reliability (i.e., inter-observer agreement).

A series of application questions assessing each of the identified skill areas was developed (e.g., 20 questions to assess objectively defining a behavior). A multiple-choice format was chosen for probes. Multiple-choice questions allow participants to
apply learned information as opposed to solely recalling facts (Salvia et al., 2007). In order to make the items application based, participants were asked to respond to classroom-based scenarios. Each scenario consisted of a maximum of four sentences. Eight possible choices were generated for the response set. Eight was chosen as the number of potential choices to make the questions more challenging (Salvia et al.) and also to reduce the chance of participants answering through the process of elimination.

One response served as the keyed response (i.e., the correct answer) while six responses functioned as distractors. Because participants were not necessarily exposed to any or all of the content at a given time, as the purpose of the probes was to monitor progress over time, the eighth response was “I don’t know.” As per Salvia et al., responses were kept short and approximately the same length. The first seven responses were placed alphabetically under the stem, which helped to vary the position of the keyed response.

Once the pool of probe items was established for each skill area, individual probes were created. Number of questions for each skill area was determined by how often each skill area was covered throughout the curriculum (see Figure 1) and the approximate length of coverage in popular behavior management texts (e.g., Alberto & Troutman, 2006; Cooper et al., 1987; Salvia et al., 2007; Sprick & Howard, 1995; Van Houten & Hall, 2001). For example, because dimensions of behavior is a topic found in many of the courses and deemed important to later successful use behavior management strategies (Van Houten & Hall), it was weighted more heavily (i.e., more questions appeared) than other topics that emerged less frequently (e.g., recording behavior). See Table 2 for a listing of the number of questions for each skill area.
The pilot probes consisted of twenty questions (Espin et al., 2001). This number was chosen as CBM probes must achieve a balance between efficiency (brevity) and technical adequacy (contain enough items to represent the domain being assessed). Questions from each skill area pool were selected randomly and then placed on the probe in an arbitrary order on a paper (21.6 cm x 27.9 cm). Five probes were created in this manner. Once probes were created, to ensure content and face validity, three content experts with experience teaching the identified skill areas were contacted and asked to examine the probes. Probes were modified based on the recommendations from the experts (e.g., reword questions, provide a variety of distractors, use different terminology).
Table 2

*Skill Areas Prior to Validation Study*

<table>
<thead>
<tr>
<th>Skill Areas</th>
<th>Number of Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – dimensions of behavior</td>
<td>8</td>
</tr>
<tr>
<td>2 – defining behavior</td>
<td>4</td>
</tr>
<tr>
<td>3 – observational recording systems</td>
<td>2</td>
</tr>
<tr>
<td>4 – writing goals and objectives</td>
<td>2</td>
</tr>
<tr>
<td>5 – interpreting data using graphs</td>
<td>2</td>
</tr>
<tr>
<td>6 – calculating reliability</td>
<td>2</td>
</tr>
</tbody>
</table>

*Validation of Predictor Variables*

A two-step process was used to validate probes as predictors. First, to determine the appropriate administration time for the probes, three students (two undergraduate students and one graduate student) previously exposed to this content completed three probes each. These students were chosen because they demonstrated mastery of the content (i.e., earned an A in the Classroom Management Course). They independently took the probes in an office setting, and the following directions were read: “*When I say begin, read each scenario carefully. Circle the best answer. If you are having much difficulty with a question and feel as though you do not have a best guess, you may circle the option I don’t know. Are there any questions? You may begin.*” A stop watch was activated, and duration to complete each probe was monitored. Students completed
probes in an average of 13 minutes with a minimum of 90% accuracy. Thus, 15 minutes was determined as an appropriate time limit for probes.

Next, a pilot study was conducted to document the equivalence of each of the five probes. Participants were recruited from a large undergraduate class comprised of general education majors. This course, Teaching Exceptional Students in General Education Settings, was required of all education majors (with the exception of special education majors). The pilot study was conducted at the beginning of a class period. First, students were given a packet containing two consent forms (the first was signed and returned to the investigator and the second retained by the participant) and one probe. Second, probe versions (1-5) were randomly distributed within the class. After the students had the opportunity to read and sign the consent forms, the investigator read the following directions aloud: “When I say begin, read each scenario carefully. Circle the best answer. If you are having much difficulty with a question and feel as though you do not have a best guess, you may circle the option I don’t know. Are there any questions? You may begin.” Students had 15 minutes to complete the probe, and all students completed the probe within the allotted time frame. A total of 126 students participated in the pilot study. A chi square test was used to document equivalence of the number of students across probes (Probe 1 \( n = 25 \), Probe 2 \( n = 23 \), Probe 3 \( n = 26 \), Probe 4 \( n = 27 \), Probe 5 \( n = 25 \)). There were no differences in the number of students across probes, \( \chi^2(1, N = 126) = .349, p = .986 \).

Analyses of variance (ANOVA) was used to analyze probe equivalence data. The assumptions of ANOVA (i.e., independence of observations, homogeneity of variance, and normality) were evaluated prior to the analysis (Glass & Hopkins, 1996). Each of the
assumptions was met, indicating ANOVA as an appropriate form of analysis for the pilot data.

The initial analysis yielded differences, indicating that probes were not equivalent, 

\[ F(1, 121) = 4.58, p = .002 \] (Probe 1 \( M = 11.64, SD = 2.36 \), Probe 2 \( M = 9.26, SD = 2.09 \),

Probe 3 \( M = 10.50, SD = 2.69 \), Probe 4 \( M = 9.44, SD = 1.60 \), and Probe 5 \( M = 10.2, SD = 1.90 \)). Follow-up analyses using Tukey post hoc tests revealed differences between Probes 1 and 4 and also Probes 1 and 2. Probes were then subjected to an item-level analysis in order to determine which specific questions differed across probes. A series of 20 ANOVAs (Lunney, 1970), which were adjusted using a Bonferroni correction procedure \( (p < .0025) \), were conducted to determine possible differences across questions. Five questions that yielded differences were removed from probes.

The analysis was conducted a second time after the five questions had been removed. Means across probes were similar (Probe 1 \( M = 8.42, SD = 1.83 \), Probe 2 \( M = 7.30, SD = 1.56 \), Probe 3 \( M = 8.35, SD = 1.74 \), Probe 4 \( M = 7.60, SD = 1.50 \), and Probe 5 \( M = 7.68, SD = 1.39 \)). No significant differences emerged, indicating that the probes were equivalent, 

\[ F(1, 121) = 1.78, p = .137 \]. The final version of the probe contained 15 questions (see Appendix C for all probes) from five skill areas (see Table 3).

_procedures_

Probes were administered to special education teacher candidates by the investigator four times during the fall semester (Probe 1 in late August, Probe 2 in early October, Probe 3 the beginning of November, and Probe 4 at the start of December). The final probe was given once during the spring semester (early January). Two probes
Table 3

*Skill Areas After the Validation Study*

<table>
<thead>
<tr>
<th>Skill Areas</th>
<th>Number of Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – dimensions of behavior</td>
<td>8</td>
</tr>
<tr>
<td>2 – defining behavior</td>
<td>3</td>
</tr>
<tr>
<td>3 – observational recording systems</td>
<td>0</td>
</tr>
<tr>
<td>4 – writing goals and objectives</td>
<td>2</td>
</tr>
<tr>
<td>5 – interpreting data using graphs</td>
<td>1</td>
</tr>
<tr>
<td>6 – calculating reliability</td>
<td>1</td>
</tr>
</tbody>
</table>

(Probes 1 and 3) were administered in a Classroom Management Course while two (Probes 2 and 4) were given in a course entitled Observing in Exceptional Settings. The CBM post-test was administered in the spring in an Assessment Course. Administration of probes occurred approximately every 4 to 5 weeks. Instructors were contacted in advance to set up dates and times for probe administration.

During each session, standardized directions were read to the participants: “*When I say begin, read each scenario carefully. Circle the best answer. If you are having much difficulty with a question and feel as though you do not have a best guess, you may circle the option I don’t know. You will have 15 minutes to complete the probe. Complete as many questions as you can. When you finish the probe, please bring it to me. Are there any questions? You may begin.*” A stop watch was activated, the session times for 15 minutes, and participants were told to stop working.
If a participant missed a probe, the participant was contacted via email and asked if s/he could make-up the probe. If s/he did not make up the probe, his/her data were recorded as missing data and were not accounted for on that particular probe. Three participants in total were asked to makeup the probe; however, none did so.

**Inter-rater Reliability**

The investigator of the study scored all probes. A second rater, a graduate student with advanced standing in the Special Education doctoral program, independently scored 20% of the probes. Point-by-point agreement was calculated by dividing the number of agreements by number of agreements plus disagreements. Then, this number was multiplied by 100 to obtain a percentage. Reliability was 100%.

**Procedural Integrity**

Procedural integrity data were collected to ensure that each probe was administered in a standardized manner. A graduate student collected procedural integrity data using a checklist outlining administration procedures as well as the administration script (see Appendix D). Procedural integrity data were collected during one administration (i.e., 20%) and was 100%.

**Criterion Measures**

Grades from the two courses that contained the content of the CBMs and a post-test CBM were used as criterion measures. Participants’ course grades (i.e., numerical grade) were used from each of the targeted courses (Observing in Exceptional Settings and Classroom Management) in several ways. First, course grades from the fall semester were averaged to determine if probe scores predict the overall average of course grades. Second, participants’ individual course grades were used as a criterion measure to
determine if probes could accurately predict the outcome of each targeted course. Finally, individual course exams and activities related to the content area of the CBMs were analyzed individually.

A post-test CBM was also used to tease out the effects of content outside of the defining and measuring behavior strand. The purpose of this criterion measure was to assess the accuracy of probes in predicting outcomes of the same content.
CHAPTER 4

Results

Data Analyses

Data were analyzed using SPSS software. Results are grouped according to each research question. When appropriate reliability coefficients were compared to the criterion $r = .60$. Salvia et al. (2007) suggest that a reliability coefficient of $r > .60$ is sufficient for group administered tests. Additionally, results were evaluated at the conventional level of statistical significance ($p < .05$).

Assumptions for linear regression and an analysis of variance (ANOVA) were considered prior to data analysis. First, in this study, a series of linear regression analyses were used to assess the adequacy of probes as predictors of criterion measures. Therefore, before conducting statistical tests, the four assumptions of regression (i.e., linearity, independence, homoscedasticity, and normality) were analyzed and met, indicating that it would be appropriate to evaluate the data using this method. Second, to examine potential differences among groups, ANOVAs were used. All assumptions of ANOVA (i.e., independence of observations, homogeneity of variance, and normality) were evaluated and met. Thus, it was also acceptable to use ANOVAs to analyze these data (Glass & Hopkins, 1996).

Were application probes internally consistent?

An analysis of internal consistency determines the extent to which the content is similar and measures the same construct within each probe (Popham, 2006; Salvia et al., 2007). Cronbach’s alpha was used to analyze the internal consistency of the five probes (Huck, 2000; Salvia et al.). The median coefficient alpha among the five probes was $r$
= .644. Coefficients ranged from $r = .588$ to .696. Four of the five coefficients (80%) were above $r = .625$ (see Table 4).

*Were alternate-forms of application probes equivalent?*

Alternate-forms reliability determines if individual probes are parallel and examines the degree to which the same construct was measured across probes (Huck, 2000; Popham, 2006; Salvia et al., 2007). The results from the pilot study, discussed in the methodology section, indicated that alternate forms of both the probes were equivalent (see Chapter 3).

*Are application probes predictors of criterion measures?*

Criterion-related validity compares scores on one assessment to scores on a related, more established criterion measure to determine the extent to which the new instrument provides accurate prediction of performance (Huck, 2000). To establish the criterion-related validity of the probes, the investigator conducted a series of linear regression analyses. Probes served as predictors and the average of course grades, course grades individually, the post-test CBM (i.e., final probe), project grades, and exam scores functioned as criterion measures.

First, the mean score across four probes served as the predictor variable (Hildebrand & Lee, 2007), and the numerical average of the two targeted courses together (i.e., Classroom Management and Observing in Exceptional Settings) functioned as the criterion measure. The average across the four probes did not surface as a significant predictor, $t = 1.36, p = .19$ (see Table 5 for additional statistics from linear regression analyses). Next, each individual probe was used as the predictor while the course average again functioned as the criterion measure. Probe 3 was the only probe that
Table 4

*Internal Consistency*

<table>
<thead>
<tr>
<th>Probe number</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.696</td>
</tr>
<tr>
<td>2</td>
<td>.588</td>
</tr>
<tr>
<td>3</td>
<td>.655</td>
</tr>
<tr>
<td>4</td>
<td>.644</td>
</tr>
<tr>
<td>5</td>
<td>.625</td>
</tr>
</tbody>
</table>

emerged as a significant predictor of participants’ average across courses. It accounted for approximately 30% of the variance, $t = 2.64, p = .02$.

Next, the mean of the four probes was assessed as a potential predictor, while the course grades individually functioned as the criterion measure. In both targeted courses, the mean across four probes was not a significant predictor of course average (i.e., in the Classroom Management Course $t = 1.63, p = .12$ and in Observing in Exceptional Settings $t = .46, p = .65$). However, again, a series of linear regressions was conducted to determine if individual probes served as a predictor for the overall average in each course. When the average from the Classroom Management course was the criterion, Probe 3 emerged as a significant predictor and accounted for approximately 29% of the variance, $t = 2.57, p = .02$. When using each probe as an individual predictor and the average of the Observing an Exceptional Settings Course was used as the criterion, no probe
Table 5

*Results from Linear Regression Analyses*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Criterion</th>
<th>$R^2$</th>
<th>Coefficient</th>
<th>SE</th>
<th>t-Value</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Across Probes (1-4)</td>
<td>Course Average</td>
<td>.09</td>
<td>1.20</td>
<td>.88</td>
<td>1.36</td>
<td>.19</td>
</tr>
<tr>
<td>Probe 3</td>
<td>Course Average</td>
<td>.30</td>
<td>1.37</td>
<td>.52</td>
<td>2.64</td>
<td>.02*</td>
</tr>
<tr>
<td>Average Across Probes (1-4)</td>
<td>Average: Classroom Management</td>
<td>.13</td>
<td>2.05</td>
<td>1.26</td>
<td>1.63</td>
<td>.12</td>
</tr>
<tr>
<td>Average Across Probes (1-4)</td>
<td>Average: Observation Course</td>
<td>.01</td>
<td>.35</td>
<td>.77</td>
<td>.46</td>
<td>.65</td>
</tr>
<tr>
<td>Probe 3</td>
<td>Average: Classroom Management</td>
<td>.30</td>
<td>2.32</td>
<td>.89</td>
<td>2.61</td>
<td>.02*</td>
</tr>
<tr>
<td>Average of Probes (1-4)</td>
<td>CBM post-test</td>
<td>49</td>
<td>.80</td>
<td>.20</td>
<td>3.90</td>
<td>.001*</td>
</tr>
<tr>
<td>Probe 1</td>
<td>CBM post-test</td>
<td>20</td>
<td>.27</td>
<td>.14</td>
<td>1.95</td>
<td>.07</td>
</tr>
<tr>
<td>Probe 2</td>
<td>CBM post-test</td>
<td>46</td>
<td>.68</td>
<td>.19</td>
<td>3.57</td>
<td>.003*</td>
</tr>
<tr>
<td>Probe 3</td>
<td>CBM post-test</td>
<td>47</td>
<td>.66</td>
<td>.18</td>
<td>3.64</td>
<td>.002*</td>
</tr>
<tr>
<td>Probe 4</td>
<td>CBM post-test</td>
<td>19</td>
<td>.43</td>
<td>.22</td>
<td>1.82</td>
<td>.09</td>
</tr>
</tbody>
</table>

*Note. $R^2$ = coefficient of determination, SE = standards error, and * denotes significance at the $p < .05$ level.*
emerged as a predictor of course performance. Probes did not predict exam scores or project grades.

In order to control for the effects of other content on the criterion measure, the final probe was used as the post-test CBM. When using the average of Probes 1-4 as the predictor, the probe average was a significant predictor and accounted for 49% of the variance, $t = 3.90, p = .001$. Additionally, when using each probe individually as the predictor and the post-test CBM as the criterion, Probes 2 and 3 emerged as significant predictors. Probe 2 accounted for 46% of the variance, $t = 3.57, p = .003$. Probe 3 accounted for 47% of the variance, $t = 3.64, p = .002$. Neither Probe 1 ($t = 1.95, p = .07$) nor Probe 4 ($t = 1.82; p = .09$) surfaced as predictors of the post-test CBM.

*Can probes discriminate among performance levels?*

This question addresses the utility of using CBM to identify participants in need of assistance by examining weekly probe scores in relation to participants’ overall average in courses. To answer this question several analyses were conducted. First, the investigator calculated means and standard deviations of each probe across all participants (see Table 6). The means for each probe were then graphed (see Figure 2). As with CBM in k-12 settings, an aim line was drawn (Hosp et al., 2007) and set at 93% (i.e., 14 correct out of 15) to signify mastery. It is a common practice in CBM for instructors to compare the data path to the aim line. If several data points are below the aim, an intervention is needed (Hosp et al.). When examining the data path representing the mean of participants, the trend had an ascending trend and ran parallel to the aim line indicating positive gains over time.
Table 6

Means and Standard Deviations of Probes

<table>
<thead>
<tr>
<th>Probe Number</th>
<th>N</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>7.45</td>
<td>2.48</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>12.47</td>
<td>1.50</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>13.33</td>
<td>1.61</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>13.00</td>
<td>1.67</td>
</tr>
<tr>
<td>Post-test</td>
<td>17</td>
<td>12.33</td>
<td>1.53</td>
</tr>
</tbody>
</table>

Note. N = number of participants, M = mean, and SD = standard deviation.

Data were then disaggregated and individual participants were grouped according to the overall course grades (i.e., Classroom Management and Observing in Exceptional Settings) and by the post-test CBM scores to determine if the probes can discriminate among performance levels. Participants obtaining an average of an A were categorized as the high group; participants earning a B were classified as the middle group; and participants with a mean lower than a B formed the low performing group. However, in some cases, due to the small number of participants, only two groups could be formed – the high group (those earning an A average) and the low group (those earning lower than an A average). After computing the averages across probes, a series of ANOVAs were conducted in order to compare mean probe scores (1-4) across performance groups. No differences emerged among the performance levels when examining the overall average of course grades (p = .112), the grade in the Classroom Management Course (p = .06), or
the grade in the Observing in Exceptional Settings Course ($p = .79$), indicating that CBM could not discriminate among performance levels based on course grades.

Similarly, data were grouped according to the CBM post-test score. Participants earning between a 90-100% formed high group, those obtaining an 80-89% were the medium group, and those with less than a 79% were classified as the low group. Data were similarly disaggregated and participants grouped according to their overall CBM post-test score to determine if probes could accurately discriminate among groups. An analysis of variance was used to examine possible differences among the groups.

*How early can application probes predict performance?*

This question further examined the utility of CBM by determining how early application probes could predict performance. That is, can participants who are having
difficulty be identified early in a given semester? Using the disaggregated data according to participants’ course grades and post-test score, (i.e., performance levels) data were graphed (see Figures 3, 4, 5, and 6) and analyzed. Only when the CBM was used as the post-test were there differences between performance groups early in the semester. For instance, there were differences on the very first probe, $F(2, 16) = 7.05, p = .006$ (see Table 7). More specifically, an instructor would be able to use these data early in the semester to address potential student issues.
Figure 3

*Data Disaggregated by Overall Course Average*
Figure 4

Data Disaggregated in Classroom Management

![Graph showing data disaggregated in classroom management.](image-url)
Figure 5

Data Disaggregated in Observing in Exceptional Settings

[Graph showing data points and trends labeled as 'high' and 'low' over 'maintenance' with 'Probes Number' on the x-axis and 'Number Correct' on the y-axis.]
Figure 6

Data Disaggregated by Post-test

Number Correct

1 2 3 4 5
Probe Number

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

medium

low

high

maintenance
Table 7

*Differences Among Groups Across Probes*

<table>
<thead>
<tr>
<th>Probe</th>
<th>df</th>
<th>F</th>
<th>p</th>
<th>Tukey Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2, 16</td>
<td>7.05</td>
<td>.006*</td>
<td>(M &gt; L)</td>
</tr>
<tr>
<td>2</td>
<td>2, 15</td>
<td>6.95</td>
<td>.007*</td>
<td>(H &gt; L)</td>
</tr>
<tr>
<td>3</td>
<td>2, 15</td>
<td>5.70</td>
<td>.014*</td>
<td>(H &gt; L)</td>
</tr>
<tr>
<td>4</td>
<td>2, 15</td>
<td>1.49</td>
<td>.257</td>
<td></td>
</tr>
<tr>
<td>Post-test</td>
<td>2, 15</td>
<td>92.50</td>
<td>.001*</td>
<td>(H &gt; L); (M &gt; L); (H &gt; M)</td>
</tr>
</tbody>
</table>

*Note. df = degrees of freedom, F = F ratio, p = significance value, and * indicates significance at the p < .05.*
CHAPTER 5

Discussion

Given the need for highly effective teachers, teacher preparation becomes a key variable in producing positive student outcomes. Before programs can chart a course for the future, it is important to be able to look back with clear lenses. Thus, a more systematic and objective way to evaluate critical content knowledge and skills covered in teacher preparation programs is needed (Wise, Darling-Hammond, McLaughlin, & Bernstein, 1986).

Curriculum-based measurement may provide a method to verify teacher candidates’ acquisition of important knowledge and skills prior to exiting teacher preparation programs. Furthermore, CBM provides for multiple “check-points” allowing teacher educators to evaluate the progress of individual teacher candidates, as well as evaluate the overall quality of a teacher preparation program as a whole. As a first step in establishing CBM as a method to evaluate teacher preparation programs, the primary purpose of this study was to determine if a series of CBM application probes, focusing on one strand of a curriculum (i.e., defining and measuring behavior), were reliable and valid predictors of student performance. Results suggested that application probes are reliable and valid predictors of performance on a post-test designed to assess application of key behavior management skills. In the remainder of this chapter the technical characteristics of CBM relative to the other methods of evaluation are presented. In addition, a research agenda for CBM as a method to evaluate teacher preparation programs that parallels that of k-12 is proposed.
Technical Adequacy

Curriculum-based measurement enjoys a high level of research support in the k-12 literature, with much of this research documenting the psychometric properties of instruments (e.g., Espin et al., 2001; Madelaine & Wheldall, 1998; Thurber, Shinn, & Smolkowski, 2002). Before CBM can be implemented on a wide-scale basis in higher education, similar studies must establish the measures as reliable and valid. Without these two technical characteristics, the adequacy of the measures for decision-making cannot be determined (Charles, 1995).

Reliability. Reliability refers to the ability of results to generalize from a sample to a larger domain (Salvia et al., 2007) and can be defined as the consistency, accuracy, and stability of tests scores (Gall et al., 2003). These different types of reliability can be framed as questions. First, does each probe measure the same skill? Second, are various forms of probes equivalent? Third, can probes be scored so that various raters agree on the outcomes?

Based on the results of this study the answers to the above questions are all affirmative. Content within each probe was determined to be very consistent and measured the same skill. The various forms of the probes were also equivalent, indicating similar content across probes (Salvia et al., 2007). Similarly, inter-rater reliability for scored probes was very high (i.e., 100%), indicating that probes produced a consensus with regard to results. These findings are also similar to those documented by other researchers across settings and content areas (Espin et al., 2001; Hildebrand & Lee, 2007).

When examining the reliability results, it is likely that two of CBM’s distinctive features, objectivity and curricular match, helped establish the accuracy and consistency
of this instrument. Objectivity refers to the extent to which tests are administered and scored in an unbiased manner (Gall et al., 2003). Unlike many other instruments used to evaluate teacher preparation programs, CBMs are administered and scored using standardized procedures (Hosp et al., 2007). It is this standardization in both administration and scoring that leads to the enhanced reliability of CBM. A second variable, curricular match, also helps to enhance the reliability of CBM probes. CBM is designed to test what is taught in a given curriculum. This direct link between the evaluation system and content to be assessed enhances consistency across instructors and ultimately allows for data-based decision making.

Validity. Validity refers to the extent to which an instrument measures what it claims to measure (Salvia et al., 2007). Similar to reliability, the types of validity can be framed in terms of questions. First, does the content of the instrument match the area to be assessed? Second, does the instrument predict important outcomes accurately? Finally, does this instrument have the ability to discriminate among high, medium, and low performers?

In answering the first question, content validity was established by a panel of experts in teacher preparation. These experts examined each probe to ensure that questions assessed key areas of teacher preparation (e.g., defining and measuring behavior). Indeed, each probe contained content and skills that one would expect beginning teachers to master.

In this study, predictive validity was differentially affected by the choice of criterion variables. Overall, probes were not accurate predictors of mean course grades. In fact, only Probe 3 surfaced as a significant predictor of the average of course grades, as
well as the average in one course (i.e., Classroom Management). It was only when unrelated content was removed from the criterion measures (i.e., using the CBM post-test) that probes become significant predictors of outcomes. Course averages were comprised of a variety scores from exams, projects, presentations, and included information from several strands of the curriculum, not only defining and measuring behavior. It is possible that probes would become more accurate predictors if they contained more content related to final grades. However, this particular study was not designed to evaluate this relationship. Therefore, this statement remains speculation and must be validated empirically.

A related question was whether probes could discriminate among high, medium, and low performers. Again, no differences in performance groups emerged when course grades were used as the grouping variable. That is, while there were clearly different course outcomes (e.g., course grades of A, B, C) probes did not discriminate among these groups prior to the end of the course. It was not until the content of the outcome variable was focused entirely on defining and measuring behavior and groups were formed based on that variable that probes could accurately determine performance group status early in the semester. Importantly, these differences on performance levels could be detected as early as Probe 1, allowing instructors/students the opportunity to intervene if needed.

Implications

This study serves as an important extension of two different literatures (i.e., the teacher preparation program effectiveness literature and the CBM literature). On the teacher preparation front, this study provides additional evidence that frequently administered objective probes can be used to predict and document important training
outcomes. If teacher candidates are not making collective progress over time on key skill areas, teacher educators could reexamine their program to ensure that important skills are being taught, reviewed, mastered, and subsequently applied.

In addition, CBM gives instructors information about individual student performance and allows for systematic change when necessary. If candidates are not making progress at all or if only small changes are noted, instructors can recommend an intervention. Perhaps these candidates need to alter study habits, take notes when reading, or review material several times a week. Further, candidates who demonstrate a lack of proficiency over time may wish to choose another major.

Finally, and perhaps most importantly, CBM has the potential to change the way teacher preparation programs are evaluated. Many of the methods used by both internal and external reviewers (e.g., surveys, portfolios, observations) to evaluate program effectiveness lack reliability and/or validity data supporting their continued use. In addition, most current measures are summative in nature (e.g., surveys, examinations, and portfolios) and allow little or no time to intervene if there are gaps in the program or if teacher candidates have inadequate content knowledge or underdeveloped skills. Given the empirical evidence supporting CBM across a variety of settings, these measures have the potential to be implemented in teacher preparation programs and incorporated throughout the accreditation process in both internal and external reviews.

This study also extends the CBM literature in two important ways. First, traditional CBM probes often consist of fact and knowledge-level assessments. The present study moved beyond the assessment of facts to an examination of application of skills. This is a key finding as teacher must not only have the knowledge of behavior
management practices, but they must be able to apply their knowledge to solve real classroom-based problems. Until this point, investigations of CBM have been limited to vocabulary matching tasks (e.g., Hildebrand & Lee, 2007; Larson & Ward, 2005), oral reading, math computation, and writing probes. Results from this study are positive suggesting that application probes are viable.

A second methodological extension was that this study helped establish a new method to determine alternate-forms reliability (i.e., equivalence of probes) in CBM. Previous researchers administered weekly probes throughout the course of their studies and correlated results of adjacent probes to determine equivalent forms (e.g., Espin et al., 2001). A potential problem with this method is that student progress actually decreases the reliability of this instrument. In this study, equivalence was established by administering all probes at the same time to a large group of students. This method is advantageous in that it is efficient (i.e., can be completed in under 30 minutes), allows developers to refine a various forms of an instrument prior to administration with a target group, and also controls for learning effects.

Limitations

Although promising, these results must be viewed within the limitations of the study. First, the sample of 20 students used in this study is relatively small for an instrument validation study. However, special education is a relatively small with few students entering the field each year. Perhaps future researchers could replicate these findings across institutions, which would not only increase the sample size, but would help enhance generality of findings.
Second, a relatively narrow band of content was used in this study. Although important, defining and measuring behavior is not representative of the entire special education curriculum. As such, the effects of additional content on the technical characteristics of the probes are unknown.

*Future Research Directions*

The k-12 literature on CBM provides a blueprint for conducting research in this area. Early research on CBM sought to establish the technical adequacy of instruments relative to other more established methods of assessment. After the technical adequacy was established, researchers began to investigate how CBM could be used as a more general evaluation of student progress (e.g., could CBM predict performance on high-stakes exams?), and using CBM to evaluate the effects of interventions (e.g., as part of a responsiveness to intervention model).

Research on CBM as a method to evaluate teacher preparation programs should follow a similar course. Given the promising findings from this study, future researchers may first consider incorporating questions from additional strands of content into probes. In doing so, it is important that the methods used to develop subsequent CBM measures replicate the procedures from prior research. That is, after probes are created, technical adequacy should be re-examined. Once researchers establish these measures as reliable and valid, implementation should occur on a larger scale. It is only then that instructors should use this assessment tool in order to make decisions and intervene as necessary.

As noted earlier, the methods used to evaluate teacher preparation programs require additional study. Currently, many methods currently used have little or no research support. Part of this research could include investigations of the utility of CBM
as part of the review process. More specifically, does use of objective formative data inform reviewers such that teacher candidates are better prepared to meet the demands of classrooms? A related question, does the use of CBM increase the efficiency of the review process?

Finally, CBM research can provide a foundation for later intervention studies. Once an accurate measure of student performance is established, interventions designed to help struggling students can be developed. For example, low performing students may be given immediate corrective feedback (Scheeler, Ruhl, & McAfee 2004), taught to set goals and self-graph their data (Alberto & Troutman, 2006). More specifically, information gained through CBM may allow for teacher candidates to examine their own performance in a given course and adjust study strategies accordingly.

Summary

Teacher preparation programs are a key link between research on best teaching practices and k-12 classrooms. Effective teacher preparation programs ensure that their teacher candidates have the essential content knowledge, pedagogical skills, and dispositions prior to graduation. These skills help create effective teachers who have the potential to positively impact the academic and social skills of their students (Greenwood & Maheady, 1997). Unfortunately, the research based on assessing teacher candidates’ mastery of key content is sparse (Cochran-Smith, 2003). Moreover, much of the data used to make decisions about teacher preparation programs (e.g., examinations, portfolios, observations) lacks critical technical support.

Using formative measures, such as CBM, within a program allows teacher educators to monitor the acquisition and development of content knowledge and skills
over time. This study serves as a first step to addressing the limitations in current evaluation methods by validating the use of formative application probes, relying on the principles of CBM, in a teacher preparation program. Results were promising as probes were reliable and valid indicators of performance on measures encompassing material related to defining and measuring behavior. Given the empirical evidence supporting CBM across a variety of settings, these measures should be further developed to include additional strands of a given curriculum and hopefully, one day, used to determine the effectiveness of an entire program.
REFERENCES


http://www.teac.org/


APPENDIX A

Recruitment Script

“The purpose of this research project is to determine the relationship between quiz scores and overall achievement for students enrolled in a teacher training program at Penn State University. In order for us to validate this assessment as a useful tool for educators training teachers, we would like to ask for your participation in a research study. You will be asked to take a total of six quizzes, which will include content that you will learn in this program. Then, we would like to see if the quiz scores predict overall achievement as indicated by your final course grades.

For the study, your name will be removed from the quizzes, and the data will be assigned a specific number, which only Katie Hildebrand and David Lee will have access. All information will be confidential.

Please take a moment to review the informed consent form. If you decide to participate, we will give you a copy of the form, which you may retain for your records. If you have any questions, please do not hesitate to contact David Lee and Katie Hildebrand as the contact information listed on the consent document.

Thank you for your time.”
APPENDIX B

Informed Consent

Informed Consent Form for Social Science Research
The Pennsylvania State University

Title of Project: The Relationship Between Progress Monitoring and Overall Achievement

Principal Investigator: Katie E. Hildebrand
122 CEDAR Building
University Park, PA 16802
keh228@psu.edu
814.863.3117

Advisor: David L. Lee
226C CEDAR Building
University Park, PA 16802
davidlee@psu.edu
814.865.3567

1. Purpose of the Study: The purpose of this research is to determine the relationship between quiz scores and overall achievement as indicated by your course grades.

2. Procedures to be followed: You will be given a series of 6 quizzes pertaining to the content of your current classes over a period of several weeks. The researchers will collect these quizzes and determine how well they predict grades in your current classes (i.e., course grades). In addition to taking the six quizzes we ask that you allow us to access your class scores from your instructors.

3. Duration: You will be asked to take a series of 6 quizzes that will take no longer than 15 minutes over a period of several weeks. Therefore, this research study will take a total of one hour and thirty minutes.

4. Statement of Confidentiality: Your participation in this research is confidential. In the event of a publication or presentation resulting from the research, no personally identifiable information will be shared. Your course instructors will not be given individual results of the quizzes.

Upon giving consent for the study and after the semester is over, a codebook will be developed that will assign a number to all persons’ names. This codebook will also be stored at 122 CEDAR in a locked file. Your name will be removed from
any data and will be replaced with the assigned number in the codebook. Katie Hildebrand and David Lee will be the only people who will have access to the codebook and your data.

5. **Right to Ask Questions:** You can ask questions about this research. Contact Katie E. Hildebrand at (814) 863.3117 with questions.

6. **Payment for Participation:** You will receive $25 at the completion of the study.

7. **Voluntary Participation:** Your decision to be in this research is voluntary. You can stop at any time. You do not have to answer any questions you do not want to answer. Refusal to take part in or withdrawing from this study will involve no penalty or loss of benefits you would receive otherwise.

You must be 18 years of age or older to consent to take part in this research study. If you agree to take part in this research study and the information outlined above, please sign your name and indicate the date below.

You will be given a copy of this signed and dated consent form for your records.

Participant Signature ___________________________ Date __________

Print Name ___________________________ Date __________

Email of Participant ___________________________

Person Obtaining Consent ___________________________ Date __________
APPENDIX C

Application Probes

Version A

Directions: Read the following scenario carefully. Circle the best answer.

1. The principal, Mrs. Kirby, reads over the math lesson plans from grade 1. Which teacher wrote an objective containing all four components?
   a. Given a worksheet, the students will be motivated to solve the problems correctly for two out of three days willingly.
   b. Given a worksheet with addition problems, the students will complete it without help from the teacher enthusiastically.
   c. Given a worksheet with 20 addition problems, the students will be eager to complete the problems with at least 90% accuracy for 2 out of 3 days.
   d. Given a worksheet with 20 1x1 digit addition problems, the students will write the answers with at least 90% accuracy for 2 out of 3 days.
   e. The students will earn a passing grade by answering the problems on an addition worksheet independently.
   f. The students will be motivated to complete an addition worksheet with at least 90% accuracy for 2 out of 3 days independently.
   g. The students will write the answers to difficult addition problems on a worksheet with at least 90% accuracy for 2 out of 3 days independently.
   h. I don’t know.

2. Which of the following sentences includes a well-defined behavior?
   a. Anna did not fall during the jump rope game.
   b. Bruce is mad that he has three homework assignments.
   c. Dennis is exhibiting aggressive behaviors.
   d. Jack is sad because he misses his mom.
   e. Miss Mary is pleased with the students.
   f. Ruby does not laugh at the clown.
   g. Stan writes his name on the top of the paper.
   h. I don’t know.
3. The behavior analyst stopped by to see Juanita who exhibits hand sucking behaviors. Juanita was sucking her thumb so hard that she broke the skin on her thumb. What dimension of behavior is the behavior analyst using?
   a. duration
   b. frequency
   c. latency
   d. locus
   e. magnitude
   f. rate
   g. topography
   h. I don’t know.

4. Miss Jackson is teaching her third graders to write in cursive. She sees that Danielle has difficulty with the overcurve stroke whereas Max has trouble with the undercurve stroke. What dimension of behavior is Miss Jackson using to assess cursive handwriting?
   a. accuracy
   b. duration
   c. frequency
   d. latency
   e. locus
   f. magnitude
   g. topography
   h. I don’t know.

5. Which of the following is a NOT an observable behavior?
   a. clap
   b. cry
   c. giggle
   d. sing
   e. smile
   f. think
   g. type
   h. I don’t know.
6. Mrs. Hua wants to decrease Magda’s hand-flapping behaviors. Which sentence below matches the graph?

- The behavior generalized.
- The behavior stayed the same.
- The intervention did nothing.
- The teacher never implemented an intervention.
- There is an ascending trend.
- There is a descending trend.
- There is no trend.
- I don’t know.

7. The teacher told Emilia to sit in her chair immediately. After the teacher’s prompt, it took Emilia 45 seconds to sit. What dimension of behavior is the teacher using?

- accuracy
- duration
- frequency
- latency
- locus
- magnitude
- retention
- I don’t know.

8. Sally is learning to solve addition facts with regrouping. Her teacher notices that she writes the digits in the wrong column. What dimension of behavior is Mr. Hoffman using?

- duration
- frequency
- inter response time
- latency
- locus
- magnitude
- rate
- I don’t know.
9. Which contains all four components of a behavioral objective?
   a. During math class, Darius will demonstrate mastery by finishing the test within 45 minutes.
   b. Once the directions are provided, Kari will think about the word problems for two minutes before she begins to solve them.
   c. Once the directions are orally given, Marisol will answer 100% of the problems in 5 minutes.
   d. When given a worksheet, Ida will write the answers at a rate of 50 correct digits per minute for 2 out of 3 trials.
   e. When given a worksheet, Jamie will complete it with 90% accuracy for 2 out of 3 trials.
   f. When given directions, Melissa will solve the problems at a rate of 55 correct digits per minute.
   g. When given manipulatives, Hannah will make a pattern using at least half of the blocks.
   h. I don’t know.

10. Jason interrupts the teacher every few minutes. The teacher records the time between two interruptions. What dimension of behavior is the teacher using?
   a. accuracy
   b. duration
   c. inter response time
   d. locus
   e. magnitude
   f. rate
   g. retention
   h. I don’t know.

11. Which of the following is NOT an observable and measurable behavior?
   a. bounce
   b. concentrate
   c. exercise
   d. gallop
   e. pinch
   f. roller blade
   g. tumble
   h. I don’t know.
12. Miss Cahill decided to record Christopher’s out of seat behavior over a period of 5 minutes. She asked the paraprofessional to observe Christopher simultaneously so that she could calculate interobserver agreement. If a + means the behavior occurred and a – means the behavior did not occur, how often do they agree?

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Paraprofessional</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

a. 20%
b. 33%
c. 40%
d. 70%
e. 75%
f. 80%
g. 100%
h. I don’t know.

13. Mrs. Witter records how long Jacob’s first tantrum of the day lasts. What dimension of behavior is she recording?

a. duration
b. frequency
c. inter response time
d. latency
e. locus
f. magnitude
g. rate
h. I don’t know.

14. Ms. Patterson consistently collects data for 20 minutes three days a week on Sean’s out of seat behavior. During one observation, Ms. Patterson determines that he gets out of his seat 7 times during the 20-minute periods. Which dimension of behavior is Ms. Patterson using?

a. duration
b. frequency
c. latency
d. locus
e. magnitude
f. retention
g. topography
h. I don’t know.
15. Shelby tantrums multiple times throughout the day. The behavior specialist observed her during a six-hour school day and determined that Shelby had 1.5 tantrums per hour. What dimension is the behavior specialist using?
   a. duration
   b. force
   c. frequency
   d. latency
   e. locus
   f. retention
   g. topography
   h. I don’t know.
**Version B**

**Directions:** Read the following scenario carefully. Circle the best answer.

1. Mrs. Lyons collects baseline data on how quickly Alexis writes the answers to multiplication facts. Then Mrs. Lyons has the paraprofessional work with Alexis for an extra 10 minutes a day. Choose the description that best matches the graph.

   ![Graph](graph.png)

   a. No data were collected.
   b. She writes the answers at an appropriate rate.
   c. The intervention was never implemented.
   d. The rate stayed the same.
   e. There is a descending trend.
   f. There is an ascending trend.
   g. There is no trend.
   h. I don’t know.

2. Miss Anderson decided to record whether or not Joseph had a tantrum during each hour of the school day. She asks the paraprofessional to also observe Joseph so that she can calculate interobserver agreement. If a + means that Joseph had a tantrum during that hour, and a – means that he did not have a tantrum during that hour, how often do they agree?

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Paraprofessional</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

   a. 0%
   b. 20%
   c. 40%
   d. 60%
   e. 75%
   f. 80%
   g. 100%
   h. I don’t know.
3. Robin has difficulty completing 2 x 2 digit addition problems with regrouping. It takes her 35 minutes to complete the independent work. What dimension of behavior is being used?
   a. accuracy
   b. duration
   c. frequency
   d. latency
   e. locus
   f. magnitude
   g. topography
   h. I don’t know.

4. Which of the following is a NOT an observable and measurable behavior?
   a. eat
   b. not fly
   c. laugh
   d. run
   e. sigh
   f. snort
   g. whisper
   h. I don’t know.

5. Which contains all four components of a behavioral objective?
   a. After each word is presented orally, Rick will write the words on the line with 90% accuracy.
   b. After each word is given orally, Charlie will write that word on his paper with enthusiasm.
   c. Destiny will demonstrate mastery by completing the spelling test within the allotted time.
   d. During a spelling test, Bette will show that she knows all of the words by getting 100% on the test.
   e. Given a worksheet with misspelled words, Donna will demonstrate mastery by identifying them.
   f. When given a spelling test, Jack will be motivated to finish it with 100% accuracy.
   g. When given the directions, Sophia will complete the test independently within 25 minutes.
   h. I don’t know.
6. Ms. Morrison observes Kevin pushing down on the paper with his pencil so hard that Kevin’s knuckles are white. What dimension of Kevin’s writing behavior is Ms. Morrison observing?
   a. duration
   b. frequency
   c. latency
   d. locus
   e. magnitude
   f. rate
   g. retention
   h. I don’t know.

7. Which of the following sentences includes a well-defined behavior?
   a. Amanda has ADHD and is hyperactive.
   b. Craig has a high IQ and is very smart.
   c. Elijah is anxious for his test tomorrow.
   d. Fred is angry because his sister made fun of him.
   e. Mr. Montag is thrilled with his students’ behaviors.
   f. Rob reads the story orally and correctly.
   g. Talia does not laugh at the teacher’s joke.
   h. I don’t know.

8. The behavior specialist consistently observes Zoe every Monday morning for 45 minutes. Zoe has 2 tantrums during that observation period. What dimension is the behavior specialist using?
   a. accuracy
   b. duration
   c. frequency
   d. latency
   e. locus
   f. magnitude
   g. retention
   h. I don’t know.
9. It took Kyla 20 minutes to get quiet after Mr. Cotter put Kyla in a time-out. What dimension of behavior is being used by Mr. Cotter?
   a. duration
   b. frequency
   c. inter response time
   d. latency
   e. locus
   f. magnitude
   g. rate
   h. I don’t know.

10. Which of the following is NOT an observable behavior?
    a. cries
    b. flies
    c. grabs
    d. imagines
    e. throws
    f. whines
    g. yells
    h. I don’t know.

11. Jackie is in her teacher-training program and is learning to write behavioral objectives. Which objective contains all four components?
    a. After the lesson, the teacher will ask the students to recognize all of the shapes (circle, square, triangle, and rectangle) they learned in class.
    b. During math class, the students will complete the addition worksheet in 15 minutes without any help from the teacher.
    c. During math class, the students will earn a passing grade by answering all of the problems correctly in 20 minutes.
    d. Given a 2x2 digit addition worksheet, the students will write the answers at a rate of 55 correct digits per minute for 2/3 days.
    e. When given verbal instructions, the students will be motivated to finish all of the problems on the worksheet independently and enthusiastically.
    f. When given manipulatives, the students will use them to determine all of the answers to the problems given on a worksheet.
    g. When prompted, the students will begin solving the problems independently and the first student to finish will earn a sticker.
    h. I don’t know.
12. David laughs periodically throughout the day. His teacher, Mrs. Shugg, records the time between two of his laughs. What dimension of behavior is the teacher using?
   a. accuracy
   b. inter response time
   c. locus
   d. magnitude
   e. rate
   f. retention
   g. topography
   h. I don’t know.

13. Ryan has difficulty staying in his seat. He often walks to the door and stares down the hallway. What dimension of behavior is being used?
   a. duration
   b. frequency
   c. latency
   d. locus
   e. magnitude
   f. rate
   g. retention
   h. I don’t know.

14. Craig puts four of his fingers in his mouth all the way up to his knuckles. What dimension of behavior is being used to observe Craig?
   a. duration
   b. frequency
   c. inter response time
   d. latency
   e. magnitude
   f. rate
   g. topography
   h. I don’t know.
15. The teacher, Mrs. Foster, assessed Tanya’s reading. She found that Tanya reads 70 correct words per minute with 3 errors. What dimension of behavior is Mrs. Foster using?
   a. duration
   b. inter response time
   c. latency
   d. locus
   e. magnitude
   f. rate
   g. topography
   h. I don’t know.
Version C

Directions: Read the following scenario carefully. Circle the best answer.

1. Greg speaks so softly that the little boy next to him cannot hear him. Even when the teacher says to Greg, “Speak up,” he remains quiet. What dimension of behavior is the teacher using to observe Greg.
   a. duration
   b. frequency
   c. latency
   d. locus
   e. magnitude
   f. rate
   g. topography
   h. I don’t know.

2. Brad’s teacher is trying to reduce some of his stereotypic behaviors. She decides to target his rocking behaviors. She records the time between two rocking behaviors. Which dimension of behavior is the teacher using?
   a. accuracy
   b. inter response time
   c. locus
   d. magnitude
   e. rate
   f. retention
   g. topography
   h. I don’t know.

3. Which of the following sentences includes a well-defined behavior?
   a. Donna crosses out the misspelled word.
   b. Ellen is nervous for her calculus exam.
   c. Lou feels happy because he is the line leader.
   d. Marguerite thinks about the math problem.
   e. Ron analyzes his peer’s essay.
   f. Vinny is ecstatic because he is going to the prom.
   g. Wanda knows how to do her homework.
   h. I don’t know.
4. During a 45-minute math class, Mr. Jacobs notices that Taylor asks to go to the bathroom two times. What dimension of behavior is Mr. Jacobs using?
   a. accuracy
   b. duration
   c. frequency
   d. latency
   e. locus
   f. magnitude
   g. retention
   h. I don’t know.

5. Which of the following is NOT an observable behavior?
   a. bite
   b. daydream
   c. hit
   d. kick
   e. scratch
   f. smile
   g. tickle
   h. I don’t know.

6. Professor Kubo is teaching her students to write objectives. Which student wrote an objective containing all four components?
   a. At the beginning of the school day, Jenny will say, “Hello Mrs. Smith,” for 3 out of 5 days during the week.
   b. At the beginning of the school day as she enters the room, Jenny will happily say, “Hello Mrs. Smith.”
   c. At the beginning of the school day, Jenny will look and say, “Hello Mrs. Smith. How are you today?”
   d. At the beginning of the school day, say, “Hello Mrs. Smith, for 3 out of 5 days during the week.
   e. Jenny will say, “Hello Mrs. Smith,” to the teacher immediately when she arrives to school.
   f. Jenny will enthusiastically say, “Hello Mrs. Smith,” for 3 out of 5 days during the week.
   g. When Jenny arrives to school, she will say, “Hello Mrs. Smith,” and then continue the conversation with Mrs. Smith.
   h. I don’t know.
7. At the end of a 10 minute period, both the teacher and the paraprofessional look to see if Daniel is staring out the window. A + means Daniel is staring out the window, and a – means that he is not. The teacher and the paraprofessional then compare their records to calculate interobserver agreement. How often do they agree?

<table>
<thead>
<tr>
<th></th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Paraprofessional</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

a. 0%
b. 20%
c. 40%
d. 50%
e. 60%
f. 80%
g. 100%
h. I don’t know.

8. In October the principal compliments the students and says, “Wow, all of you are walking down the hall appropriately. You look like fifth graders.” What dimension of behavior is the principal targeting?

a. duration
b. frequency
c. latency
d. locus
e. magnitude
f. rate
g. topography
h. I don’t know.

9. The cafeteria monitor recorded Simon screaming for a total of 18 minutes throughout the lunch period. What dimension of behavior was the cafeteria monitor using?

a. duration
b. frequency
c. latency
d. locus
e. magnitude
f. rate
g. topography
h. I don’t know.
10. Which contains all four components of a behavioral objective?
   a. When given a 5th grade reading passage, Missy will demonstrate mastery by comprehending the passage when asked questions.
   b. When given a difficult 5th grade reading passage, Missy will glance at the passage consistently for three minutes and then respond to questions about it.
   c. When given a 5th grade reading passage, Missy will orally read at a rate of 100 correct words per minute with less than 3 errors for 2 out of 3 weeks.
   d. When given a 5th grade reading passage, Missy will peruse the passage at a rate of 100 correct words per minute with less than 3 errors for 2 out of 3 weeks.
   e. When given a 5th grade reading passage, Missy will silently read at a rate of 100 correct words per minute with less than 3 errors for 2 out of 3 weeks.
   f. When given a 5th grade reading passage, Missy will study the passage for only 5 minutes and then be motivated to answer all questions correctly.
   g. When given a 5th grade reading passage, Missy will understand the passage after reading it independently for only 8 minutes.
   h. I don’t know.

11. The teacher told Damon to take his fist out of his mouth. It took Damon 1 minute and 30 seconds to do so. What dimension of behavior was the teacher using?
   a. duration
   b. frequency
   c. inter response time
   d. latency
   e. locus
   f. magnitude
   g. rate
   h. I don’t know.

12. During a tantrum, the behavior analyst notes that Tiara goes to the very back of the room when she has a tantrum. What dimension of behavior is the behavior analyst using?
   a. accuracy
   b. duration
   c. frequency
   d. latency
   e. locus
   f. magnitude
   g. rate
   h. I don’t know.
13. Which of the following is NOT an observable and measurable behavior?
   a. to circle
   b. to name
   c. to number
   d. to point to
   e. to underline
   f. to wonder
   g. to write
   h. I don’t know.

14. Mrs. Morgan tries to increase the amount of time Kevin raises his hand. Which sentence below best matches the graph?

   ![Graph]

   a. The behavior generalized.
   b. The behavior stayed the same.
   c. The intervention did nothing.
   d. The teacher never implemented an intervention.
   e. The trend is flat.
   f. There is a descending trend.
   g. There is an ascending trend.
   h. I don’t know.

15. Shannon calls out rather than raising her hand. Mr. Budin observed Shannon calling out 6 times per hour. What dimension of behavior is Mr. Budin using to observe Shannon?
   a. duration
   b. inter response time
   c. latency
   d. locus
   e. magnitude
   f. rate
   g. topography
   h. I don’t know.
Version D

Directions: Read the following scenario carefully. Circle the best answer.

1. Rose scratches herself frequently but only scratches her right forearm. What dimension of behavior is the teacher using to observe this behavior?
   a. duration
   b. frequency
   c. latency
   d. locus
   e. magnitude
   f. rate
   g. topography
   h. I don’t know.

2. Which of the following is NOT an observable behavior?
   a. to cross out
   b. to draw
   c. to fill in
   d. to like
   e. to number
   f. to remove
   g. to shade
   h. I don’t know.

3. Jamie often screams in the middle of science class. The science teacher records the time between two screams. What dimension of behavior is the science teacher using?
   a. accuracy
   b. inter response time
   c. force
   d. magnitude
   e. rate
   f. retention
   g. topography
   h. I don’t know.
4. Which contains all four components of a behavioral objective?
   a. Dana will be motivated to give a $5 bill to the cashier and say thank you for 3 of 4 trials.
   b. Dana will be happy to give money to the cashier and willingly take her change for 1 of 2 trials.
   c. When given a physical prompt, Dana will give all of her money to the cashier.
   d. When asked to pay the cashier, Dana will give a $5 bill to the cashier for 3 of 4 trials.
   e. When asked to pay the cashier, Dana will remember to pay for her meal for 3 of 4 trials.
   f. Upon finishing her meal, Dana will give a bill to the cashier and remember to take her change.
   g. Upon finishing her meal, Dana will give all of the money to the waitress.
   h. I don’t know.

5. Brian is a student in Ms. Russo’s class and asks excessive questions. Ms. Russo determined that Brian asked .5 questions per minute. What dimension is Ms. Russo using?
   a. accuracy
   b. duration
   c. latency
   d. locus
   e. magnitude
   f. rate
   g. topography
   h. I don’t know.

6. Choose the description that best matches the graph.
   ![Graph Image]
   a. No data were collected.
   b. The behavior generalized.
   c. The behavior maintained.
   d. The behavior stayed the same.
   e. There is a descending trend.
   f. There is a flat trend.
   g. There is an ascending trend.
   h. I don’t know.
7. Miss Wynn looks up after every 3 minutes to see if Colleen is on-task (i.e., reading silently with head down facing the book). Miss Wynn asks Mr. Alberta to observe Colleen as well in order to calculate interobserver agreement. A + indicates that Colleen is on task, and a – indicates that she is not. How often do Miss Wynn and Mr. Alberta agree?

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>6</th>
<th>9</th>
<th>12</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miss Wynn</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>Mr. Alberta</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

a. 0%
b. 20%
c. 50%
d. 60%
e. 80%
f. 90%
g. 100%
h. I don’t know.

8. During every 30-minute lunch period, the lunch monitor observes Colin pinching Jake three times. What dimension of behavior is the lunch monitor using?

a. duration
b. frequency
c. inter response time
d. latency
e. locus
f. magnitude
g. topography
h. I don’t know.

9. Mrs. Vogel observes Darius screaming so loudly that all of the students covered their ears as he went by the classroom. What dimension of behavior is Mrs. Vogel using?

a. duration
b. frequency
c. inter response time
d. latency
e. locus
f. magnitude
g. topography
h. I don’t know.
10. The head of the history department is examining teacher lesson plans. Which teacher wrote an objective containing all four components?
   a. Given a map of the United States, each state will be labeled with 85% accuracy for 2 out of 3 consecutive days.
   b. Given a map of the United States, Samantha will be motivated label each state correctly.
   c. Given a map of United States, Samantha will label each state with 85% accuracy for 2 out of 3 consecutive days.
   d. Label each state in its proper position with 85% accuracy for 2 out of 3 consecutive days.
   e. Samantha will label each state to demonstrate that she knows the geography of the United States.
   f. Samantha will enthusiastically label each state with 85% accuracy 2 out of 3 consecutive days.
   g. Samantha will learn the location of each state by writing the name of each state on a United States map.
   h. I don’t know.

11. Which of the following sentences includes a well defined behavior?
   a. Candice knows the answer to $2x = 6$.
   b. Diane thinks about the weather.
   c. Gary imagines himself as president.
   d. Joe is friends with Don.
   e. Rob does not attend school often.
   f. Tyler orally repeats the teacher’s exact directions.
   g. Vera wonders about her report card.
   h. I don’t know.

12. The behavior specialist observes Eric and notices that Eric shouts at his peers and waves his arms up and down furiously. The behavior specialist is using which behavioral dimension during this observation?
   a. duration
   b. frequency
   c. inter response time
   d. latency
   e. rate
   f. retention
   g. topography
   h. I don’t know.
13. Which of the following is **NOT** an observable and measurable behavior?
   a. bounce  
   b. catch  
   c. dig  
   d. exercise  
   e. fantasize  
   f. tackle  
   g. walk  
   h. I don’t know.

14. Ms. Hynes told Megan that she needed to hand her book report in immediately. It took Megan 4 days to hand it in after Ms. Hynes explicitly asked her to bring it in for review. What dimension of behavior did Ms. Hynes use?
   a. duration  
   b. frequency  
   c. inter response time  
   d. latency  
   e. locus  
   f. magnitude  
   g. rate  
   h. I don’t know.

15. During a 45-minute period, Debbie whined for 32 consecutive minutes. What dimension of behavior was used to record how long Debbie whined?
   a. accuracy  
   b. duration  
   c. frequency  
   d. latency  
   e. locus  
   f. magnitude  
   g. retention  
   h. I don’t know.
**Version E**

**Directions:** Read the following scenario carefully. Circle the best answer.

1. Mrs. Chwan signaled the class to take out math materials and place them on their desks. It took Laurel 4 minutes to take out her materials. What dimension did Mrs. Chwan use to assess Laurel’s behavior?
   a. duration
   b. frequency
   c. inter response time
   d. latency
   e. locus
   f. magnitude
   g. rate
   h. I don’t know.

2. Henry’s teacher records the time between his hand-flapping behaviors. What dimension of behavior is being used?
   a. duration
   b. inter response time
   c. force
   d. magnitude
   e. rate
   f. retention
   g. topography
   h. I don’t know.
3. Which contains all four components of a behavioral objective?
   a. Madison will be happy to read 20 cvc words orally with 100% accuracy for 2 out of 3 days.
   b. Madison will read orally with 100% accuracy for 2 out of 3 days and then will answer a series of questions.
   c. Read them orally with 100% accuracy for 2 out of 3 days and be motivated to do so independently.
   d. When presented with flash cards containing cvc words, Madison will read them orally for a long period of time.
   e. When presented with flash cards, Madison will read each one orally with 100% accuracy for 2 out of 3 days.
   f. When presented with 20 flash cards containing cvc words, Madison will read each one orally.
   g. When presented with 20 flash cards containing cvc words, read each one orally with 100% accuracy for 2 out of 3 days.
   h. I don’t know.

4. Which of the following behaviors is NOT observable and measurable?
   a. to bite
   b. to draw
   c. to experience
   d. to hum
   e. to march
   f. to push
   g. to sit
   h. I don’t know.

5. Which of the following sentences includes a well-defined behavior?
   a. Becca is annoyed at Keith for being late.
   b. Grace likes the colors yellow and orange.
   c. Hilda loves going to school with her friends.
   d. Jack writes the spelling word on the line.
   e. Kay does not know her multiplication facts.
   f. Linda has two cats – Lexie and Velcro.
   g. Sebastian does not cry when he gets hurt.
   h. I don’t know.
6. Brenda calls out 2 times per minute. What dimension of behavior is the teacher using to observe this behavior?
   a. accuracy
   b. duration
   c. frequency
   d. latency
   e. locus
   f. magnitude
   g. rate
   h. I don’t know.

7. Mrs. Ely collects the following data on Jacob calling out during class. Which sentence best describes the data?
   ![Graph showing an ascending trend]
   a. No data were collected
   b. The behavior generalized.
   c. The behavior maintained.
   d. The behavior stayed the same.
   e. There is an ascending trend.
   f. There is a descending trend.
   g. There is a flat trend.
   h. I don’t know.
8. Mr. James records Scott’s noisemaking behaviors for a period of 5 minutes. To calculate interobserver agreement, Mr. James also asks Ms. Henry to observe Scott simultaneously. If a + indicates that Scott is engaging in the target behavior, and a – means that he is not, how often do they agree?

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr. James</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Ms. Henry</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

a. 0%
b. 10%
c. 20%
d. 40%
e. 60%
f. 80%
g. 100%
h. I don’t’ know.

9. Which of the following is NOT an observable behavior?
   a. to analyze
   b. to cry
   c. to dance
   d. to drink
   e. to eat
   f. to sing
   g. to whimper
   h. I don’t know.
10. Larry keeps track of how many times he raises his own hand throughout the 6-hour day. He consistently raises his hand 15 times throughout each school day. What dimension of behavior did Larry use?
   a. accuracy
   b. duration
   c. frequency
   d. latency
   e. locus
   f. magnitude
   g. retention
   h. I don’t know.

11. Leslie was humming so loudly in the art room that students walking in the hallway could hear her. What dimension of behavior were the other students using to assess this behavior?
   a. duration
   b. frequency
   c. latency
   d. locus
   e. magnitude
   f. rate
   g. topography
   h. I don’t know.

12. During a 45-minute math period Steven was out of his seat for 2 minutes, then 3 minutes, then 4 minutes, and 6 minutes for a total of 16 minutes. What dimension of behavior was used to calculate how long he was out of his seat?
   a. duration
   b. frequency
   c. inter response time
   d. latency
   e. rate
   f. retention
   g. topography
   h. I don’t know.
13. Mr. Jones circulates the classroom and notices that Amanda is writing all of her Ps backwards. What dimension of behavior is Mr. Jones using to assess Amanda’s letters?
   a. duration
   b. frequency
   c. latency
   d. locus
   e. magnitude
   f. rate
   g. topography
   h. I don’t know.

14. The behavior specialist observes Brian climbing into the very bottom of his cubby in the back of the classroom. What dimension is the behavior specialist using to observe Brian?
   a. duration
   b. frequency
   c. inter response time
   d. latency
   e. locus
   f. magnitude
   g. rate
   h. I don’t know.

15. The principal is examining lesson plans. Which teacher wrote an objective containing all four components?
   a. During recess and without a prompt from the teacher, Suneeta will socialize with peers for 2/3 days.
   b. In the morning, Jasmine will be happy to greet her peers for 3/5 days.
   c. In Algebra class, Dana will demonstrate her understanding by getting an A on the test for two consecutive weeks.
   d. Maxine will complete the problems at a rate of 55 correct digits per minute for 3 consecutive days.
   e. When given a 4th grade reading passage, Sue will orally read at a rate of 100 correct words per minute.
   f. When given a worksheet, Grace will be motivated complete it with 90% accuracy independently.
   g. When given a series of cvc words (e.g., cat, dog, pig), Mitchell will read each word orally.
   h. I don’t know.
### APPENDIX D

Procedural Integrity Checklist

<table>
<thead>
<tr>
<th>Description of Task</th>
<th>Completed?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y/N?</td>
</tr>
</tbody>
</table>

1 – Participants were instructed to leave probes face down on their desks.

2 – Probes were distributed to all participants.

3 – Directions were read precisely.

“When I say begin, read each scenario carefully. Circle the best answer. If you are having much difficulty with a question and feel as though you do not have a best guess, you may circle the option I don’t know. You will have 15 minutes to complete the probe. Complete as many questions as you can. When you finish the probe, please bring it to me. Are there any questions? You may begin.”

4 – Stopwatch was activated.

5 – When participants completed the probe, they submitted it to the investigator.


VITA

Katie E. Hildebrand

Education
2004-present            Ph.D. Candidate in Special Education, The Pennsylvania State University
2001                  B. A. in Special and Elementary Education, Juniata College

Publications


Professional Experience
2001-2004            Special Education Teacher, Students with Multiple Disabilities, East Rutherford, NJ; Students with Mild Disabilities, New Milford, NJ