

TOWARDS AN OPERATIONAL DEFINITION OF PHARMACY CLINICAL
COMPETENCY

By

CHARLES ALLEN DOUGLAS

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2011

© 2011 Charles Allen Douglas

To my daughter Sophia, to show her that with courage and perseverance,
one can achieve their dreams at any age

ACKNOWLEDGMENTS

This dissertation study has been a wonderful transformative experience. I would like to thank the members of my dissertation committee, Dr. Ried, Dr. Beck, Dr. Kimberlin, and Dr. Miller. I would like to thank my wife, Martha for supporting me through the successes and hard times as well that accompanies the completion of a Ph.D. program. Without your love and support, this would not have been possible.

TABLE OF CONTENTS

	<u>page</u>
ACKNOWLEDGMENTS.....	4
LIST OF TABLES.....	9
LIST OF FIGURES.....	13
ABSTRACT	14
CHAPTER	
1 INTRODUCTION	16
1.1 Background.....	16
1.2 Statement of the Problem	18
1.3 Purpose of Study	20
1.4 Theoretical Frameworks	21
1.4.1 Validity.....	21
1.4.2 Reliability.....	22
1.5 Research Questions	22
1.6 Significance of Problem	23
2 LITERATURE REVIEW	25
2.1 Competency.....	25
2.1.1 Defining Competency	26
2.1.2 Competency and Performance.....	28
2.1.3 Describing Competent Performance.....	28
2.1.4 Competency Summary	30
2.2 Assessment	31
2.2.1 Assessment Function	31
2.2.2 Assessment Structure	33
2.2.3 Performance Domain.....	36
2.2.4 Role of the Preceptor.....	37
2.2.5 Assessment Summary.....	39
2.3 Validity	40
2.3.1 A Defensible Assessment.....	40
2.3.2 Content Validity	41
2.3.3 Reliability	44
2.3.4 Accuracy.....	46
2.3.5 Validity Summary.....	46
2.5 SUCCESS Overview.....	47
2.6 Research Questions	49
2.6.1 Research Question 1.....	50
2.6.2 Research Questions 2 and 3.....	51

2.7 Literature Review Summary.....	52
3 METHODS.....	55
3.0 Methods Introduction	55
3.1 Phase I Methods.....	55
3.1.1 Study Participants.....	55
3.1.2 Survey Instrument Development	56
3.1.3 Data Collection	59
3.1.4 Data Analysis	60
3.1.5 Phase I Methods Summary	60
3.2 Phase II Methods.....	60
3.2.1 Study Participants.....	61
3.2.2 Simulation Development.....	62
3.2.3 Data Collection	73
3.2.4 Data Analysis	73
3.2.5 Phase II Methods Summary	76
4 RESULTS.....	79
4.0 Results Introduction	79
4.1 Delphi Panel	79
4.1.1 Round One Panel.....	79
4.1.2 Domain Specification Results.....	80
4.1.3 Performance Criteria Results	86
4.1.4 Round Two Panel.....	99
4.1.5 Domain Specification Results.....	100
4.1.6 Performance Criteria Results	101
4.1.7 Summary of Delphi Panel Results.....	110
4.2 Phase II Video Simulation.....	111
4.2.1 Expert Panel.....	111
4.2.2 Expert Panel Results.....	112
4.2.3 Summary Expert Panel Results.....	115
4.2.4 Preceptor Panel.....	117
4.2.5 Preceptor Panel Results.....	118
4.2.6 Summary Preceptor Results.....	127
5 DISCUSSION	141
5.0 Overview.....	141
5.1 Phase I - Delphi Panel	141
5.1.1 Domain Specification.....	142
5.1.2 Performance Criteria	145
5.1.3 The Delphi Panel	146
5.2 Phase II - Video Simulation.....	150
5.2.1 Video Production	150
5.2.2 Preceptor Panel.....	152

5.2.3 Reliability Results	153
5.2.4 Accuracy Results.....	159
5.2.5 Rating Scale Comparison.....	161
5.3 Future Research	163
5.3.1 Content Validation	163
5.3.2 Video Simulation - Strategy	165
5.3.3 Video Simulations - Production	166
5.3.4 Training	167
5.3.5 Rating Scale Analysis.....	169
5.3.6 Performance Levels.....	170
5.3.7 Beyond Graduation.....	171
5.3.8 National Validation Study	171
5.4 Summary and Conclusions	172

APPENDIX

A SUCCESS COMPETENCIES.....	179
B DRUG THERAPY EVALUATION AND DEVELOPMENT	180
C IRB DOCUMENTS.....	183
D DELPHI PANEL ROADMAP	188
E ANALYSIS PATH AND ASSESSMENT RUBRIC	191
F SKILL STATEMENTS AND CHECKLISTS	194
G PERFORMANCE CRITERIA GLOSSARY.....	196
H DELPHI PANEL (ROUND ONE) RESULTS	198
I DELPHI PANEL (ROUND TWO) RESULTS.....	204
J CASE STUDY SUMMARIES	209
K EXPERT PANEL SCRIPT TARGETS AND RESULTS.....	212
L PRECEPTOR PANEL RESULTS	215
M PRECEPTER RELIABILITY RESULTS	217
O PRECEPTOR ACCURACY RESULTS.....	220
P SUPERVISION PROXY RATING SCALE.....	223
Q RATING SCALE COMPARISION	224
LIST OF REFERENCES	227

BIOGRAPHICAL SKETCH..... 244

LIST OF TABLES

<u>Table</u>	<u>page</u>
3-2 Video vignettes times	77
3-3 Landis & Koch's ¹⁵⁰ "Strength of Agreement" table	77
4-1 Delphi panel (round one) demographics	133
4-2 Delphi panel (round one) practice site characteristics	134
4-3 Delphi panel (round two) demographics.....	135
4-4 Delphi panel (round two) practice site characteristics	136
4-5 Expert panel demographics	137
4-6 Expert panel - practice site characteristics	138
4-7 Preceptor panel demographics	139
4-8 Preceptor - practice site characteristics	140
H-1 Responses for Skill A	198
H-2 Responses for Skill B	198
H-3 Responses for Skill C	198
H-4 Responses for Skill D	198
H-5 Responses for Skill E	198
H-6 Responses for Skill F.....	198
H-7 Responses for Skill G	199
H-8 Responses to question 8	199
H-9 Responses to question A1	199
H-10 Responses to question A2.....	199
H-11 Responses to question B1	199
H-12 Responses to question B2.....	199

H-13 Responses to question C1	200
H-14 Responses to question C2	200
H-15 Responses to question D1	200
H-16 Responses to question D2	200
H-17 Responses to question E1	200
H-18 Responses to question E2.....	200
H-19 Responses to question F1	201
H-20 Responses to question F2.....	201
H-21 Responses to question G1	201
H-22 Responses to question G2	201
H-23 Responses to complexity question	201
H-24 Responses to reliability question	201
H-25 Responses to deficient performance question.....	202
H-26 Responses to efficiency question	202
H-27 Responses to entry-level performance question.....	202
H-28 Responses to excellent performance question	202
H-29 Responses to performance criterion question	202
H-30 Responses to quality question.....	202
H-31 Responses to supervision question	203
I-1 Responses to question A1	204
I-2 Response to question B1	204
I-3 Responses to question D1	205
I-4 Collapsed responses to question D1.....	205

I-5 Responses to question D2	205
I-6 Responses to question E1	206
I-7 Responses to question E2	206
I-8 Responses to question G1	206
I-9 Responses to question G2	207
I-10 Responses to complexity question.....	207
I-11 Responses to deficient performance question	207
I-12 Responses to excellent performance question	208
K-1 Diabetes performance targets from the script.....	212
K-2 Diabetes assessments by expert panel	212
K-3 Heart Failure performance targets by the script.....	213
K-4 Heart Failure assessments by expert panel.....	213
K-5 Anticoagulation performance targets by the script.....	214
K-6 Anticoagulation assessments by expert panel	214
L-1 Diabetes assessments by preceptor panel	215
L-2 Heart Failure assessments by preceptor panel.....	215
L-3 Anticoagulation assessments by preceptor panel	216
M-1 Reliability – competent vs. not-competent.....	217
M-2 Reliability – excellent vs. entry-level.....	217
M-3 Reliability of scale items – Skill A	218
M-4 Reliability of scale items – Skill B	218
M-5 Reliability of scale items – Skill D/E	219
M-6 Reliability of global assessment – competent vs. not-competent	219

N-1 Accuracy – Skill A.....	220
N-2 Accuracy – Skill B.....	220
N-3 Accuracy – Skill D/E	221
P-1 Skill A – Comparing rating scale with proxy scale.....	224
P-2 Skill B – Comparing rating scale with proxy scale.....	224
P-3 Skill D/E – New rating scale vs. proxy supervision scale	225

LIST OF FIGURES

<u>Figure</u>	<u>page</u>
2-1 Miller’s model of clinical performance	52
2-2 APTA’s CPI rating scale	53
2-3 APTA’s CPI criterion for entry-level assessment.....	53
2-4 SUCCESS rating scale	54
3-1 Number of video simulations for 40 preceptor raters.....	78
3-2 CE course objectives	78
3-3 2x2 student assessment tables	78
B-1 Seven competency skills	180
E-1 Delphi panel endorsement criteria	191
E-2 Data collection and analysis pathway	191
N-1 Skill A – Comparing expert and preceptor panels.....	221
N-2 Skill B – Comparing expert and preceptor panels.....	222
N-3 Skill D/E- Comparing expert and preceptor panels.....	222
P-1 Skill A – Comparing rating scale with proxy scale.....	225
P-2 Skill B – Comparing rating scale with proxy scale.....	226
P-3. Skill D/E – New rating scale vs. proxy supervision scale	226

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

TOWARDS AN OPERATIONAL DEFINITION OF PHARMACY CLINICAL
COMPETENCY

By
Charles Douglas

August 2011

Chair: L. Douglas Ried
Major: Pharmaceutical Outcome and Policy

The scope of pharmacy practice and the training of future pharmacists have undergone a strategic shift over the last few decades. The pharmacy profession recognizes greater pharmacist involvement in patient care activities. Towards this strategic objective, pharmacy schools are training future pharmacists to meet these new clinical demands. Pharmacy students have clerkships called Advanced Pharmacy Practice Experiences (APPEs), and these clerkships account for 30% of the professional curriculum. APPEs provide the only opportunity for students to refine clinical skills under the guidance of an experienced pharmacist.

Nationwide, schools of pharmacy need to evaluate whether students have successfully completed APPEs and are ready treat patients. Schools are left to their own devices to develop assessment programs that demonstrate to the public and regulatory agencies, students are clinically competent prior to graduation. There is no widely accepted method to evaluate whether these assessment programs actually discriminate between the competent and non-competent students. The central purpose of this study is to demonstrate a rigorous method to evaluate the validity and reliability of APPE assessment programs. The method introduced in this study is applicable to a

wide variety of assessment programs. To illustrate this method, the study evaluated new performance criteria with a novel rating scale. The study had two main phases. In the first phase, a Delphi panel was created to bring together expert opinions. Pharmacy schools nominated exceptional preceptors to join a Delphi panel. Delphi is a method to achieve agreement of complex issues among experts. The principal researcher recruited preceptors representing a variety of practice settings and geographical regions. The Delphi panel evaluated and refined the new performance criteria.

In the second phase, the study produced a novel set of video vignettes that portrayed student performances based on recommendations of an expert panel. Pharmacy preceptors assessed the performances with the new performance criteria. Estimates of reliability and accuracy from preceptors' assessments can be used to establish benchmarks for future comparisons.

Findings from the first phase suggested preceptors held a unique perspective, where APPE assessments are based in relevance to clinical activities. The second phase analyzed assessment results from pharmacy preceptors who watched the video simulations. Reliability results were higher for non-randomized compared to randomized video simulations. Accuracy results showed preceptors more readily identified high and low student performances compared to average students. These results indicated the need for pharmacy preceptor training in performance assessment. The study illustrated a rigorous method to evaluate the validity and reliability of APPE assessment instruments.

CHAPTER 1 INTRODUCTION

1.1 Background

Medications help improve the quality of people's health and well-being. Advances in medicine have yielded a continually-expanding body of therapies, many requiring complex administration and vigilant monitoring.¹ According a report by the Institute of Medicine (IOM), *To Err Is Human: Building a Safer Health System*²

Because of the immense variety and complexity of medications now available..., the pharmacist has become an essential resource in modern hospital practice..., and thus access to his or her expertise must be possible at all times.

The pharmacist is the most accessible health care provider and patients visit community pharmacies more than any other health care setting.³ Pharmacy comprises the third largest health profession in the United States.^{4, 5} Pharmacists are considered an essential component of the health care system in both institutional settings as well as local community practices.

The public expects educational institutions to meet their social obligation to graduate competent professionals who can provide safe and effective health care. Rising concern over the competency of health care professionals has increased public pressure for accountability^{6, 7} of educational institutions. This public outcry coincides with widely publicized reports of drug safety problems⁸ and the uneven delivery of quality health care across the nation.^{9, 10}

In response to growing public concern over the competency of health care professionals, IOM report *Health Professions Education: A Bridge to Quality*¹¹ concluded that "reform around the edges" of educational institutions will not address US health care safety issues. Rather, the IOM is promoting a change toward competency-

based education for all new health care graduates. Competency-based education focuses on integrating competency into all facets of training and assessment.¹²

Institutional leaders in the pharmacy profession have taken initiatives toward adoption of competency-based education. The Accreditation Council for Pharmacy Education (ACPE) has been engaged in a decade-long process of fundamental reform of the educational model and its goals.¹³⁻¹⁵ Based on professional standards, the Center for the Advancement of Pharmaceutical Education (CAPE) Educational Outcomes report sets the standards for accreditation, guides pharmacy education, and establishes competencies for graduating pharmacy students. The 2004 CAPE Educational Outcomes report¹⁶ categorizes the practice of pharmacy into three broad dimensions, each containing a set of competencies:

- Pharmaceutical care – Provide pharmaceutical care based upon sound therapeutic principles and evidence-based data
- Systems management – Effectively manages and uses health care system resources
- Public health – Promote improvement in health, wellness, and disease prevention in the community

The first dimension, pharmaceutical care, reflects the ongoing change in the scope of the pharmacy practice towards patient-centered drug therapy. According to the Commission to Implement Change in Pharmaceutical Education^{13, 14}, “The mission of pharmacy is the provision of pharmaceutical care” and this competency reflects a central objective of the educational mission. According to Zlatic¹⁷, “The challenge of pharmacy education today is to design, implement, and assess curricula that integrate the general and professional abilities that will enable practitioners to be responsible for drug therapy outcomes and the well-being of patients.” The objective of competency assessment is appraising the student’s performance in the clinical setting and

determining if the student is sufficiently competent to enter the profession of pharmacy.¹⁸⁻²⁰

1.2 Statement of the Problem

The goal of the IOM's competency-based reforms is graduation of competent practitioners who provide safe and effective health care to the public. The ACPE shares the goal of graduating competent pharmacists and each pharmacy program is required to assess student educational outcomes including advanced pharmacy practice experiences (APPE) performance. The 2008 American Association of Colleges of Pharmacy (AACCP) president²¹ and the American College of Clinical Pharmacy (ACCP) Educational Affairs Committee have called for a standard APPE assessment instrument.²² In order to meet this need, competency assessment needs to be applicable not only across practice settings but also across geographical regions. However, several issues must be addressed in order to realize valid and reliable assessment of student performance during APPE rotations.

First, the ACPE does not provide any guidance for the validation of an APPE assessment. Many schools of pharmacy have developed proprietary APPE assessment instruments. Preceptors may assess students differently depending on which assessment instrument is used. This complicates any attempt by governing bodies to ensure the competency of graduating students across institutions. Developing guidelines for a rigorous process to establish the validity of a national standard assessment instrument is essential.

Second, the CAPE Educational Outcomes report defined the competencies expected of an entry-level practitioner. However, does not describe the behavior a preceptor would observe of a competent or incompetent student. This opens the issue

of how to operationalize performance criteria usable by preceptors. Pharmacy preceptors should assess students based on a comparison of observed behavior with professional performance standards.²³⁻²⁵ Findings suggests preceptors may simply apply their professional judgment and this is a major reason for variation between preceptors since assessment may not necessarily use the profession's performance standards.^{26, 27} Performance criteria based on professional standards are a central element of a valid assessment instrument and the pharmacy profession has not addressed this need.

Third, the ACPE does not provide any guidance on acceptable preceptor inter-rater reliability, nor does it define acceptable accuracy levels for an assessment instrument. Inter-rater reliability is an integral aspect of a valid assessment instrument, which has little value if it produces inconsistent results between preceptors. For example, a preceptor in one APPE rotation may assess students as being not competent when in fact they are competent. Then, a similar APPE rotation with a different preceptor may assess students competent when they are not. Thus, a standard method to estimate preceptor accuracy and inter-rater reliability needs to be established.

Finally, implementing regional assessment instrument requires the burden of teaching different assessment programs to busy preceptors who are often volunteers.²⁸ The American Physical Therapy Association (APTA) requires all students and preceptors to pass an examination demonstrating knowledge of the APTA assessment system.²⁹ Taking a page from the APTA, the methods developed in this study to

estimate accuracy and inter-rater reliability can be turned around and used to train and test whether individual preceptors reliably and accurately assess APPE performances.

1.3 Purpose of Study

In response to the need for a single national assessment instrument for advanced pharmacy practice experiences, this study will take initial steps toward developing a validation method that is applicable for a national APPE assessment instrument. Based on the CAPE Educational Outcomes report, the SUCCESS instrument assesses student performance during APPE rotations³⁰ and defines thirteen broad competencies for assessment. SUCCESS, which is an acronym for Universal Clinical Competency is a continuing collaboration between all each of the four Florida schools of pharmacy. A list of the thirteen broad competencies appears in Appendix A. The focus of this study is assessment of Drug Therapy Evaluation and Development competency, which is one of the thirteen competencies. This competency is comprised of seven competency skills and a complete description appears in Appendix B. Drug therapy is essential to pharmacy education and the pharmacy profession.

To illustrate the validation process, the study will demonstrate steps to introduce a multifactorial rating scale and other new performance criteria with the Florida assessment system. These steps will provide evidence toward content validity and measure preceptor inter-rater reliability and accuracy. First, a panel will identify relevant competency skills, and evaluate and refine the performance criteria. Second, this study will develop methods to measure reliability and accuracy of clinical faculty assessments.

1.4 Theoretical Frameworks

1.4.1 Validity

Based in part on Messick's (1995)^{31, 32} unitary concept of construct validity, the Standards for Educational and Psychological Testing³³ (Standards) describe validity as the "degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests." The Standards³³ categorized five sources of validity evidence: 1) Content, 2) Response Process, 3) Internal Structure, 4) Relation to Other Variables, and 5) Consequences. These different categories of evidence are interdependent and may be complementary.

A content validation study is the first essential step in accumulating evidence of validity and represents a link between the hypothetical construct of competency and measurable indicators. Subject matter experts (SMEs) judge content validity. They evaluate the degree to which elements of an assessment instrument are relevant and representative of the hypothetical construct for a particular assessment purpose.^{34, 35} Findings suggest appropriate selection of experts to participate in the content validation study contributes to the strength of the validity evidence. For the purpose of this study, relevance refers to whether "the test items reflect the construct domain to be measured",^{34, 36, 37} and representativeness is described as the "degree to which its elements are proportional to the facets of the targeted construct."^{32, 34, 35, 38}

Messick (1996)³² describes two major threats to validity: construct irrelevant variance and construct underrepresentation. When descriptions are too narrow, the assessment may under represent the construct of interest. On the other hand, when descriptions are too broad, the assessment may introduce construct irrelevant variance.

The principal researcher will address issues of construct variance and construct underrepresentation.

1.4.2 Reliability

Reliability is integral to judging validity since the reliability of assessment scores is a factor in the interpretation of assessment scores. Reliability is necessary but not sufficient evidence for a claim of validity. There are different types of reliability measures important in assessment measurement. This study is primarily interested in estimating the inter-rater reliability of preceptor assessment scores.

Inter-rater reliability refers to the degree to which assessment scores are consistent and provide clear information to differentiate between individuals for the purpose of the assessment.³⁹ Acceptable levels of reliability vary according to the objectives of the assessment. Estimates of inter-rater reliability are dependent on the population of preceptors, the population of students, the instrument, the number of times students are rated,³⁹ and the sample size.⁴⁰ Reliability is associated with variance observed in assessment scores. Based on classical test theory, variability in a preceptor's assessment scores is the sum of the variability of the true score and that due to random and systematic error. The sources of variance in assessment scores can be separated into the variance attributed to students, preceptors, and random error individually. The proportion of variation attributed to preceptors is used to quantify the inter-rater reliability of assessments.

1.5 Research Questions

Research Question 1: What is the degree of agreement regarding content validity evidence among a panel of experts for the SUCCESS' Drug Therapy Evaluation and Development competency?

Research Question 2: Can preceptors use the modified SUCCESS instrument to make reliable assessments with in Drug Therapy Evaluation and Development competency?

Research Question 3: Can preceptors use the modified SUCCESS instrument to make accurate assessments with Drug Therapy Evaluation and Development competency?

1.6 Significance of Problem

Literature suggests more time and resources have been spent on creating competency statements for various health care professions than have been allocated to developing a sound scientific foundation for determining the validity of assessment instruments.^{41, 42} This study will address the question: Are we actually measuring with the professional standards of performance for a given competency? The outcomes of the Delphi panel will address the content validity of the operationalized performance measures used by preceptors. The methods used in this study will demonstrate a scientifically sound process to conduct content a validity study and establish a standard for the content validation of other pharmacy competencies.

APPEs account for at least 30% of the professional curriculum.^{28, 43, 44} The APPEs provide the only opportunity for students to demonstrate their competency in a clinical setting, under the guidance of an experienced preceptor pharmacist. This comprises a significant portion of students' training and provides the only opportunity for authentic assessment. This study will establish a standard method to estimate preceptor accuracy and inter-rater reliability for this critical training period. The degree of reliability and accuracy are of critical importance, since they have an effect on the interpretation of the assessment scores. The measurement of inter-rater reliability and accuracy from this

study may be used to create a benchmark for comparison and contribute to preceptor training and qualification standards nationwide.

According to the Commission to Implement Change in Pharmaceutical Education, “Professions exist to serve society and the mission of the pharmacy profession must address the needs of society and individual patients.”¹³ A key factor in the profession’s commitment to both society and the individual patient is ensuring the fitness to practice of new graduates.⁴⁵ The current fragmented and ad hoc assessment strategy does not adequately demonstrate to the public whether pharmacy graduates are actually fit for practice. An assessment instrument supported by strong evidence of validity will be a significant advancement towards the realization of the competency-based educational goals of the pharmacy profession. This study addresses essential elements related to meeting the goal of ensuring that the nation’s pharmacy schools graduate competent pharmacists and fulfill their obligation to the public’s health and safety.

CHAPTER 2 LITERATURE REVIEW

2.1 Competency

Assessing fitness for practice helps maintain professional standards and a key requisite of patient safety is ensuring fitness for practice of all health professionals.⁴⁵ The public demands safe and effective healthcare from all pharmacists, including new graduates. According to the Commission to Implement Change in Pharmaceutical Education, “Professions exist to serve society and the mission of the pharmacy profession must address the needs of society and individual patients.”¹³ Eraut (1994)⁴⁶ proposes that a profession define itself through the development and assessment of entry-level competency. Governing institutions have the responsibility to establish national professional standards for entry into the profession.

The Accreditation Council for Pharmacy Education (ACPE) is the institution responsible for accrediting pharmacy schools in the United States. The ACPE Board of Directors is comprised of major pharmacy organizations such as the AACP, the American Pharmacists Association (APhA), and the National Association of Boards of Pharmacy (NABP). In addition, the American Council on Education (ACE) has a pharmacy representative on their board of directors. Through the ACPE, the pharmacy profession establishes the professional standards for all pharmacy educational programs. Sponsored by the AACP, the CAPE educational outcomes have been incorporated into ACPE’s Standards for Professional Degree Programs in Pharmacy.⁴⁴ The CAPE competencies represent professional standards and guide pharmacy education, establish competencies for new graduates, and represent an important public statement about the role of pharmacists in today’s health care system.

Assessing fitness for award aims to differentiate high-performing students and programs. Patients may select practitioners based on the reputation of the practitioner's educational institution or individual academic achievement. High-quality programs are expected to graduate individuals who progress into leadership roles in the profession. Governing and educational institutions need to measure training programs in order to identify high-performing students and high-quality programs. In addition, measurement of educational outcomes help educational institutions understand the learning and teaching process.⁴⁷

There are two key reasons for assessing clinical competency prior to graduation. One, is fitness for practice and the other is fitness for award.⁴⁷ Assessing fitness for practice helps maintain professional standards and patient confidence. Assessing fitness for award helps identify high-performing students and high-quality programs.

2.1.1 Defining Competency

Clinical competency is a hypothetical construct that is useful in conceptualizing how well an individual performs in the clinical setting.^{24, 48-51} Gonczi²³ proposes three perspectives for conceptualizing competence: technical, underlying, and a combination of both. The technical competence perspective focuses on detailed and measurable performance tasks. Using this perspective, an assessment measures if the individual is capable of completing specific tasks related to their profession. The underlying traits perspective conceptualizes a set of underlying traits that infers the capacity of competency. Competency is inferred to be present from observation of student behavior. These underlying traits are separate from any individual technical tasks. The final perspective combines the previous two and conceptualizes competency as an amalgam of both technical skills and underlying traits.

According to Eraut (1994)⁴⁶, the ability to exercise a specific set of technical skills is based in a tradition of the skilled trades. However, this technical competency perspective has been criticized as being too narrow and attempts to “atomize” the skills of a given profession.^{46, 50} Decomposing pharmacy competencies into an exhaustive list of specific skills does not address the holistic character of a competent pharmacist since these skills are not discrete or independent, and typically several of these technical tasks are performed simultaneously.²⁴ Technical skills alone do not capture the holistic richness of professional practice.^{23, 49, 50} However, specific technical skills are easier to define and assess than underlying traits.⁵²

Eraut (1994)⁴⁶ argues the perspective of conceptualizing competence as a set of underlying traits is a development of management research and this perspective focuses on the ability of an individual to perform in the workplace. Grounded in this perspective, affective qualities such as initiative and critical thinking are viewed as predictive factors of an individual’s ability to perform competently in the occupational setting.⁴⁶ However, there are technical skills required of any competent pharmacist and these technical skills need to be considered along with underlying traits in a definition of competency. Thus, competency can be conceptualized as a combination of specific technical skills and underlying traits.²³

Eraut (1994)⁴⁶ proposes two dimensions of competence for professionals. The first dimension is, “scope of practice” relates to the range of professional roles, tasks, and situations in which competence is expected. The second dimension is “quality,” which conceptualizes competence on a continuous scale ranging from novice to expert. A newly graduated pharmacist should be autonomous and competent within their scope of

practice. The CAPE competencies represent the “scope of practice” of the pharmacy profession.

2.1.2 Competency and Performance

Miller’s (1990)⁵³ widely-cited model of clinical performance (Figure 2-1) implies there is a step-by-step progression towards competent clinical performance and is represented in the “Does” layer at the apex. The “knows” layer illustrates knowledge and the “knows how” layer represents skills. The “shows how” layer illustrates competency and the “does” layer represents competent performance in the clinical setting. Miller suggests knowledge and skills are the foundation for competency, however competent performance emerges in the presence of a patient.⁵³ Literature suggests there is a distinction between competency and competent performance and proposes that competency is a necessary but insufficient requirement for competent performance in a clinical setting.^{18, 20} Competence has been described as how individuals perform under ideal conditions and knowing that they are being assessed if they have the requisite knowledge, skills and attitudes¹⁸. On the other hand, performance describes behavior in the clinical setting^{18, 54}.

Literature suggests that assessment should focus on performance and not competence.^{18, 20} This proposition is consistent with the two main rationales for competency assessment: fitness for practice and fitness for award. However, this does not address the challenge of describing competent performance appropriate for entry-level pharmacists.

2.1.3 Describing Competent Performance

Miller (2000)⁵⁵ suggests that a complete description of competent performance needs to be realistic, general, and representative of good practice. This study assumes that

professional practice standards are representative of good practice. The operationalization of the performance criteria and rating scales for assessment is the process of writing descriptions of observable behaviors indicative of competent performance. Performance is conceptualized in terms of knowledge, skills, and attributes in the context of realistic professional tasks.²⁴

Bloom's taxonomy is a widely-used classification system that describes educational objectives in terms of knowledge, skills, and attitudes, and is not related to any specific learning theory.⁵⁶ Comparison with other taxonomies is not in the scope of this study. Three domains occupy the first level of this hierarchical system: cognitive, affective, and psychomotor. The cognitive domain refers to the intellectual aspect of performance and includes knowledge, critical thinking skills, and recall of facts, procedures, and concepts. It has six major categories: knowledge, comprehension, application, analysis, synthesis, and evaluation. Each category in the cognitive domain operates in a developmental progression.^{56, 57} The affective domain refers to the emotional and attitudinal aspects of competent performance and includes feelings, values, motivations, and attitudes. There are five major categories which also operate in a developmental progression: receiving phenomena, responding to phenomena, valuing, organizing, and internalizing values.^{56, 57} Third, the psychomotor domain refers to physical skills of competent performance. Psychomotor aspects include physical movement, coordination, and use of motor skills. These skills are measured according to speed, precision, and techniques in execution.⁵⁶⁻⁵⁸

The CAPE Educational Outcomes report identifies competencies that encompass the scope of pharmacy practice. These CAPE outcomes are consistent with Eraut's⁴⁶

(1994) first dimension of competency, namely the 'scope of practice' in which competence is expected. The objective of the operationalized performance criteria for each competency is to capture the holistic richness of professional practice.^{23, 49} This is consistent with Epstein and Hundert's (2002)⁵⁹ assertion that "Professional competence is more than a demonstration of isolated competencies." Michael Kane (1992)⁶⁰ defined clinical competence as the degree to which an individual can use the knowledge, skills, and judgment associated with the profession to perform effectively in the domain of possible encounters defining the scope of professional practice. Assessments of student performance in APPEs are grounds to generalize professional competence upon graduation.

2.1.4 Competency Summary

To address the public's need for safe and effective health care, the pharmacy profession has established competency standards for entry into the profession. APPE assessments are responsible for discrimination between students who are competent and students, who are not competent, as well as identifying high-performing students and programs. APPE rotations are an important part of the training future pharmacists.

Eraut⁴⁶ proposed two dimensions of competency. The first dimension is performance within the 'scope of practice' and the second is level of performance. The CAPE guidelines define the scope of pharmacy practice. However, the pharmacy profession does not provide guidance addressing the important issue of describing the observable student behaviors that are indicative of a competent entry-level pharmacist.

2.2 Assessment

2.2.1 Assessment Function

The terms “measurement,” “examination,” “evaluation,” and “assessment” have similar meanings in everyday language. It is important to assign specific definitions to these terms for this study. Measurement is the process of empirically observing and assigning numbers to an observed attribute, characteristic, or phenomenon according to established rules. Here, the instrument is the examination or test used to collect information. Evaluation is a judgment on the information collected.⁶¹ Assessment integrates the measurement, examination, and judgment processes. Specifically, assessment scores provide information about the student, the preceptor, and the program.

2.2.1.1 Assessment Purpose

Assessments are either formative or summative. Each has a different perspective and generates different outcomes.⁶² The assessment’s purpose needs to be clearly established and consistent with the overall aim of the APPE program.

Formative assessment aims to support the student in the learning process.⁶³ Formative assessments will include some type of feedback about the student’s progress compared to predetermined educational goals. The feedback should identify areas where the student excels and those that require remediation. Formative assessments are used when the educational aim is to further skill development.⁶³

Summative assessments aim to provide a definitive account of a student’s competence. Summative outcomes address the presence or absence of competent performance. This function is consistent with the gate keeping responsibility of APPE

assessments. Summative assessments are consistent with the purpose of assessment for fitness to practice and fitness for award.

Formative assessments focus mainly on teaching responsibilities, while summative assessments deal primarily with gate keeping. Findings suggest adjusting to new or multiple purposes diminishes the preceptor's processing of assessment information and suggests mixing formative and summative functions within a single assessment process may create conflicts.²⁶ Preceptors may experience a conflict between the relationships developed with students through teaching and their responsibilities for providing a definitive assessment of performance.^{47, 64} Students may become overly focused on demonstrating performance and, as a result, overlook learning opportunities.^{26, 63}

2.2.1.2 Assessment Reference

Interpretation of assessment scores varies depending on whether the assessment is norm-referenced or criteria-referenced. Therefore, the assessment's reference needs to be clearly stated and consistent with the overall aim of the APPE program.

Norm-referenced assessments compare an individual student's score with the average score of a reference group. Interpretation of norm-referenced assessments scores will inform how well the student performed compared to the reference group. This approach may help motivate students to assess their strengths and weaknesses from feedback in the norm-referenced assessment. This approach may help students develop the capacity for self-assessment. However, norm-referenced assessments will not provide information whether students have mastered a particular competency.

Criteria-referenced assessments compare an individual student's scores to external criteria. This external reference may be comprised of descriptions of competent

performance. Criteria-referenced assessments may be used to select students for advanced educational programs and to predict future performance. Sax⁶⁵ proposes that criteria-referenced assessments can determine whether a student has demonstrated competent performance. The purpose of the APPE assessment is to ascertain fitness for practice and fitness for award. This purpose suggests that criteria-referenced assessments are appropriate for APPEs.

2.2.2 Assessment Structure

2.2.2.1 Checklists

Findings suggest that checklists improve the reliability of assessments. The reliability of medical and nursing preceptor scores who used performance criterion improved compared to those who used an open-ended evaluation form.⁶⁶⁻⁶⁹ A substantial portion of errors by medical students in the clinical setting are errors of omission and the use of performance criteria may help preceptors identify omissions^{70,70}. Holmboe (2008)⁷¹ argues performance criterion help preceptors frame the observation of student behaviors. Checklists used in this study list the activities that a competent pharmacist would be able to complete for a given competency and expert preceptors will evaluate and refine these checklists.

2.2.2.2 Rating Scales

Norman (2005)⁷² cautions that checklists are not a replacement for rating scales in assessment instruments for medical students. Norman argues that checklist scores do not directly assess competency and adds that, "Competence ... cannot be captured by checklists, which reward thoroughness, not expertise." Various rating scales are

reported in the literature including visual analog scales (VAS), behavioral observation scales (BOS), global scales, and behaviorally anchored scales (BAS).⁷³

With a VAS scale, preceptors mark a spot on a line that identifies the student's level of performance. Descriptions may be included along the VAS line.⁷³ Using BOS, preceptors would use the simple behavior descriptions contained in the scale and rate the frequency of the behavior on a Likert-type scale.⁷³ With global scales, the preceptor rates the degree to which a student demonstrates a general characteristic.⁷⁴ BAS scales have anchors, which are behavioral descriptions for the preceptor to observe. A behavioral description is available for each performance level for a particular competency and are usually developed through a consensus process.⁷³

Literature discusses several advantages of BAS scales as compared to other assessment scales. The BAS performance criteria are descriptive, specific to the position, and introduce less construct irrelevant variance.⁷⁴⁻⁷⁶ Kingstrom and colleagues (1980)⁷⁵ concluded BAS scales do not have superior validity compared to other rating scale formats. However, Gomez-Mejia (1988)⁷⁴ proposes that the lack of difference observed between other scale formats and the BAS scale in the Kingstrom et al. study may be due to the use of overly complicated BAS performance criteria. The resulting preceptor burden may have resulted in less discriminating assessments. Gomez-Mejia's explanation is consistent with research related to the assessment of student nurses in the clinical setting. Findings from two Bondy studies (1983, 1984)^{66, 67} of nursing assessment suggest that clarity of BAS behavior definitions is a significant source of error.^{66, 67}

Literature also suggests that preceptors may use the BAS performance criteria as a simple checklist and not as the criteria-referenced assessment intended when presented with low-quality descriptors.^{76, 77} There is evidence that when the BAS performance criteria are changed to observed qualities of performance, rather than a performance criteria or simple checklist, the validity of assessments increases.⁷⁷ The advantage of BAS may result from how the behavior descriptors provide preceptors with a clearer understanding of the rating task and what behaviors to observe. Wolf (1995)⁷⁷ showed behavior descriptions were interpreted variably unless supported by illustrations of what they actually look like in practice.

Landy and Farr's (1970)⁷³ seminal article reviewed performance assessment literature and argue that there is ample evidence that BAS has an advantages in validity and reliability compared to numerical scales or adjective phrases. Landy and Farr suggest that, in addition to rigorously developed anchor descriptions, effective performance scales may also reflect participation by individuals' representative of those who will actually use the assessment scales.

2.2.2.3 Rating Levels

One of the observations established by the literature is that too few or too many categories can negatively affects scale reliability.^{73, 78} Bending (1954)^{79, 80} investigated this problem and indicated a decrease in reliability with less than three or more than seven categories. Considering the effects on both scale reliability and rater reliability, Bending argues that there is no gain in efficiency when the number of categories is increased from five to nine.^{73, 79, 80}

Evidence suggests that preceptors avoid rating students on either extreme of a rating scale. Due to end aversion bias, researchers³⁹ argue that a 5 point scales might

actually be used as a 3 point scale in practice. There is a belief that loss of response categories tends to decrease both efficiency and reliability.³⁹

The purpose of the APPE assessment is to ascertain fitness for practice and fitness for award. Assessments for fitness for practice necessitate a scale of at least two categories. These two categories would differentiate between students who are competent and students who are not-competent. Likewise, assessments of fitness for award necessitate a scale of at least two categories for competent students, which would differentiate between high-performing students and other competent students. Addressing both fitness for practice and fitness for award implies performance assessments need to differentiate at least three levels of performance: non-competent, competent, and high-performing. This suggests a scale of at least three but no more than five categories.

2.2.3 Performance Domain

Literature suggests common themes across the medical, veterinary, occupational therapy, and physical therapy healthcare professions for identifying attributes that qualify students for entry-level practice and the preceptor decision-making process according to Jette et al. (2007)⁸¹. Preceptors approached assessment in a holistic manner and without prioritizing any particular student attribute. Jette et al.(2007)⁸¹ proposes that the preceptor decision process is multifactorial. Preceptors synthesize a number of student attributes and decide whether a student has demonstrated entry-level performance.

Jette et al. commented that the student attributes revealed in their study of the physical therapy preceptor decision-making process were similar to the American Physical Therapy Association's (APTA) Clinical Competency Instrument (CPI), which

several of the participants used for student assessment. The CPI is the product of a national effort by the APTA and has been in use since 1997. Following a major update in 2006, 152 out of 212 physical therapy schools have adopted the CPI.²⁹ The CPI contains 18 competencies each containing multiple skill statements. Each competency is rated relative to entry-level practice. The CPI used BAS with six performance levels: Beginning Performance, Advanced Beginner Performance, Intermediate Performance, Advanced Intermediate Performance, Entry-level Performance, and Beyond Entry-level Performance.

The CPI multidimensional scale is based on unpublished analysis²⁹ of student data, suggesting that performance is comprised of five dimensions (Figure 2-2). Preceptors “score” the performances observed per the five performance dimensions. Scores reflect percentages (%) of necessary behavior students are capable of performing. These scores determine on which of the six performance levels the student should be assessed (Figure 2-3). If a student does not complete all of the dimensions for a particular performance level (e.g. Entry-level), then the student will not be assessed at that level. Unpublished findings from the APTA²⁹ suggest that the psychometric properties of this multidimensional scale were strong. Preceptors were able to discriminate among all six-performance levels. There was no evidence of clustering or ceiling effects.

2.2.4 Role of the Preceptor

Pharmacy preceptors are the practicing professionals who assess student performance during APPE’s and have the important responsibility of acting as gatekeepers. Pharmacy preceptors are expected to assess students based on a

comparison of observed behavior with professional performance standards as outlined in the CAPE Educational Outcomes report.²³⁻²⁵

According to the literature, preceptors may apply personal standards to their evaluations of individual students, and this is a major reason for variation between preceptors since assessments are not grounded in professional standards.^{26, 27} The Jette et al. (2007)⁸¹ study found that preceptors commonly use an intuitive decision making framework to assess students , commenting that "... gut feeling" seems to represent their cognitive integration of those characteristics into a decision about the overall adequacy of performance."⁸¹ Cross and Hicks (1997)⁸² conclude that preceptors commonly use implicit criteria in the decision making process. Preceptors would ask themselves, "Would I hire this student?" for their decision-making framework rather than use clinically based objective measures. Findings in a study by Alexander et al. (1997)⁸³ suggest preceptors assess whether the student's performance is representative of the desirable characteristics of an entry-level practitioner. However, this decision-making framework is affected by impressions of previous students and a personal perception of what constitutes an entry-level practitioner.

Preceptor experience positively affects quality of performance ratings and those preceptors who are judged as better clinicians are better at rating the job performance of others.⁷³ In addition, findings suggest preceptors with little experience or substandard clinical skills have greater idiosyncratic assessment scores that increase score variation.^{71, 84} Findings showed different student assessments between nursing educators and nursing clinical preceptors.^{66, 85} This study suggests that the educators

who often develop assessment instruments have different values and conceptualization of competency compared to practicing clinicians.

2.2.5 Assessment Summary

Assessment integrates the measurement, examination, and judgment processes. Researchers have concluded that APPE assessments should be criteria-referenced rather than norm-referenced.⁷ The assessment process should provide a summative judgment of whether a student qualifies for entry-level practice.⁷ There are a number of professional attributes to evaluate in the assessment process and are consistent with the five performance domains used in the CPI instrument. The five performance domains used in the CPI are based on analysis of physical therapy students. Some of these attributes may be a value in APPE assessment and should be reviewed by pharmacy experts.

Preceptors bring considerable experience and their participation is crucial to the assessment process. The assessment process is multifactorial and preceptors are faced with the difficult task of synthesizing a number of student behaviors. The assessment instrument should assist the preceptors in conducting reliable assessments based on professional standards.

The structure of the instrument affects the accuracy and reliability of assessments. Findings suggest the use of checklists improves assessment reliability; however, checklists are not a replacement for rating scales. APPE rating scales are most reliable when they range from three to five levels. There appears to be advantages of BAS over other scale formats.^{73, 86} Expert clinician preceptors should be recruited to develop the BAS behavior anchors. Literature also suggests that student participation is beneficial. Behavior descriptions need to be clearly written and unambiguous.³⁵ The development

of a glossary of terms may also promote a more uniform understanding of the behavior descriptions used in the instrument,⁸⁷⁻⁸⁹ such that these different items help reduce variation due to differing interpretations of behavior descriptions.^{12, 90} Assessment variation due to the preceptor rater effects decrease inter-rater reliability and accuracy results.

2.3 Validity

2.3.1 A Defensible Assessment

Assessments in healthcare education require evidence of validity to possess meaningful interpretations.⁹¹ Assessment instruments unto themselves are not strictly valid or invalid; rather, the assessment scores have more or less evidence to support a specific interpretation, such as passing or failing an APPE competency. The collection of evidence in support of validity must form a structured and coherent argument based on accepted standards of test measurement.^{18, 92}

Based in part on Messick's (1995)^{31, 32} unitary concept of construct validity, the Standards for Educational and Psychological Testing³³ (Standards) describe validity as the "degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests." The Standards³³ categorized five sources of validity evidence: 1) Content, 2) Response Process, 3) Internal Structure, 4) Relation to Other Variables, and 5) Consequences. These different categories of evidence are interdependent and may be complementary. Briefly stated, the categories of evidence are:

Content: – The "relationship between the test's content and the construct it is intended to measure."

Response Process – “Analyses of responses, including the actions, strategies, and thought processes of individual respondents or observers. Differences in response processes may reveal sources of variance that are irrelevant to the construct being measured.”⁹³

Internal Structure – “The degree to which individual items within the instrument fit the underlying constructs.”⁹³

Relations to Other Variables – “The relationship between scores and other variables relevant to the construct being measured.”⁹³

Consequences – The evaluation of unintended effects that can support or challenge the validity of score interpretations.⁹³

In addition, Messick (1996)³² describes two major threats to validity, construct irrelevant variance and construct underrepresentation. When assessments are too narrow, the assessment may under represent the construct of interest. On the other hand, when assessments are too broad, the assessment may introduce construct irrelevant variance. These should be taken into consideration in developing a case for validity. Evidence from multiple sources are required to support or challenge a case for validity.⁹¹

2.3.2 Content Validity

A content validation study is the first essential step in accumulating evidence of validity and is a major focus of this study. Evidence supporting content validity is not grounded in empirical data but rather the degree to which these scores can be assumed to tap the targeted construct.⁷⁰ An important purpose of a content validation study is to

minimize potential error variance due to poorly defined constructs of clinical competency or poorly operationalized student behaviors.

Several descriptions of content validity are available in literature and hold the same essential meaning.^{33, 35, 94, 95} Haynes et al. (1995)³⁴ provides the following definition: “Content validity is the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose.” Two key objectives of the content validation process are evaluation for relevancy and representativeness. The relevance of an assessment instrument refers to the appropriateness of its elements for the targeted construct and function of assessment.^{34, 94, 96, 97} Representativeness refers to the “degree to which its elements are proportional to the facets of the targeted construct.”^{32, 34, 35, 38} This study focuses on important elements associated with assessment of Drug Therapy and Development competency within the SUCCESS instrument (Appendix B).

Haynes et al. (1995)³⁴ points out that within the definition of content validity that the phrase, “the degree to which,” indicates that content validity is a quantitatively-based judgment⁹⁷⁻⁹⁹ A panel of preceptors who are subject matter experts (SMEs) in a particular area would evaluate a specific part of an assessment instrument related to their expertise. When the mean score of the panel’s rulings meet a predetermined threshold, the item is accepted. In failing to meet this threshold, the expert panel may refine the item for reevaluation or remove it from consideration entirely. The appropriate selection of experts to participate in the content validation study contributes to the strength of the validity evidence. The panel should be comprised of recognized SMEs from different practice settings and geographical regions in order to provide confidence

that the assessment scores and interpretations can be generalized across practice settings and geographical regions.

The tasks of the expert panel can be broken down into two parts: 1) domain specification and 2) evaluation of the behavior domains. Literature suggests construct specification is the first task in a content validation study.^{34, 100-103} The targeted competency should be unambiguously defined. The expert panel will evaluate whether all seven-skill statements are relevant and represent all the important facets of the Drug Therapy and Development competency in the SUCCESS instrument. Guion (1977)⁹⁷ argues, "the boundaries of the construct should be clear enough that the panel will understand the measurement issues and recognize when operationalized performance indicators are inside or outside those boundaries. Haynes, et al. (1995)³⁴ affirms the importance of this step and proposes, "a construct that is poorly defined, undifferentiated, and imprecisely partitioned will limit the content validity of the assessment instrument."

Two key objectives for the expert panel are the evaluation of behavior domains within the instrument for relevancy and representativeness. The panel needs to assess the relevancy of the scores for the student behaviors described in the instrument for targeted skill statements. Are all the scores from the student behaviors described in the instrument necessary to the targeted skill statement? On the other hand, are the scores for the student behaviors described in the instrument comprehensive and account for the necessary of knowledge, skills, and personal attributes for the targeted skill statement?

Guion (1977)⁹⁷ believes representativeness is observed when the “small sample of behavior observed in the measurement procedure represents the whole class of behavior that falls within the boundaries defining the content domain.” The process of assessment entails a preceptor observing samples of student behavior described in an instrument. A representative sample of student behaviors should contain the essential characteristics of the universe of possible behaviors, in their proper proportion and balance.¹⁰⁴ Lennon (1956)¹⁰⁴ argues we are really after is test score variance which is attributable to the same sources as the variance in the performance criteria. The matching of specificity appears to be a key aspect in evaluating representativeness.¹⁰²

2.3.3 Reliability

Reliability is an integral aspect of validity evidence since the reliability of assessment scores is a factor in the interpretation of assessment scores. There are different types of reliability measures that are important in assessment measurement. Downing(2005)¹⁰⁵ argues that inter-rater reliability is an essential component of validity evidence for all assessments using raters.¹⁰⁵ This study will estimate preceptor inter-rater reliability.

Inter-rater reliability refers to the degree to which assessment scores are consistent and provide clear information useful for differentiating between individual students for the purpose of the assessment.³⁹ Literature suggests that measures of inter-rater reliability in the clinical setting are often low and estimates of .80 or above are rarely achieved.¹⁰⁶ Acceptable levels of reliability vary according to the objectives of the assessment and the objective of the study is to establish benchmarks for comparison.

APPE assessment needs to discriminate between students who are competent and students who are not-competent students, as well as discriminate between high-

achieving students from the remaining student cohort. APPE assessments are based on an external criteria and the numerical differences among the preceptor scores are important in comparing inter-rater reliability benchmarks.³⁹

Based on classical test theory, the variability of a student's assessment scores may be modeled as the sum of the variability of the student's "true score", the variability of the preceptor's ratings and the variability due to error. The term "true score" is conceptualized as the average of all the observed student scores if there were a infinite number of observations and scoring was error-free.³⁹

Error is commonly categorized into two groups, random error and systematic error.¹⁰⁷ Random error is measurement error and may be due to inattention or tiredness. This error may equally increase or decrease the score. Random error is conceptualized such that these score increases and decreases over an infinite number of observations are self-canceling.¹⁰⁷ Systematic error is one type of construct-irrelevant variance. One source of may display excess reliable variance and is associated with measuring constructs other than the construct targeted by the assessment.¹⁰⁸ Preceptors may also introduce systemic error in the form of leniency/severity error, halo error, central tendency error, or other forms of bias.⁷¹

The different sources of variance in the assessment scores may be separated into the variance attributed to students, preceptors, and random error. The ratio of the variation attributed to preceptors to the total variation is used to quantify the inter-rater reliability of preceptor assessments. Estimates of inter-rater reliability are dependent on the population of preceptors, the population of students, the assessment instrument, and the sample size.^{39, 40}

2.3.4 Accuracy

A study by Noel et al. (1992)⁶⁸ measured the accuracy of student assessments by medical faculty (n=203). All participants viewed two clinical evaluation (CEX) case simulations on tape. Using a validated scoring system, overall accuracy was calculated at 32% for the group using the open-ended evaluation form and 60% for the group using the structured evaluation form. When all participants were ranked by accuracy scores, 19 out of the top 20 most accurate participants were from the group that used the structured evaluation form. For the 20 least accurate participants, 15 participants used the open-ended form and 5 participants used the structured evaluation form. Over half of all the participants rated the student performances in the two scenarios as satisfactory or superior. In both case simulations, however, enough errors were purposely included such that all participants should have rated the student performances as less than satisfactory. Thus, accuracy was less than desirable even with the improved inter-rater reliability measures for the group using a structured evaluation form. This study illustrates two points. First, a higher inter-rater reliability measure (in this case, one group compared to the other) by itself does not support a case for validity. Second, this study demonstrates the use of videotaped simulations as a sound method to measure preceptor inter-rater reliability and accuracy.

2.3.5 Validity Summary

In Sereci's (1998)³⁶ review of the development of content validity the author reported, "More than sixty years ago, it was recognized that purely empirical approaches to instrument validation were insufficient for supporting the use of a test for a particular purpose." On the other hand, proposing a case of validity based solely on content validation is also insufficient. Supporting a case for validity, an assessment

process needs a balanced of evidence and should include content validity data and empirical data.

Reasonably high preceptor inter-rater reliability estimates are required to support the meaningful interpretation of ratings. However, high reliability measurements alone do not necessarily provide evidence for an equally high level of validity. Reliability is necessary but does not provide sufficient evidence to claim validity, and one reason is that even preceptors who agree may be completely wrong.¹⁰⁵

The use of video simulations is generally accepted as an effective and economical method to evaluate assessment strategies.^{109, 110} Videotapes may be supplemented with simulated physician orders, laboratory reports, and other documents. The objective is to simulate an authentic encounter with a student during APPE rotations and provide the types of evidence a preceptor needs to make a valid assessment. The aim is to generate evidence supporting validity that is congruent with the APPE environment.

Healthcare education experts argue the need for reliable and valid student assessments.¹¹¹ Assessment instruments need to be defended with rigorous scientific methods, which demonstrate that students are capable of performing in real world clinical settings.¹¹²⁻¹¹⁴ Despite this widely held position, the pharmacy profession has yet to establish a national policy outlining acceptable validation criteria for APPE assessment instruments. Establishing a sound process for validity that is generalizable across practice settings and geographical regions will advance the practice of evidence-based education in the pharmacy profession.

2.5 SUCCESS Overview

SUCCESS is the result of the collaboration among all of the Florida schools of pharmacy. In 2001, representatives from University of Florida, Florida A&M, and Nova

Southeastern University (there were three Florida pharmacy schools in 2001) attended the AACP spring training institute to jointly developed a Web-based APPE assessment instrument. SUCCESS is an acronym for the System of Universal Clinical Competency Evaluation in the Sunshine State. The ASHP residency evaluation system served as a model for the development of SUCCESS. Beta testing was initiated in 2004 by preceptors from two of the four Florida schools.¹¹⁵

Based on the CAPE Educational Outcomes guidelines, the Florida schools developed a set of 13 competency statements that reflect an over-arching category of knowledge, skills, attitudes, and behaviors that qualify a student for entry into the profession (Appendix A). Within each of the thirteen competency statements are a number of skills. There are 96 skill statements in the system. An assessment scale with behavior anchors for three levels of performance are described for each skill. The scale assesses for one dimension of behavior: the level of supervision relative to an entry-level pharmacist.¹¹⁵ The rating scale is illustrated in Figure 2-4. Preceptors select from among the list of skill statements they expect the students to encounter during the APPE rotation. Preceptors score student performances for each skill statement based on the behavior description contained in the BAS scale. Preceptors score students at one of three levels of performance: deficient, competent, or excellent. Preceptor grades for the rotation are for comparison. The SUCCESS system assigns the students letter grade for the clerkship rotation. In addition, for safety concerns, certain skills were labeled as “critical” skills. Unsatisfactory performance in such areas could result in harm to a patient or to the practice site. Collectively, these “critical skills” act as a red flag to indicate the potential need for remediation.¹¹⁵

Since its implementation, the SUCCESS instrument has been continuously used by all the Florida schools of pharmacy. In addition, the instrument documented University of Florida student experiences in the United Kingdom, Spain, Ecuador, and Mexico. Findings from the Ried et al. (2007)³⁰ study suggest that the letter grades for the clerkship rotation provided by the preceptor correlated with the SUCCESS system-assigned grade. In addition, students incrementally achieved higher competency scores as they progressed through their scheduled rotations.

2.6 Research Questions

This literature review has drawn together information regarding the nature of competence and assessment. Wolfe and Gitomer (2001)⁷⁷ remind us that the principles for the design of performance assessments are very much in their infancy. Healthcare education experts argue the need for reliable and valid student assessments.¹¹¹ However, the pharmacy profession has yet to establish a national policy outlining validation criteria for APPE instruments. Evidence supporting or challenging assessment instruments need to be grounded in rigorous scientific methods. The aim of this study is advance evidence-based education by establishing a sound process for validation of a national APPE assessment instrument.

In the role of addressing the public's need for safe and effective health care, the CAPE Education Outcomes guidelines outline competency standards for entry into the profession. Preceptors assess student performance compared to the operationalized performance criteria from the instrument. However, the CAPE guideline does not describe the operationalized performance criteria that predict competent entry-level practice. The instrument should assist preceptors in conducting reliable and valid

summative assessments as to whether a student is qualified for entry into the pharmacy profession.

Assessment of student APPE performance in drug therapy is the focus of this study. The Drug Therapy Evaluation and Development competency in SUCCESS is an essential aspect of both pharmacy education and the pharmacy profession. Literature suggests that a valid national APPE assessment instrument should incorporate certain key features and processes.

2.6.1 Research Question 1

First, Florida's SUCCESS assessment instrument has been successfully used in the state for over four years. However, there is need to evaluate use beyond the state of Florida. This study will facilitate an expert panel to evaluate and refine the operationalized performance criteria for relevance and representativeness based on professional standards that cross practice settings and national regions. Findings suggest that the addition of a checklist improves the reliability of assessment among preceptors. An expert panel will evaluate the process of incorporating a checklist into the existing performance criteria. Currently, SUCCESS employs a one-dimensional scale and student performances are assessed by the degree of supervision required compared to that of an entry-level practitioner. Findings suggest addition performance dimensions may be indicative of competent entry-level practice. The expert panel will evaluate incorporating multiple dimensions into the existing scale. Findings suggest variation in scores attributable to differing preceptor interpretations of student performance descriptions. Literature suggests that the development of a glossary will promote a uniformed understanding among preceptors.⁸⁷⁻⁸⁹ The expert panel will

develop a glossary that will contribute to uniformed interpretation of the performance criteria.

These processes will contribute evidence to support content validity. The process will refine the operationalized performance criteria that may be generalizable across practice settings and national regions. This aspect of the study will establish guidelines for content validity studies of a national APPE assessment instrument.

Research Question 1: What is the degree of agreement regarding content validity evidence among a panel of experts for the SUCCESS' Drug Therapy Evaluation and Development competency?

2.6.2 Research Questions 2 and 3

Second, literature suggests that measurement of inter-rater reliability is an essential component of validity evidence¹⁰⁵ and measurement of accuracy would contribute evidence of validity. There are no published measures of preceptor inter-rater reliability or accuracy in the body of pharmacy literature. This study will include an estimate of preceptor inter-rater reliability and accuracy using video simulations. Video simulations are generally accepted as an effective and economical method to evaluate assessment instruments^{109, 110}. The aim is to generate evidence-supporting validity that is congruent with the APPE environment. This aspect of the study will establish guidelines for measuring preceptor inter-rater reliability and accuracy. These measures would contribute validity evidence for a national APPE assessment instrument.

Research Question 2: Can preceptors use the modified SUCCESS instrument to make reliable assessments with Drug Therapy Evaluation and Development competency?

Research Question 3: Can preceptors use the modified SUCCESS instrument to make accurate assessments with Drug Therapy Evaluation and Development competency?

2.7 Literature Review Summary

This study will accumulate evidence to support the inferences for competency in Drug Therapy Evaluation and Development of student APPE performance with the SUCCESS. Research questions that address research gaps have been presented. An expert panel will provide content validity evidence that will represent various practice settings and geographical regions. The analysis of preceptor scores will demonstrate whether SUCCESS can produce reliable and valid assessments, and whether the operationalized performance criteria are readily usable.

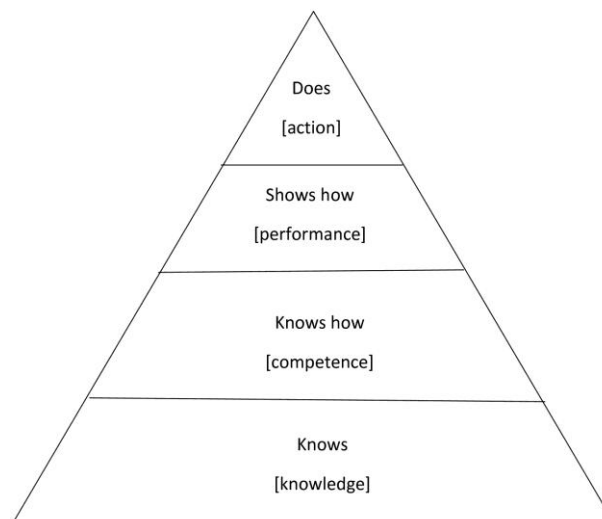


Figure 2-1. Miller's model of clinical performance

Performance Dimensions	Description
Supervision/guidance	Refers to the level and extent of assistance required by the student to achieve entry-level performance.
Quality	Refers to the degree of knowledge and skill proficiency demonstrated
Complexity	Refers to the number of elements that must be considered relative to the patient, task, and/or environment.
Consistency	Refers to the frequency of occurrences of desired behaviors related to the performance criterion
Efficiency	Refers to the ability to perform in a cost-effective and timely manner.

Figure 2-2. APTA's CPI rating scale

CPI Criterion for Entry-Level Performance
A student who is capable of functioning (100%) without guidance or clinical supervision managing patients with simple or complex conditions.
At this level, the student is consistently (100%) proficient and skilled in simple and complex tasks for skilled examinations, interventions, and clinical reasoning.
Consults with others and resolves unfamiliar or ambiguous situations (100%).
The student is capable of maintaining 100% of a full-time physical therapist's caseload in a cost effective manner.

Figure 2-3. APTA's CPI criterion for entry-level assessment

COMPETENCY- OVERALL KNOWLEDGE, SKILLS AND ATTITUDES

Skill statement-

- What should the student know?
- What should the student do?
- What does the student value?

Excellent- Practices at a level above an entry-level pharmacist into the profession (an excellent student).

Competent- Practices at the level of an entry-level pharmacist in the profession (an average student).

Deficient- Cannot practice at the entry level.

Figure 2-4. SUCCESS rating scale - Adapted from the SUCCESS web site (2009)

CHAPTER 3 METHODS

3.0 Methods Introduction

This chapter will address the motivations for the research questions, and will specifically, describes the participants, instrument development, data collection, and analysis process. There are two phases associated with this study. Each phase addresses distinct research questions and samples different populations. Both phases include a pilot study.

3.1 Phase I Methods

Phase I has a descriptive purpose with a survey design. A non-probabilistic, purposeful sample of expert preceptors was recruited. The purpose was to collect evidence to support content validity for the modified Drug Therapy and Development competency within the modified SUCCESS instrument.

3.1.1 Study Participants

Literature suggests a Delphi panel of professionals requires the participation of 15-30 individuals.¹¹⁶⁻¹²¹ Directors of experiential programs from ten schools of pharmacy will be asked to nominate ten preceptors with recognized expertise in drug therapy who also have experience as pharmacy preceptors. Schools were selected due to some degree of experience with SUCCESS. This previous experience took different forms and included partial implementation of SUCCESS for APPE assessment or participation in a SUCCESS workshop.

Nominated preceptors represented various regions across the nation and practiced in both public and private institutions. The objective was to build a non-probabilistic, purposeful panel with of expert clinicians with a mix of Board of

Pharmaceutical Specialists (BPS) and non-BPS practitioners representing a variety of practice settings, regional locations, and educational institutions. Literature suggests the panel should be comprised of the types of individuals who will use the instrument, namely practicing clinicians who are preceptors.⁷³ Literature also suggests the quality of the experts has a greater impact on the panel's outcome than the size of the panel. The nomination process defines the inclusion criterion and will ensure high quality expertise within the panel.¹²² A high dropout rate is anticipated due to the perceived burden placed on the participation of busy professionals and the study will attempt to recruit thirty participants. Participation notices and reminders were adapted from the contact scheme outlined in the Tailored Design Method by Dillman et al.(2009)¹²³ and used to promote a high recruitment rate. If funding becomes available, compensation will be used to encourage participation.

Demographic characteristics were collected and summarized to describe participants and the practice setting. Questions concerning gender, age, degree, years of practice, years precepting students, primary role with students, institution type, pharmacy practice setting, and population (of practice setting) were adopted from the AACP Annual National Survey of Volunteer Pharmacy Preceptors.²⁸ However, comparison with the AACP survey is limited due to a low response rate by the convenience sample.

3.1.2 Survey Instrument Development

The Delphi method was originally developed by the Rand Corporation in the 1950s as a means of extracting opinion from a group of experts.¹²⁴ According Linstone and Turoff (1992)¹²⁴, there are four key elements of this method. The first element is anonymity; panel members are unaware who contributed a given statement or opinion.

This is to avoid peer influences and to allow panel members to change their opinions without fear of social consequences for changing opinions. The second element is iteration. Panel member interaction is carried out through series of questionnaires. Discussion items are refined in each successive round. There may be several survey rounds and the panel members will complete essentially the same survey each round. The third element is group feedback. Each successive survey round contains a summary of the responses and opinions from the previous round. Group feedback reports the percent supporting each opinion in the survey. The final element is endorsement. A definition of approval from the panel is set prior to the panel's start. When this threshold is achieved, the item is adopted. An item under discussion will be continually refined, based on panel input, until the approval threshold is achieved.

Conducting a Delphi panel appears to be more art than science. There are no concrete guidelines for the panel inclusion criteria such as, panel size, number of rounds, feedback or agreement definitions.^{116-120, 124-128} It appears that there are no published accounts of a Delphi panel with the exact same purpose as this study, although there are several similar studies.¹²⁹⁻¹³⁷ The Delphi method provides for structured communication among panel members and designed to exploit the advantages of group-based work while overcoming its disadvantage.¹²⁴ This goal is accomplished by focusing attention on the free exchange of ideas along with hindering any single group or individual from dominating the conversation. The design of the Delphi survey instrument is based on the Reactive Delphi model. The Reactive Delphi method, as defined by McKenna (1994)¹¹⁸, is a method that requests panel members to react to previously prepared material rather than ask them to generate original material.

With this strategy, panel members were asked for their professional opinion on the materials presented.

In this study, the Delphi panel was asked for their opinion on specific competency skills for evaluation and the performance criteria for assessment. Some questions are associated with more than one target subject and reflect the cross method strategy of the Delphi survey. This gave the investigator an opportunity to evaluate for major themes, discrepancies, help provide a robust description of the study question. The competency skills were adapted from the Drug Therapy and Development competency. The performance criterion included the checklist, rating scale, and performance levels. A physician at the University of Florida Shands Medical Center developed the checklist. The rating scale and performance levels were adapted from the physical therapy assessment instrument. The intention was to seed the discussion with the new items in the context of Drug Therapy and Development competency assessment.

The principles of survey development outlined in the Dillman et al. (2009)¹²³ Tailored Design Method were used to create the survey instrument. In addition, literature suggests participants should be provided with materials related to the conceptual basis for competency and the operations of the survey.¹²² A brief participant's guide is available in Appendix C. Literature also suggests that the quality of the feedback from previous survey rounds is a crucial feature for optimal participation. The objective of the feedback process is to provide relevant data that the panel can readily interpret. This feedback is both quantitative and qualitative. The resulting summary of the percent response for a single survey question from previous Delphi survey rounds is an example of quantitative feedback. On the other hand, an

opportunity for panel members to contribute comments is provided for every survey question, and the researcher conducts a thematic analysis of the panel members' comments, which are coded according to key words or phrases. . This summary comprises the qualitative feedback.

A pilot test was conducted to uncover any operational issues and assess the readability of the instrument. The University of Florida Director of Experiential Programs and the AACCP Director of Academic Affairs and Assessment were asked to nominate preceptors for the pilot test. Three experienced preceptors were recruited for the pilot study.

3.1.3 Data Collection

The UF Institutional Review Board (IRB 02) approved the study and the approval letters are presented in Appendix C. The principal researcher used the third-party web service Survey Monkey¹³⁸ to administrate the on-line survey. The on-line survey contained the informed consent form, an explanation of the study, and the actual survey.

When a survey round was distributed, panel members were asked to submit their responses within 10 calendar days. Distribution of the next survey with feedback was scheduled within three calendar days of the previous survey's closure. This cycle was scheduled to continue up to three survey rounds or completion of the Delphi panel tasks. Literature suggests that at least two survey rounds will be necessary for completion of the Delphi panel's agenda. This is consistent with several similar Delphi studies with a greater participation burden that required three Delphi rounds.^{131, 135-137,}

139, 140

3.1.4 Data Analysis

The definition of panel endorsement was set prior to starting the Delphi process. Several methods are cited in healthcare literature to quantify content validity evidence.¹⁴¹⁻¹⁴³ Literature suggests that percent agreement is a reasonable definition of panel endorsement for this application and illustrated. This study used an 80% agreement threshold for panel endorsement of an item.

3.1.5 Phase I Methods Summary

The intention of this study was to collect content validity evidence for the modified SUCCESS competency Drug Therapy Evaluation and Development. The principal researcher used an online Delphi method since it is an effective and economical way to facilitate discussion among experts. Preceptors were nominated by schools of pharmacy to form a diverse group of recognized experts. The Delphi panel evaluated and refined the domain specification and performance criteria (Figure 3-1). The aim of Phase I, is to demonstrate a scientifically sound process that could be replicated by other academy members for conducting a content validation study for a national APPE assessment instrument.

3.2 Phase II Methods

Phase II used a non-experimental, descriptive design. Data collected in Phase II addressed Research Questions 2 and 3, namely establishing benchmarks of reliability and accuracy for future comparisons. The objective of the second phase is to collect data from preceptors to estimate assessment reliability and accuracy. Preceptors observed simulated student performances, and used the new rating scale to assess simulated student clinical performances.

3.2.1 Study Participants

A non-probabilistic convenience sample of typical preceptors was recruited to participate in an on-line Continuing Education (CE) course. In addition to ACPE credit, participants were offered a \$100 eGift certificate from AMAZON.COM. The Texas Pharmacy Association and the principal researcher marketed the CE course. The Texas Pharmacy Association distributed promotion material on the TPA WEB site and was included on regular email lists. The principal researcher contacted a number of schools to distribute promotional materials. Except for two Florida-based schools, all schools declined to forward promotional materials. These two schools provided the researcher with preceptor mailing lists. Citing limited resources, the schools asked the researcher to distribute promotional materials directly to preceptors.

The target number of preceptors to recruit was based on the requirements of the analytical methods. This study used Intra-Class Correlation (ICC) to estimate reliability and the Fisher Exact Test to estimate accuracy. The Fisher Exact Test is an exact test and not sensitive to small sample sizes. However, the ICC is sensitive to a number of factors and requires sample size estimation. The classical approach to sample size calculations are based on statistical power and minimal detectable difference. There is a growing trend to estimate sample size based on confidence intervals rather than classical hypothesis testing, and this is the strategy used in this study.¹⁴⁴

Sample size was calculated based on a modification of a formula described by Bonett (2002).¹⁴⁵ The study selected an α of 0.05, β of 0.2 and a confidence interval of 0.5. Figure 3-2 illustrates the number of video simulations needed with 40 preceptor raters and an ICC ranging from 0.5 to 0.9 in order to assess 9 examples of each skill. The unit of analysis is the assessment score based on the new rating scale. The

objective of the sample size estimate is to allow the study to compare and contrast assessment scores based on the new rating scale between all skills for all students. This would allow analysis of the results of each competency skill illustrated by each of the nine student vignettes.

3.2.2 Simulation Development

3.2.2.1 Expert Panel

A non-probabilistic convenience sample from the Delphi panel was recruited. Invitations were sent to 12 preceptors based on the quality and quantity of comments in the Delphi panel. The study needed preceptors who were able to provide usable advice with a rationale for their opinions and did not shy away from giving advice. Participants were offered \$400 compensation for their time and effort. Ten preceptors accepted my invitation to join the expert panel, however only seven completed the project.

The expert panel had two main tasks. The first task was to review and edit scripts prior to filming. The objective of each video vignette was to simulate an authentic encounter between a preceptor and a student during APPE rotations. The filmed simulations and supporting materials were evaluated for authenticity and appropriateness. The video script provided evidence a preceptor needs to make a valid assessment. In the second task, the expert panel watched the video vignettes and scored the performance using the new performance criteria. These assessment scores were compared with a group of typical preceptors.

3.2.2.2 Case Study Development

The objective of the study was to evaluate a new performance criterion for the assessment of the seven skills under the SUCCESS Drug Therapy and Development competency. The Delphi panel evaluated and revised the competency skills and

performance criteria. However, producing video vignettes that illustrate all the competency skills would have made the videos prohibitory long. This would hamper recruitment of typically busy preceptors in the next step in the study. In the researcher's opinion, illustrating only Skills A, B, and D/E would provide the most complete opportunity for assessment within a reasonable time limit. The expert panel was asked to focus on the quality of the scripts and not to be concerned about any time limitations.

The expert panel was sent summaries of six case studies and asked to select their top three preferences. Panel members were asked for their opinion on which case studies held the best potential as a vehicle to illustrate Skill A, Skill B, and Skill D/E. The poll was informal and intended to start a dialog. The expert panel preferred to use a case study on diabetes, heart failure, and anticoagulation. The first two cases focus chronic conditions and the third case offered an opportunity in an inpatient setting. Panel members chose to develop scripts for these subjects to offer an opportunity to consider an inpatient setting. The three cases not selected dealt with hypertension, chronic obstructive pulmonary disease, and pain management.

The expert panel was consulted regarding an appropriate setting for the video vignettes. Panel members who were available for consultation suggested capturing the give-and-take dialog between a student and their preceptor. The vignette would capture the responses of the student from preceptor questions. One preceptor suggested a particular framework of preceptor questions. Developed by the Association of Health System Pharmacists, this strategy outlines a series of inquiries to help pharmacists evaluate drug therapy. The objective of the video vignettes was to show different student behaviors for assessment. On purpose, the variation in dialog within each

vignette illustrates different levels of student clinical performance and not the performance of the preceptor. To accomplish this goal, the preceptor prompts for each case study were the same regardless of the student or the performance level illustrated.

3.2.2.3 Script Development Process

Scripts were distributed to the expert panel within the period for completion. After the responses were compiled, the new version was sent back to the expert panel review. The principal investigator called several members of the panel for clarification of terms and comments. At the end of each script, the expert panel was asked to respond to four questions. A summary and sample of comments follows.

Question 1: Do these vignettes illustrate a typical interaction between the preceptor and the student?

Question 2: Is there enough information illustrated in the script to judge student performance?

Question 3: How do you tell the difference between a competent student performance (e.g. Excellent and Entry-Level) and a not competent student performance (e.g. Deficient)?

Question 4: How do you tell the difference between an “Excellent” student performance and a regular “Entry-Level” student performance?

Question 1: All members of the expert panel agreed on the scripts, which, for the purpose of this study, represented typical interaction between a student and a pharmacy preceptor. Many comments suggested that preceptors spend considerable time asking questions to elicit responses from students. The behavior of the preceptor is not depicted in the vignettes. A sample of comments follows.

1”. ... I have to prompt students to get them to talk.”

2. “For an [Entry-level] student I would expect a whole lot more prompting from the preceptor.”

2. “Typically [even for Entry-level] adequate students are not confident in what they know so I have to ask tons of questions but after they realize they really did know the info they just didn't put it all together in the beginning.”

Question 2: All members of the expert panel agreed on the scripts, which, for the purpose of this study, presented enough information for a typical preceptor to assess performance. The panel agreed that the scripts were ready for filming. There was some reservation that assessment of one example) of a student, performing a skill may be difficult. The first draft of the script included a scene where the preceptor gives a history student performance that is not in the video. This was an attempt by the researcher to give the impression of multiple performances to help make assessment feel more realistic. However, the panel uniformly rejected this scene. The panel explained that this prevented the preceptor from basing their assessment on their observations and introduced bias. A sample of comments follows.

1. “I think there is enough info illustrated in the script to judge student performance. However, one caveat is, the preceptor will have judged multiple cases to determine their level by the end of a rotation. Sometimes only one is hard.”

2. “... definitely yes”

3. “Yes with modifications... “

4. “The preceptor summary introduces bias...”

Question 3: The expert panel related several behaviors they want to observe for evaluation of whether a student is competent or not. First, the competent student should understand what the patient’s medications and the therapeutic objective. Medication therapy is seen as a balance of benefits and risks based on the principles of evidence-based medicine. Knowledge of medication should be consistent and detailed. Beyond a

technical understanding of therapeutics, the competent student has a patient-centered perspective. They will inquire about patient compliance and investigate barriers to access to medication. A small sample of comments follows.

1. "At some point (either history or in discussion later) I would like [the competent student] to make some sort of comment on her [the patient's] compliance. . . . No point in adding meds if the patient isn't going to take them."
2. " [For a competent student], I would also want to see her bring in [patient] her glucometer records to download them and follow up with her that she is monitoring at home at least monthly and not wait 3 months for the HgbA1C. "
3. "I would expect [the non-competent student] to just to handout the pamphlets..."
4. "When the students are giving their recommendation, I never let them get away with "intermediate acting insulin" they would need to specify which product the patient will be receiving."
5. "[a competent student] . . . understands what the medications are and what they are for. "
6. ". . . I drive for understanding of the risks and benefits. And also include compliance and financials. "
7. "Have him [the non-competent student] prioritize a different problem. . . ."
8. "Have him [the non-competent student] pick a very expensive insulin (like lantus) and not have considered if her insurance would pay for it or what her copay would be or justify it by him stopping the other generic meds which have low co-pays."
9. "The student [the non-competent student] has the right meds but an out of this world dose. Have him recommend a much higher dose since he is stopping all of her meds. (like ummm I think 1 Unit per kilogram BID) but not know why he chose that dose."
10. ". . . giving me [by the non-competent student] a long list of factors that may affect INR [laboratory results] but are not relevant to this case."

Question 4: Every member of the expert panel expressed difficulty-describing behavior that was different between Excellent and Entry-Level performance. Both the Excellent and Entry-level students understand the patient's medications are and the therapeutic objective. They perceive medication therapy as a balance of benefits and risks based on the principles of evidence based medicine. Knowledge of medication should be consistent and detailed. However, there were several suggestions to show examples in which Entry-level students ask for assistance, or in which the Excellent student recognizes when medical documentation does not contain all the answers and one needs to ask the patient. It was also suggested that the script show how the Excellent student reached out to connect with the patient to ensure compliance, verify educational needs, and recognize barriers. Comments suggested that Excellent students should have a grasp for the strengths and weakness of laboratory tests and an advanced understanding of practice protocols as compared to other students. A small sample of comments follows:

1. "Perhaps also a comment of what [patient's] insurance looks like or product costs (i.e., If deciding between 2 insulins, which would be better for that reason). "
2. "An excellent student would verify whether the patient needs diabetic education. . . just because Joyce is a nurse, I would assume she would need more diabetic education, she may think she knows what to eat but nurses tend to be the worst and she would benefit from education. "
3. ". . . I may even suggest [the excellent student] phone the patient at home after 1 week to see how she is doing with her new regimen. "
4. " [For an excellent student], I would also want to see her bring in her glucometer records to download them and follow up with her that she is monitoring at home at least monthly and not wait 3 months for the HgbA1C. "

5. “ When talking about self management activities, excellent students may mention compliance and written info, but perhaps also suggest other compliance devices like a pill box or give him a chart to record his meds as he takes them may be of benefit also for an excellent student. “
6. “[For an excellent student]. . . “since he has CAD, I would probably continue his aspirin but there has been some controversy surrounding ASA recently. Is that what you are currently recommending here at your clinic?”
7. “Considers impact of non medication factors,. . . consistent diet is the key [in the anticoagulation case] “
8. “. . . focusing on the patient at hand. He is not significantly elderly [in the anticoagulation case]. After the patient details, I as a pharmacist would be asking her, why do you think today his INR is 4.6? “
9. “He [an excellent student] needs to talk more about diet, signs and symptoms of clotting, lifestyle issues with warfarin, compliance, etc. What questions do we need to ask him? “

3.2.2.4 Rating Scale and Student Behavior

The expert panel was asked for input on student behaviors that would illustrate different levels of performance for each of the five dimensions (see Appendix F and G) of the new rating scale. The poll was informal and intended to help the researcher effectively use the feedback from the panel. Brief summaries of the strategy used to illustrate behavior for the different levels of performance with each of the rating scale follows:

Supervision: This is the level and extent of assistance required by the student to achieve entry-level performance. The rating has four levels of assessment (see Appendix F and G). The supervision rating is different from all the other scale behaviors. The preceptor bases the degree of supervision on behaviors captured in the other rating scale items. For example, the preceptor may observe a student recommending the wrong dose or missing important information in the medical records. In this case, the

preceptor would make the appropriate assessment for consistency or complexity in the rating scale based on these observations. The preceptor may also assess the necessary degree of supervision differently based on these observations, which is, in effect, a global assessment of the student.

Many comments suggested that preceptors spend considerable time asking questions in order to elicit certain responses from students and that the degree of prompting is strongly associated to the degree of supervision needed by the student. Previously explained, preceptor prompts were identical regardless of the student performance level. This is unfortunate since many preceptors use the degree of prompting as a strong indicator of the student's level of performance. A sample of comments follows:

1. "Typically my worst students take the longest to get through a case because they miss so much or I have to drag the information out of them. "
2. "Students who are adequate are likely looking for positive reinforcement and will ask me what I would normally do if they are not confident in their response and this would indicate a little more supervision is being required. Perhaps a response from the student as "Can you explain what you mean by this? "
3. ". . . I have to prompt students to get them to talk. "
4. For an [entry-level] student I would expect a whole lot more prompting from the preceptor.
5. "Typically adequate students are not confident in what they know so I have to ask tons of questions but after they realize they really did know the info they just didn't put it all together in the beginning. "

Quality: The objective of this scale is to rate the degree of knowledge and skill demonstrated by the student's performance. The rating scale has three levels of assessment (see Appendix F and G). As previously mentioned, the competent student

understands what the therapeutic objective for a patient's medication. Many students can recite textbooks; however, preceptor comments suggest several key behaviors. First, the student should understand the medications therapeutic objective. Second, the student should understand the benefits and risks of medication therapy. Lastly, the student should understand the therapeutic objective of the medication for the patient. A sample of comments follows:

1. "I typically don't care for [students] to tell me statistics if they can tell me why they need to do x or y for a patient. "
2. "I drive for understanding of the risks and benefits. "
3. "When the students are giving their recommendation, I never let them get away with "intermediate acting insulin" they would need to specify which product the patient will be receiving. "
4. "So when discussing the problems I would expect some statement from the student that the [patient] has signs and symptoms which require drug therapy. . ."

Complexity: The objective of this scale is to rate the proficiency with simple or complex cases. The number of elements of the case (e.g., simple or complex) must be considered relative to the patient, task, and environment. The rating scale has two levels for assessment (see Appendix F and G). It was suggested that a deficient student would fail to address all the relevant factors with a question for a complex patient. A sample of comments follows.

1. "When Joseph can't list all the drugs that exacerbate heart failure, the preceptor may want to encourage but then respond with the other classes or may request he look into this and report back more thoroughly by the end of the day. "
2. "I may include some encouraging statements throughout – (e.g., when she is listing all the diff meds that exacerbate heart failure, to indicate she got most if not all of them."

3. "Have him talk about hypoglycemia for the question about considerations and completely miss the weight gain. "

Consistency: The frequency of occurrences of desired behaviors (e.g. infrequently, occasionally, and routinely). The rating scale has three levels of assessment (see Appendix F and G). One suggestion was to show that this behavior is to have the students give a good rationale for a medication and then select the wrong dose without any rationale. A sample of comments follows:

1. "Have him [student] pick the right insulin ... Have him recommend a much higher dose. . . (like ummm I think 1 Unit per kilogram BID) but not know why he chose that dose. "
2. "Have him [student] talk about hypoglycemia for the question about considerations and completely miss the weight gain. "
3. "Have him skip meds that he doesn't think are related (like ASA or statin) and have preceptor ask if that is all the meds. "

Efficiency: This is defined as the ability to perform in an effective and timely manner (e.g., inefficient/slow, efficient/timely). The rating scale has two levels of assessment (see Appendix F and G). The student's inability to make appropriate time management decisions were suggested as one-way to demonstrate efficiency in the video vignettes. Another example is to have the student give a messy and disorganized presentation to the preceptor. The following is a sample of participant's comments:

1. "Also, these students [the non-competent student] tend to make a lot of excuses so instead of him just saying he needs extra time for the other patients, I would expect him to say something like I haven't gotten to them as I've been working on the journal club article for next week when the patients are due for an appointment in a few minutes (terrible prioritization skills). "
2. ". . . these messy students show how unorganized they are. "
3. ". . . can't find things without Google. "

3.2.2.5 Randomization

The final step prior to filming was randomization. During script development, the performance level for each student was the same for every skill. For example, the performance level for Mary, Patricia, and Linda was Excellent for all skills. This was to make the job of editing easier. More importantly, preceptors and assessment instruments should reliably and accurately assess competency skills on the merits of the particular student's performance. A given assessment should be based on the performance criterion (e.g. skill description and the checklist) and not influenced by prior behavior of this or other students. The first 3 video vignettes were not randomized. This was done to help evaluate the effect of randomization in the study. The remaining 6 video vignettes were randomized by skill resulting in an almost equal number of examples of excellent, entry-level, and deficient performances (Table 3-1). The expert panel was blinded to the randomization.

3.2.2.6 Video Production

Nine students from the School of Pharmacy at the Southwestern Oklahoma State University (SWOSU) were recruited. The principal researcher obtained the approval of the SWOSU IRB and used the consent documentation approved by the University of Florida IRB (Appendix C). These third-year students were compensated \$100 for their time. Students completed rehearsals and filming in one day. An experienced SWOSU faculty played the role of the preceptor. The longest vignette was over 10 minutes in length, the shortest video was 5 minutes, and 10 seconds long (Table 3-2). It took 70 minutes to play all the vignettes consecutively. This was much longer than originally expected. However, no time limit was placed on development of the scripts.

3.2.2.7 Expert Panel Data Collection

The principal researcher used the third-party web service Survey Monkey¹³⁸ to administer the on-line survey. The on-line survey contained the informed consent form, an explanation of the study purpose, access to video vignettes, and the survey. Data collected from the expert panel are presented in the RESULTS chapter and are compared with the assessments collected by the preceptor panel.

3.2.3 Data Collection

The Texas Pharmacy Association's (TPA) web site was used for presenting the webinar and, as mentioned earlier, Survey Monkey¹³⁸ was used to show the video vignettes and collect assessment data. Preceptors were given CE credits and a \$100 AMAZON.COM eGift certificate for participation. The university's Institutional Review Board (IRB02) approved the study and the approval letters are presented in Appendix C. The CE program employed a variety of teaching formats, including topic presentations, video illustration of APPE performance, and expert discussion. The teaching objectives are shown in Figure 3-3.

3.2.4 Data Analysis

Figure G-1 in appendix G illustrates the data collection and analysis pathway. The expert panel and the preceptor panel members viewed students' performances from the same video vignettes. Each of the nine video vignettes illustrated student performance of Skills A, B, and D/E. Performance evaluations based on the new performance criteria were collected. Performance levels for each skill are were based on the assessment rubric discussed previously.

Assessment data is collapsed into two 2X2 tables for analysis. One table compares competent vs. not-competent assessments and the other table compares

excellent vs. entry-level assessments. For the evaluation of reliability, the single-measure Interclass Correlation Coefficient (ICC) is estimated for both 2X2 tables. The estimates of reliability from the ICC addresses research question two. For an evaluation of accuracy, the Fisher Exact test is estimated for both 2X2 tables. The estimates of accuracy from the Fisher Exact test address research question three.

Research Question 2: Can preceptors use the SUCCESS instrument to make reliable assessments of Drug Therapy Evaluation and Development competency?

Research Question 3: Can preceptors use the SUCCESS instrument to make accurate assessments of Drug Therapy Evaluation and Development competency?

3.2.4.1 Reliability test

The ICC is a widely used statistical test to estimate the degree of inter-rater reliability.¹⁴⁶⁻¹⁴⁹ Shrout and Fleiss(1979)¹⁴⁹ describes six forms of the ICC and give a decision rubric for selecting the appropriate form. The first decision is specifying whether the study should analyze the data using a one-way or a two-way model. The researcher would like to separate error into systematic and random error in order to measure single-measure reliability, and this objective requires a two-way model. Each preceptor needs to rate all nine video vignettes. The single-measure ICC compares both the reliability of raters and the agreement between raters for a criterion-based rating scale as used in this study.¹⁴⁶

The second decision is whether the study conceptualizes the preceptor raters as a fixed effect or a random effect. The aim of this study is to produce results that may be generalizable; this is why a random effects model will be utilized by the study. The selection of the unit of analysis is the third decision. In practice, only one preceptor rates student performances during a clerkship rotation, as compared to other healthcare

professions in which multiple preceptors assess student performance. The unit of analysis is the individual preceptor's score.

For the measurement of inter-rater reliability, preceptors and students are treated as random factors in a two-way random effects model. Since the three performance levels (e.g. excellent, entry-level, and deficient) are hierarchical and this is a violation of ICC assumptions, the study will collapse the scores into 2x2 tables for comparison. Competent students will be compared to non-competent students and excellent students will be compared to all other students as shown in Figure 3-4.

The ICC will measure the degree of agreement between preceptors' assessment of students scores. The null hypothesis is $H_0: \rho = 0$ and the alternative hypothesis is $H_a: \rho \neq 0$. This test will demonstrate if the reliability of assessment among preceptors is greater than by chance alone. For the purpose of discussion, Landis & Koch¹⁵⁰ created the "Strength of Agreement" (Table 3-2). This table can be used to interpret ICC results. According to the authors, these divisions are arbitrary. However, they are useful for discussion.

3.2.4.2 Accuracy test

The Fisher Exact test measured preceptors' accuracy and were compared with expert panel members' scores from the pilot study. Fisher's Exact test is an appropriate non-parametric test to compare differences in the distribution of scores between groups whose matrix may contain cells with values less than five. As previously mentioned, assessment data is collapsed into two 2X2 tables for analysis. One table compares competent with non-competent assessments and the other table compares excellent to entry-level assessments. The Fisher Exact test treats this collapsed data as nominal data.

The null hypothesis is H_0 : the probability of observing no difference $\leq 10\%$ in the scores between preceptors and experts. The alternative hypothesis is H_a : the probability of observing a difference $> 10\%$ between preceptors and experts. This test will show if typical preceptors assess students with the same accuracy, within a 10% range, as compared to the expert panel beyond chance alone. However, there are no widely accepted criteria defining acceptable accuracy for assessment. The principal researcher will note results up to the 30% level as a form of sensitivity analysis.

3.2.4.3 Rating Scale Analysis

With the SUCCESS assessment instrument, preceptors indicate the degree of supervision required for each competency skill. This supervision scale rates student performances on three levels (Figure 2-4). The rating scale introduced in this study rates student performances on five domains of behavior. To facilitate comparison between the two scales, The principal researcher created a surrogate scale to approximate the SUCCESS instrument. This approximation was accomplished by collapsing the responses for levels of supervision from the rating scale. This proxy scale will be used to estimate the three supervision currently used in SUCCESS (see appendix O). The Fisher Exact test will quantify differences between the two scales. The study will compare results for the competent vs. not-competent and the excellent vs. entry-level comparisons are shown in Tables P-1, P-2, and P-3 for competency skills A, B, and D/E respectively.

3.2.5 Phase II Methods Summary

The purpose of Phase II was to collect quantitative evidence of validity from preceptors for the SUCCESS competency Drug Therapy Evaluation and Development. Given our goal of establishing a scientifically sound method to measure reliability and

accuracy of APPE assessment instruments, our results should be generalizable to the clinical setting and establish a benchmark for comparison.

Table 3-1. Performance targets from the script

Diabetes Cases	Mary	Thomas	Susan
Skill A	Excellent	Entry-Level	Deficient
Skill B	Excellent	Entry-Level	Deficient
Skill D/E	Excellent	Entry-Level	Deficient
Heart Failure Cases	Patricia	Joseph	Dorothy
Skill A	Entry-Level	Excellent	Deficient
Skill B	Entry-Level	Deficient	Excellent
Skill D/E	Deficient	Excellent	Deficient
Anticoagulation Cases	Linda	David	Barbara
Skill A	Excellent	Entry-Level	Deficient
Skill B	Entry-Level	Excellent	Deficient
Skill D/E	Excellent	Deficient	Excellent

Table 3-2. Video vignettes times

Diabetes	Mary	Thomas	Susan
	10:13	8:19	5:17
Heart Failure	Patricia	Joseph	Dorothy
	5:31	9:05	7:35
Anticoagulation	Linda	David	Barbara
	6:32	7:39	5:10

Table 3-3. Landis & Koch's¹⁵⁰ "Strength of Agreement" table

Strength of Agreement	
<u>ICC</u>	<u>Strength of Agreement</u>
< 0.0	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost Perfect

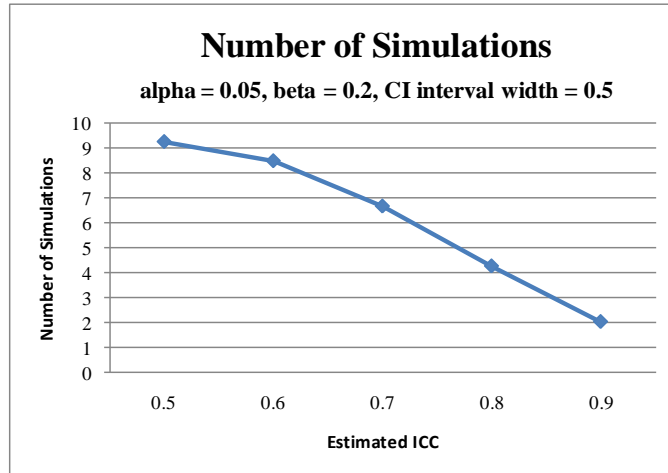


Figure 3-1. Number of video simulations for 40 preceptor raters

- CE Course Objectives**
- Compare and contrast how preceptors assess competency in other health professions.
 - Discuss Accreditation Council for Pharmacy Education (ACPE) experiential education requirements.
 - Describe the factors indicative of pharmacy competency.
 - Explain how to assess students' advanced pharmacy practice experience (APPE) performance.
 - Discuss preceptors' perceptions of the student assessment process for APPEs.

Figure 3-2. CE course objectives

	Excellent and Entry-Level	Deficient
Observed		
Expected		
	Excellent	Entry-Level
Observed		
Expected		

Figure 3-3. 2x2 student assessment tables

CHAPTER 4 RESULTS

4.0 Results Introduction

There are two study phases and each phase addresses different research questions and samples different populations. Phase I recruited a Delphi panel to obtain expert opinion from a cross section of clinical faculty. The Delphi panel evaluated and refined the competency skills and performance criteria. The Delphi panel outcomes address Research Question 1 and provide evidence of content validity. Phase II examined evidence for research questions 2 and 3. With the assistance of an expert panel, the principal researcher developed video simulations illustrating selected competency skills and performance criteria as a pharmacy student interacts with their preceptor. A preceptor panel assessed student performances with the new performance criteria. This data was used to estimate assessment reliability and accuracy.

4.1 Delphi Panel

4.1.1 Round One Panel

Pharmacy schools were selected according to the institution's familiarity with the SUCCESS assessment system. From the 10 contacted schools, the principal researcher received responses from 7 schools to email invitations to their best performing preceptors. Citing resource constraints, two schools requested that the principal researcher contact preceptors directly and provided contact information. The primary researcher made requests for the remaining 5 participating schools to resend the email invitations every two weeks, for three follow-up contacts. Reminder invitations were sent every two weeks for a total of three follow-up contacts. Of the 26 individuals who completed the on-line consent form on the survey's first page, 19 panelists

completed the entire survey and one panelist partially completed the survey. This represents a 73% response rate based on the number of individuals who completed the on-line consent form. The survey responses and demographic data reported will include the usable responses from the 20 participants of the Delphi panel and the two pilot study panelists.

Panel demographics are listed in Table 4-1. Panelists' ages ranged from 29 years to 63 years, a majority (54%) reporting 41 years or older. A slight majority (52%) of the panels were male and almost half (47%) held a doctor of pharmacy (PharmD) degree. Several panelists possessed advanced degrees beyond the PharmBS and PharmD degrees, including a MBA, a MPH, and a PhD. The majority (64%) of the panel reported that they had practiced in pharmacy for 11 years or more. The number of year's precepting students varied. The majority of panelists (73%) indicated that their primary role with students was as a preceptor.

The site demographics are listed in Table 4-2. 73% of panelists were affiliated with public institutions, while the remainder was affiliated with private institutions. Approximately half (53%) worked in hospitals and the remaining worked in various settings. Panels reported a variety of population settings. Panelists represented 8 states and most respondents were from Florida (43%), followed by Pennsylvania with (24%).

4.1.2 Domain Specification Results

The first series of questions address the domain specification of the construct Drug Therapy Evaluation and Development. This line of inquiry asks the Delphi panel to evaluate whether a competency skill is relevant for assessment. The panel reviewed the seven competency skill statements (see list in appendix B) and were asked to respond to these two questions:

- **Domain question:** Is competency Skill X relevant in defining the Drug Therapy Evaluation and Development competency?
- **Follow-up domain question:** Are there any aspects of pharmacy knowledge, attributes, or skills missing in the any of the skill statements?

Skill A – domain question: All 22-panel members indicated that competency Skill A is clinically relevant (Table H-1). Comments suggested incorporating patient input within the process of identifying problems rather than relying solely on data gathered from medical records and other documents. It was suggested that the description be modified by reducing the wording to limit the focus on identifying medication problems. The following is a sample of comments.

1. “This is vital. However, I would like to see it focus on patient MEDICATION history and the student’s ability to perform a thorough medication history by interviewing the patient. This statement could be interpreted that the student is able to just gather information from a chart.”
2. “If there are 5 pre-existing co-morbidities and no meds prior to admission. Or there are the wrong meds. . . this is simple, YES (relevant)”
3. “This statement does not include complete current medications and past medications. Would consider specifically mentioning these as they are important to identifying problems.”
4. “This skill could be expanded. To me, interviewing a patient to get a thorough patient history is much different from using a chart to look up lab data. The third skill then is summarizing into important and relevant to determine the problems. It would be nice to have more areas to assess and not lump so many together where a student can be good at one part so you don’t want to give them a low score but could use work on the other areas of this skill set.”

Skill B – domain question: All panelists (22) indicated Skill B relevant (Table H-2). Comments are as follows.

1. "Yeah. . . see above. If they are in decomp'd CHF, some decision has to be made on a hospital admission as to whether we keep them on the same regimen or change it".
2. "I believe it's very important to know the rationale, particularly when discussing a patient's medication needs with other health care providers.

Skill C – domain question: Nineteen (86%) panelists responded that Skill C was relevant and 14% (3) indicated it was "important but not relevant" (Table H-3).

Comments suggested this description is actually Skill B that is the process of prioritizing drug related problems and this competency skill actually describes a high priority event.

The following is a sample of panel comments.

1. "Relevant, yes, but perhaps impossible to truthfully answer as on most rotations, students don't encounter these events so preceptors end up assuming the student has this ability through discussions. This may not be reflective of the student's competence in a real life situation. Instead of just emergency medical attention, perhaps it would be more relevant to evaluate their ability to recognize emergency and immediate (ie. blood glucose of 1000, potassium of 6.5 opposed to just EMERGENCIES such as respiratory arrest or MI)"
2. "It is already a part of the prioritization found in statement B."
3. "This skill implies a medical diagnosis and not a drug related problem. Not sure that this should be a priority for pharmacy students. If it implies emergency drug related problems I would assume that is incorporated into skill B "prioritizing" problems."

Skill D – domain question: Competency Skill D was clinically relevant for 86% (19) of the panelists and 9% (2) panelists indicated Skill D was "important but not relevant" (Table H-4). Panelists suggested including patient needs, financial implications, and approved guidelines, since these are factors used in evaluating treatment plans. Comments recommended the skill should not be limited to pharmacokinetic data and drug formulation data and include other factors such as

pathophysiology and pharmacodynamics. This competency skill should be written in the form of a therapeutic monitoring plan. The following is a sample of comments.

1. "The "drug formulation" piece of this skill statement needs to be changed. The formulation (i.e. tablet, capsule, suspension, etc) is not relevant to Drug Therapy Evaluation and Development. Instead it would be more relevant if the statement included the phrase ". . . rational drug therapy based on known pathophysiology of disease as well as pharmacodynamics of drugs used to treat the disease(s). . . ."
2. "I would like to see this one reworded but I think that determining an appropriate solution to each of the patient's problems is important. I think more goes into a decision than just PK and drug formulation data. The solution also needs to incorporate patient needs, financial implications, approved guidelines, etc. This should be in the form of a therapeutic monitoring plan."
3. "What about patient data, patient preferences, etc.?"
4. "This statement is too limited in that it only includes pharmacokinetic and pharmaceuticals. There are many more basic and clinical sciences involved in the ability to make good decisions about optimal drug therapy. These include pharmacology, medicinal chemistry, physiology, pathology, and pathophysiology."

Skill E – domain question: Of the 21 panelists who answered this question, 86% (18) indicated competency Skill E as relevant. 10% (2) panelists indicated that Skill E was "important but not relevant" and 5% (1) respondent indicated Skill E was "not relevant" (Table H-5). A number of panelists indicated the need to consider all the factors in competency Skill E and the factors described in the Skill D before a therapeutic plan could be developed. Sample of panel comments follow.

1. "This is actually Skill D"
2. "Students can barely get the primary plan together...."
3. "This should be a part of D. Once the problem is identified, possible solutions are determined and the best solution becomes the recommendation and includes a monitoring plan for that solution. All of

these items need to be determined before the recommendation is determined.”

4. “Add using disease states "and evidence based medicine," This implies importance of rationale.

Skill F – domain question: Only fifteen (68%) panelists indicated this competency skill as relevant and 32% (7) panelists indicated it was “important but not relevant” (Table H-6). Panel members suggested that "back up" plan is not the process that occurs in the clinical setting. When the primary plan does not work, the pharmacist needs to reassess reasons why the plan did not work and use this information to develop an alternative plan. The following is a sample of comments.

1. “Backup plans are rarely performed in real life situations. If a medical situation changes, then an alternate plan is determined. I don't think I've ever included backup plans in a progress note to a physician although by nature, I have alternate plans. They are rarely communicated verbally unless necessary; therefore it is difficult for me to remember to encourage students to do this.”

2. “Again, this should be reworded to be follow-up to the monitoring plan because the monitoring plan will take into consideration what adverse effects or lack of efficacy and if it occurs one starts at the beginning of solving the problem again. I do not think back up plans are necessary.”

3. “Not sure "back up" plan is appropriate term. When primary plans do not work, pharmacist needs to reassess reasons why the plan(s) did not work and then be willing to develop alternative plan(s). This also should include alternative treatments should the recommended therapy not be accepted by the medical team”.

Skill G – domain question: Twenty (91%) panelists reported this competency skill as relevant and 9% (2) panelists indicated it was “important but not relevant” (Table H-7). Panelists recommended that the statement needed to identify the target audience. One panel suggested three audiences; 1) documentation for the patient, 2)

documentation for the chart, and 3) documentation at the pharmacy. The following is a sample of panel comments.

1. "No drug evaluation can be useful if it cannot be communicated in an effective manner regardless of the educational background of the patient."
2. "You do not state to whom it is provided. Also, if documentation does not occur, then payment does not follow. It is necessary and vital to patient care."
3. "This needs to be split into 3 areas - documentation for the patient, documentation for the chart and other HCP and documentation at the pharmacy."

Follow-up domain question: A slight majority 55% (12) indicated there were no missing aspects in the competency skill statements. However, 45% (10) panelists indicated there were missing items (Table H-8). Comments suggested that patient centered factors were missing in the skill statements and several recommended incorporating drug information from the patient and patient's values within the treatment plan. Comments indicated that skill statements lacked factors important with making evidence-based decision clinical decisions. Several panelists suggested skill statements D and F should be combined since in their practice these factors are evaluated together when developing a treatment plan. The following is a sample of panel comments.

1. "The "drug formulation" piece of this skill statement needs to be changed. The formulation (i.e. tablet, capsule, suspension, etc) is not relevant to Drug Therapy Evaluation and Development. Instead it would be more relevant if the statement included the phrase ". . . rational drug therapy based on known pathophysiology of disease as well as pharmacodynamics of drugs used to treat the disease(s). . . ."
2. "Building a rapport with the patient, respect patient's time and values Patient history - review of systems, being thorough, appropriately respond to patient concerns, prioritizing drug therapy problems quality of recommendations, identifying alternatives and then making the recommendation with care, plan including goals of therapy, what requires monitoring and appropriate time frames?. Follow-up with the patient to

monitor the plan and recommend adjustments. Documentation - in a language understandable by the patient, appropriate for other HCP and appropriately uses the systems in place at the rotation site.”

3. “There should be a statement about development of a drug monitoring plan (implied efficacy and toxicity) and communication of that plan to the patient.”

4. “... the statements do not include evidence-based decision making and do not include the aspect of being a member of a health care team.”

4.1.3 Performance Criteria Results

This series of questions asks the Delphi panel to evaluate the performance criteria (e.g. checklist, rating scale, and performance levels). These items associate observed student behavior with a degree of competent performance. The first set of questions relates to the relevance and representativeness of the performance criteria with each competency skill statement. The following and concluding series of questions address descriptions of the individual elements of the rating scale and definitions of the performance levels. Descriptions of the performance criteria are shown in appendix F and G.

- **Relevant question** - Are all the elements of the checklist and the rating scale behaviors relevant for assessing performance in Skill X?
- **Representative question** - Considering the universe of performance measures, are the elements of the checklist items and rating scale behaviors representative of the skills, knowledge, and attitudes required to assess performance in Skill X?

The Delphi panel was given the competency skill statement, checklist, rating scale, and performance descriptions for each competency skill and an explanation of their purpose. The checklist is a description of activities the

student needs to perform to complete an individual competency skill. The rating scale describes student behaviors for a particular performance level (e.g. excellent, competent, and deficient). Student performance are to be rated across five performance dimensions.

- **Supervision:** refers to the degree of assistance.
- **Quality:** refers to the degree of knowledge and skill demonstrated.
- **Complexity:** refers to the number of elements that must be considered.
- **Consistency:** refers to the frequency of desired behaviors.
- **Efficiency:** refers to the ability to perform effectively and timely.

Skill A - relevance question: All panelists (22) indicated the performance criteria relevant for assessing performance of competency Skill A (Table H-9). Panelists commented that data gathered from the patient was important to Skill A and suggested including the data gathered directly from patients would improve the checklist. A panelist commented, maintaining a complete caseload is not indicative of excellent performance. This should be excluded from the rating scale.

1. "Maintaining a complete caseload" may not be needed to achieve an excellent rating. The quality of the skill is more important than the quantity. Student will acquire the ability to do multiple tacks as they gain more knowledge and experience"
2. "I expect the students, particularly 4th year students, to go see the patients and obtain and basic (pertinent) physical exam face to face, they are asked to read the chart both before and after seeing the patient."
3. "The checklist is largely quantitative, but should also consider the severity of disease, as well as the comprehensiveness of the care provided to manage the disease(s)."

4. "Where possible the student should be taking a history from the patient and not relying on what others have in the chart or medical record. Pharmacists can often get more thorough home med lists from patients and may see side effects that other HCP miss when speaking with the patient as we are focused on these issues. Students should be able to interpret lab data and test results so this should remain to be included."
5. "Seems focused on a process where data already gathered. What about the patient interface and gathering data from the patient?"

Skill A - representative question: All 22 panelists indicated that the performance criteria were representative as defined (Table H-10). Panelists' comments suggested a need for greater emphasis on patient interaction. Panelists suggested exclude maintains a full-time caseload, seeks to assist others where needed, and capable of supervising others from rating scale. The following is a sample of panel comments.

1. "Again, there needs to be more emphasis on the one to one nature of the encounter with the patient."
2. "I would expect that the students consider more than the H and P and labs before forming a problem list, I would also like the to look at the home meds, as well as the chief complaint, history of present illness, ht, wt, allergies, (which aren't always in the history or physical), family history, social history including alcohol, drugs, tobacco, caffeine, social setting. regarding the labs, I would also expect the students to take into consideration the exams and studies prior to forming a problem list (CT, MRI, Chest Xray, EKG, etc.) specific edit suggestions: "previous history of illness" is more precisely put "past medical history" "sex" should be "gender" should also include in this section ht/wt/allergies "complain" should be complaint" "synthesizes. . ." should include the other components as discussed above. "expansive view of the profession" this isn't really clear, could be worded differently."
3. "Not sure that these should be included to receive a rating of excellent (or at least should be optional):
 - Capable of maintaining a full-time caseload and seeks to assist others where needed.
 - Capable of supervising others"
4. "Needs to be more patient focused, symptoms the patient is experiencing that may be caused by drug therapy or the lack of drug therapy."

Skill B - relevance question: Seventeen (81%) panelists indicated performance criteria were relevant and 19% (4) indicated a need for revisions (Table H-11).

Panelist's comments suggested the problems listed need to be prioritized on more factors than morbidity and mortality in the checklist. The following is a sample of panel comments.

1. "When organizing problems, it is important to understand the concerns of the patient. A condition that does not appear to be clinically important may be very important to the patient."
2. "I think assessment of home meds should come earlier in the process, but I could certainly work within this framework."
3. "Change 'Organize' to 'Prioritize.'"
4. "Problems need to be prioritized on more details than just morbidity and mortality! Patient priority should also have some bearing on this and team priority may also come into play. Looking at the current medication list also needs to be expanded to look for inappropriate frequencies, inappropriate doses (too high or too low), interactions with food, lab data, etc."

Skill B - representative question: Fourteen (76%) panelists indicated the performance criteria were representative, while 33% (14) of the panelists indicated a need for revisions (Table H-12). Comments suggested that the difference between deficient and competent performance was too great and some panel comments suggested that there was not enough of a difference between competent and excellent performance in the rating scale. Panelists suggested excluding caseload from rating scale. The following is a sample of panel comments.

1. "The difference between competent and deficient is too different, Students can't all carry a full caseload. . . even I have difficulty with this sometimes. This should be reworded to 'the assigned number of patients'. All students are not created equal; some can carry a full load and others cannot: however that one element makes it difficult to choose competent on this question."

2. "What about meds prior to this? What did they come in on?? If the patient isn't doing well, have we asked if there were meds taken prior to admission that could account for this change? Has a 'formulary substitution' been made that has been a therapeutic failure for this patients history?"
3. "I am not convinced that the rating scale category of 'excellent' is relevant in this skill. This category seems too stringent for most students to achieve. The checklist contains items that really only a qualified, licensed practitioner could achieve."
4. "Sometimes the slight variations between competent and excellent are too small to determine. It would be nice if these were a little more defined to more easily see the difference. What is a caseload defined as? It would be nice if there were more guidance - like each student in an ambulatory care rotation had to follow 10 patients in a 2 month time frame so I can know for sure if the case load is appropriate."

Skill C - relevance question: Seventeen (81%) panelists indicated performance criteria are relevant and 19% (4) panelists recommended revisions (Table H-13). Comments suggested consolidating Skill C within Skill B, since this is part of prioritizing patient problems. The following is a sample of comments.

1. "Checklist is disjointed."
2. "While emergency situations may come up in some rotations, I think this falls under the prioritizing the problems of the last skill set and can be eliminated as a separate skill."
3. "May want to reword, In my practice setting, much of the care is provided by certified nurse practitioners in additional to physicians,"
4. "Consider consolidating this with the previous skill."

Skill C - representative question: Fourteen (76%) panelists indicated performance criteria were representative. While 19% (4) of the panelists indicated a need for revisions and 5% (1), panelists indicated the rating scale behaviors were "not

relevant” (Table H-14). Panelists suggested consolidating Skill C within Skill B, since this is prioritizing patient problems. The following is a sample of panel comments.

1. “[I] would like this to address emergency medical needs within the hospital setting (this seems mostly to reflect the community pharmacy or clinic setting.”
2. “Prioritization may need to be pulled from the last skill set so one statement is on defining the problems and one is on prioritizing but I don't believe emergency situations need a separate statement.
3. “Implement any emergency interventions that are needed and a pharmacist is capable of delivering.”

Skill D - relevance question: Thirteen (62%) of the panelist indicated all the performance criteria elements were relevant and 38% (8) panelists indicated a need for revisions (Table H-15). Comments suggested the consolidating Skill D with Skill F, since more factors are considered than pharmacokinetics and formulation. Comments mentioned the checklist does not mention financial implications, patient values, and health literacy. Panelists suggested excluding caseload and supervision from the rating scale. The following is a sample of panel comments.

1. “Excellent: do not think that ‘capable of supervising others’ is relevant to performance in Skill D”
2. “Checklist does not mention financial implications, patient values, education as an option. Also more goes into the decision than PK and formulation - either leave this part out or expand and include all factors in the decision.”
3. “Should the checklist include socio-economic factors including cost, patient's readiness to accept/use therapy/dosage form, health literacy, etc.?”

Skill D - representative question: Twelve (57%) panelists indicated the performance criteria are representative. While 38% (8) of the panelists indicated a need for revisions or indicated the checklist and 5% (1), panelists indicated that the rating

scale behaviors were “not relevant” (Table H-16). Panelists suggested that “State of Health” is too narrowly defined and should be replaced with severity of illness.

Comments suggested that parts of the rating scale were “too stringent” and recommended excluding; 1) maintaining a full-time caseload, 2) seeks to assist others, and 3) managing patients with conditions beyond expected of an entry-level practitioner from the rating scale. The following is a sample of panel comments.

1. “... I would change full caseload to assigned number of patients”
2. ““Severity of illness” should be included. ‘State of Health’ is too narrowly described. “Define the “state of health” of a 20 year old college male who gets a scratch that gets red and is deemed to be community MRSA, and is given a prescription for clinda for a week versus grandma who's 75 and is in septic shock on pressors & CRRT with a VAP. . . on vanco q12 with a trough of 20 for suspected MRSA from her nursing home.”
3. “I think the deficient needs to be redefined. A student coming out of 3 years of a pharmacy program had better be able to design and evaluate treatments or they shouldn't be on rotation. Deficient should be seen as not examining all options or not taking into account patient factors (only looking at guidelines, etc). Competent would be all options included but the best one not chosen and Excellent would be choosing the most appropriate therapy for this patient using all the available information.”
4. “Need to include other disciplines beside pharmacokinetics and pharmaceuticals.”

Skill E - relevance question: Fifteen (71%) panelists indicated the performance criteria were relevant and 29% (6) panelists indicated a need for revisions (Table H-17). Comments suggested the consolidating Skill D with Skill E. Panelists mentioned the checklist does not mention financial implications, patient values, and health literacy. Several suggested excluding caseload and supervision from the rating scale. The following is a sample of panel comments.

1. “Bullet points 3-5 are not relevant to achieving an excellent rating”

2. "I think lifestyle, financial, and safety issues are only some of the issues that should be reflected in the checklist per the following. . . 'evaluates treatment regimens for optimal outcomes using disease states and previous or current drug therapy as well as'

3. "The checklist does not specifically state if there is a better medication for a problem/disease state based on clinical data (ie. study results, guidelines) it only focuses on patient psycho-social and healthcare system factors. In the rating scales I don't feel that being 'capable of supervising others' is relevant to being Excellent in this skill."

4. "This should be included in the last skill statement (D?)"

5. "Based on comments in section D, should sections D and E be merged?"

Skill E - representative question: Thirteen (62%) panelists indicated the performance criteria were representative, while 38% (8) panelists indicated a need for revisions (Table H-18). The following is a sample of panel comments.

1. "Full caseload should be changed to assigned number of patients."

2. "Religion might be considered here too. Jehovah Witness people and the no blood product (or albumin - like in LVP or epo products)."

3. "Again, missing the skills of assessing appropriateness of medications based on disease-state specific criteria (ie. current standard of care, guidelines, clinical data)"

Skill F - relevance question: Sixteen (76%) panelists indicated all the performance criteria are relevant and 24% (5) panelists indicated a need for revisions (Table H-19). Comments suggested the checklist could be improved by having the student identify patient-specific factors that would necessitate a change in drug therapy. Panelists suggested including "checkpoints" in the checklist for monitoring. The following is a sample of panel comments.

1. "Checklist #2 implies this is an ambulatory situation where information is directly shared with a patient, in an inpatient (ie. ICU) rotation statement #2 would not imply and should perhaps read 'patient/healthcare provider.'"

2. "Do not think 'capable of supervising others' is relevant to this skill."
3. "What about developing monitoring plan? The patient and team should be aware of all the clinical parameters and/or lab values to monitor and when and if they are not in the range of what should be done which may include adding a drug, increasing or decreasing a drug, discontinuing a drug, etc. Also education is important as there may be diet or lifestyle activities that can impact the plan. This needs to be made more comprehensive."
4. "This section seems less about 'back up plans' and more about monitoring therapy. Although some 'up front' discussion of medication issues and/or 'checkpoints' need to be discussed as foundation to monitoring, this skill seems more related to developing monitoring plans that can be employed during follow up care."
5. "Consider asking to the student to use current accepted therapeutic guidelines to formulate their 'plan B.'"

Skill F - representative question: Sixteen (76%) panelists indicated that the performance criteria were relevant, while 24% (5) panelists indicated a need for revisions (Table H-20). The following is a sample of panel comments.

1. "Describe patient-specific factors that would necessitate a change in drug therapy"
2. "That checklist is weak. . . . There are so many things that go into a back-up plan. 'Explain to the patient'? Not exactly gonna happen in the ICU. How about the doc?"
3. "This also requires skills and knowledge related to physical and patient assessment. The student must be able to identify appropriate therapeutic outcomes and evaluate them to accomplish this objective."

Skill G - relevance question: Thirteen (62%) panelists indicated that all the elements of the performance criteria were relevant and 38% (8) indicated a need for revisions (Table H-21). Panelist's comments suggested the need to state where the documentation should occur and that the checklist would need major revisions to be appropriate for the hospital setting. The following is a sample of comments.

1. "It isn't clear for whom this skill is intended. If it is for other health-care professionals, then the elements are too involved. If it is for patients, why bother? All of the information on the check-list is readily available in on-line patient information sheets such as those in Clinical Pharmacy."
2. "I never realize this was just referring to the plan for patients (written in LAY language). I have always graded this regarding progress notes to providers. This should be changed to include BOTH. And it needs to be rewritten to be clear on what is actually being graded."
3. "Need to state where the documentation should occur: into EMR, in patient education materials, pharmacy database, etc."
4. "Does not include documentation for team and for pharmacy, utilization of resources at the site. While it may be nice to have written the above details, I am afraid this is probably being accomplished by students printing out drug info pamphlets from Clinical Pharmacology or other software applications. I would like to see this in giving the patient the monitoring plan with written drug information and take out what it needs to include as there are many items missing from the above and others not relevant. I would want the above information covered in a verbal education session with the patient also."

Skill G - representative question: Fourteen (67%) panelists indicated the performance criteria were representative as defined. While 33% (75) panelists indicated a need for revisions (Table H-22). Panelist's comments suggested that the checklist was appropriate for patient communication, however the checklist lacked the essential elements for communicate with physicians and other health care professionals. The following is a sample of panel comments.

1. "Again, would change full caseload to required number of patients."
2. "None of that applies to the ICU - which is probably the highest dollar per patient day in the hospital."
3. "Appropriate information for the receiver of the information whether it be HCP, patient, agent or documentation at the rotation site so other pharmacy students or preceptors can continue care. Also needs to include follow up documentation."

4. “While the elements on this checklist are important for patient communication, they essentially ignore the need to effectively communicate with physicians and other health care professionals. Students have a difficult time with this aspect of practice.”

Complexity question: Sixteen (76%) panelists accepted the definition as written and 24% (5) indicated revisions were needed. No panelists indicated that the definition was not helpful (Table H-23). Comments suggested the description was not clear and made a number of suggestions to improve understanding. Several panelists suggested that examples would illustrate the behavior for assessment. The following is a sample of panel comments.

1. “This one needs clarification especially the part that reads “. . . with fewer elements under direct control by the preceptor.””
2. “the number of elements (simple or complex) doesn't seem to be just right in explaining what I think is meant by this statement.”
3. “What is an example of an increased level of complexity to the environment? May want to give an example.”
4. “It would be nice to define some of these things for guidance. E.g. is 5 comorbid conditions too many for any student? At what point should we be expecting these advances - at quarterly intervals?”
5. “The second sentence is confusing. Is “complexity” a function of how a student derived his or her answers to problems?”

Consistency question: Nineteen (90%) panelists accepted the definition as written and 10% (2) panelists indicated revisions were needed (Table H-24). Following is a sample of panel comments.

1. “is expected to progress from their baseline to routinely.”
2. “Use a common timeline (e.g., weeks or days) as an example of frequency, as many students may obsess over how fast that they need to improve. Words like ‘routinely’ have no instructive meaning for them.”

Deficient performance question: Thirteen (62%) panelists accepted the definition as written and 38% (8) indicated that revisions were needed (Table H-25). Panelists suggested that the performance as described was too narrow and made a number of suggestions to improve its utility. Several panelists suggested that examples would illustrate the behavior for assessment. The following is a sample of comments.

1. “. . . feedback might be changed to something like ‘correction.’ Everyone gets feedback. When the feedback isn't positive. . . there's the problem.”
2. “would change ‘patients’ (sentence one) to ‘tasks’”
3. “It is not clear to me what ‘clinical reasoning is performed in an inefficient manner’ really means. What is an ‘inefficient manner’ as it relates to reasoning? Actions can be inefficient but it's not clear to me how reasoning can be ‘inefficient.’ Wording that comes to my mind includes ‘reasoning below expectations’ or ‘lack of clinical reasoning’ or ‘clinical reasoning that is consistently incorrect.’”
4. “You may wish to include poor attitude and work ethic here.”
5. “Consider adding a patient safety component to this definition”
6. “We replaced ‘Deficient’ with ‘Novice’ at our school”

Efficiency question: Twenty (95%) panelists accepted the definition as written and 5% (1) panelists indicated that revisions were needed (Table H-26) There were no comments from the panel.

Entry-level performance question: Sixteen (76%) panelists accepted the definition as written. While 19% (4) panelists indicated that revisions were needed and 5% (1) panelists indicated, the definition was not helpful (Table H-26). Panelists made a number of suggestions.

1. “I don't think that entry-level performance should include complex patients. In most cases, entry-level would indicate that the student can maintain a caseload but only of simple patients.”

2. "Substitute 'student' with 'pharmacist,' and now you're making \$100k/yr. Not helpful."
3. "Need to define 'Entry-level' or use another term. 'Proficient?'"
4. "I don't think it is appropriate for students at any level to have no clinical supervision. This needs to be reworded to 'A student whose recommendations are thorough, appropriate and in agreement with the preceptor's without coaching' or something similar."

Excellent performance question: Sixteen (76%) panelists accepted the definition as written and 24% (5) panelists indicated revisions were needed (Table H-27).

Panelists made a number of suggestions, in particular suggested the removal of "supervision" and "case load" in the performance criterion. Several panelists suggested the performance criterion was too high. The following is a sample of comments.

1. "need to correct: 'with simple or highly complex' and '. . . is consistently proficient and is capable of. . .'"
2. "Strike, 'Student is capable of supervising others.' They are NOT capable of supervising others - not negotiable, period, end of discussion."
3. "with simple, highly complex" should read "with simple to highly complex" "
4. "I think the bar is set too high for almost all students to achieve. This sentence in particular is the most troubling ('student is able to maintain a full-time caseload and seeks to assist others where needed.')
5. "I don't think it is appropriate for students at any level to have no clinical supervision. This needs to be reworded to 'A student whose recommendations are thorough, appropriate and in agreement with the preceptor's without coaching' or something similar."

Performance criterion question: All panelists accepted the definition as written (Table H-28) There were no panel comments.

Quality question: Seventeen (81%) panelists accepted the definition as written and 10% (2) panelists indicated revisions were needed (Table H-29). The following is a sample of comments.

1. “ yuk - wrong term, wrong place, wrong definition.”
2. “no skill should be an option. No skill, limited skill, appropriate skill, excellent skill or something to this end maybe?”
3. “Consider adding attitude/behavior.”

Supervision question: Twenty (95%) panelists accepted the definition as written and 5% (1) indicated revisions were needed (Table H-30). The following is a sample of panel comments.

1. “It will also vary with the individual students. I'm not a fan of the ‘one size fits all’ theory”
2. “While they may be capable of independent performance, they are not licensed and still require a degree of supervision until they graduate. We are preparing them for the world after graduation but they are not there yet, even in their last rotation, they still have requirements to meet before licensure.”

4.1.4 Round Two Panel

From the 22 panelists in round one, only eight participated in round two. This low number of panelists for the second was well below the recommended sample size discussed in literature. However, the principal researcher deviated from the original plan and took the unorthodox step to recruit additional preceptors directly into round two. The principal researcher contacted 28 pharmacy schools. The study received positive responses from 14 schools to participate in the study and email invitations to their best preceptors. The researcher made requests for schools to resend the email invitations to the selected preceptors every two weeks for a maximum of three follow-up contacts.

Reminder invitations were sent directly to these preceptors every two weeks for a maximum of three follow-up contacts.

Of the 71 individuals who completed the on-line consent form on the survey's first page, 66 panelists completed the entire survey and one panelist partially completed the survey. This represents a 93% response rate based on the number of individuals who completed the on-line consent form. The number of preceptors who were nominated by pharmacy schools to participate and chose not to participate in the Delphi panel is unknown.

Panel demographics are listed in Table 4-3. Panelists' ages ranged from 29 years to 63 years with a slight majority (52%) reporting 40 years or younger. A majority (62%) of the panelists was female and (77%) earned a doctor of pharmacy (PharmD) degree. Several panels possessed advanced degrees including MBAs, MPHs, and PhDs. The majority (54%) of panelists reported that they had practiced pharmacy for 11 or more years. A majority (60%) reported precepting students for 10 or less years. The majority of panelists (56%) indicated that their primary role with students was as a preceptor, followed by 24% who were Educational Coordinators.

Panel site demographics are listed in Table 4-4. 73% percent of panelists were affiliated with public institutions, while the remainder was affiliated with private institutions. A majority (60%) worked in hospitals and the remainder worked in various settings. The majority (77%) of panelists reported working in an urban setting. Panelists represented 18 states, and most were from Texas (20%), followed by Florida (15%).

4.1.5 Domain Specification Results

The first round of the Delphi panel gave strong results in addressing the domain specification of the construct Drug Therapy Evaluation and Development. Except for

competency Skill F, panel agreement was greater than 86% percent. These results meet the threshold of panel endorsement, and domain questions were not carried over to the second round.

4.1.6 Performance Criteria Results

Items that failed to reach panel agreement of at least 80% agreement are presented again with the prior panel rating scores and informative comments. As previously described, this series of questions ask the Delphi panel to evaluate the performance criteria (e.g. checklist, rating scale, and performance levels). These items relate observed student behavior with a degree of competent performance. The first set of questions relates to the relevance and representativeness of the performance criteria with each competency skill. The next and concluding series of questions ask for evaluation of rating scale descriptions, and definitions of the performance levels.

- **Relevant question** - Are all the elements of the checklist and the rating scale behaviors relevant for assessing performance in Skill X?
- **Representative question** - Considering the universe of performance measures, are the elements of the checklist items and rating scale behaviors representative of the skills, knowledge, and attitudes required to assess performance in Skill X?

Descriptions of the performance criteria are shown in Appendices F and G. The Delphi panel was given the checklist and a rating scale for each competency skill. The checklist is a description of activities the student needs to perform to complete an individual skill statement. The rating scale describes student behaviors for a particular performance level (e.g. excellent, competent, and

deficient). Student performances are to be rated across five performance dimensions:

- **Supervision:** the degree of assistance.
- **Quality:** the degree of knowledge and skill demonstrated.
- **Complexity:** the number of elements that must be considered.
- **Consistency:** the frequency of desired behaviors.
- **Efficiency:** the ability to perform effectively and timely.

Skill A – relevant question: Of the 67 panelists in round two, 66% (44) indicated that the performance criteria were relevant. 33% (22) panelists recommended revisions and 2% (1) respondent rejected the checklist items and the rating scale behaviors as “not representative” (Table I-1). There were a number of recommendations; generally, that the rating scale should reflect a greater emphasis on the quality of data collected from the student-patient encounter and recognition of the patient's chief complaint. One panelist suggested modifying the checklist to emphasize the identification of problems based on collecting an accurate and comprehensive patient history and physical exam. The following is a sample of comments.

1. “I can't think of a single student in over 20 years of precepting who would meet all of the criteria for Excellent. If that is the goal, it is probably fine. But, if the goal is to make Excellent attainable, you may want to reconsider some elements (e.g., most new grads couldn't manage a full case load, serve as a resource, AND supervise people).”
2. “I like these revisions because I have had students that can write down everything, but not understand what is important or relevant to the main problems.”
3. “I define competent as being able to complete functions independently which is under your ‘excellent’ descriptor. I would prefer changing the excellent section to competent, the competent descriptor to ‘above average.’ I think this is beyond what the typical ‘student’ can accomplish in

the typical APPE rotations unless the student has the good fortune of having the best rotations and best rotation preceptors. For this example, I think residency is required to meet the 'competent' category."

4. "Under 'Excellent'-- would a student really be expected to or have the ability of supervising others? I think that statement is unnecessary. I realize that it says 'capable' of supervising others, but I think this is asking too much."

5 "I would include social, cultural and economic factors in the checklist for the history. The patient's ability to pay for a medication (i.e. insurance status) may contribute to the problem list and the plan. Additionally, social and cultural factors may need to be considered in a patient's history."

6. Rating Scale: Our experience with our own eval process using a different instrument but similar scale is that preceptors avoid using the rating of 'Deficient.' We ended up with really just a 2-part rating scale (Developing and Proficient). We changed our wording to 'Novice' so that it would not sound so derogatory as 'Deficient.'"

Skill B - representative question: Fifty-three (79%) panelists indicated performance criteria as representative. While 19% (13) panelists recommended revisions and 2%, (1) respondent rejected the performance criteria (Table I-2). Panelists comments suggested a number of items to add drug-related problems to the checklist, including dosage regimen, drug interactions, adherence, and compliance issues. They recommend that the checklist to include a medication taken by the pharmacy student. The following is a sample of panel comments.

1. "I believe there should be something regarding medication reconciliation."

2. "Other than not liking the excellent and competent descriptors, these are good expectations for APPE students. Unfortunately, many of the APPE rotations at our local pharmacy schools do not emphasize these skills enough. Congrats if your rotations do so!"

3. "Can something be included about medications taken prior to admission? With as big as medication reconciliation is right now, some significant drug related problems actually occurred prior to admission."

The student being able to identify that so the resumption of home meds doesn't perpetuate the problem.”

4. “I would accept as written, but would consider adding the ability to perform a medication history if not evaluated elsewhere as it is impossible to identify all potential medication problems without finding out from the patient or caretaker what they are REALLY taking.”

Skill D – relevant question: Fifteen (23%) panelists indicated that the performance criteria were relevant, while 55% (36) panelists recommended merging Skills D E into a single skill. Within the combined D and E skill statement, 20% (13) panelists endorsed including severity of illness into the checklist. Three percent (3%) recommended including pharmacodynamic data into the skill statement (Table I-3).

Collapsing the responses into two outcomes, 1) panelists accepted the performance criteria as written, and 2) panelists recommend merging Skill D and Skill E into a single skill for assessment (Table I-4). Of the 66 panelists who answered question, 23% (15) accepted the checklist items and rating scale as written. Seventy-seven percent (51) panelists recommended merging Skill D and Skill E into a single competency skill. Panelists made a number of suggestions to improve the checklist. Comments suggested changes to preceptor guidance and caseload in the rating scale.

1. How about considering therapeutic alternatives (i.e. the person should be able to identify all feasible pharmacotherapeutic alternatives available for achieving the therapeutic outcomes)? This should include both pharmacologic and non-pharmacologic therapies. It seems that Skill set E should include a combination of drug-allergy, drug-drug, drug-disease, interactions, patient age, organ impairment (renal/hepatic), adverse effects, poor compliance, pregnancy, and cost.”

2. “Sorry, can not get past your definition of competent to include the need for preceptor's guidance. Competence to me means able to complete independently. I agree with adding both severity of illness and pharmacodynamic data. The descriptions for this section are pretty vague compared to the previous skills and thus open to interpretation. I do not agree with combining this one with Skill E”.

3. "In the Checklist, the term "possible therapeutic options" should be clarified to suggest that alternative plans need to be considered when developing a treatment plan. For example, with the revised skill statement, what if a patient cannot afford the medication proposed in the initial plan?"

4. "Severity of Illness will assist with the entire clinical picture of the patient."

Skill D – representative question: Thirty-one (47%) panelists indicated the performance criteria as representative, while 52% (34) panelists recommended removing the last three items in the Excellent rating scale. Two percent (1) panelist rejected the checklist items and the rating scale behaviors as "not representative" (Table I-5). Comments suggested that the rating scale for "Excellent" performance is too narrowly defined and would be applicable to a small minority of students. Some panelists suggested the "Excellent" rating scale was applicable to pharmacy residents rather than pharmacy students. The following is a sample of panel comments.

1. "I don't agree with the "beyond expected of entry-level practitioner" for any of the excellent ratings for any of the skills. Very few students (less than 5%) graduate with skills at this level and most require time on the job or a residency to achieve ratings as you describe under the 'excellent' rating."

2. "Let's face it: economics play a part in therapy! We must be aware of cost associated with care and introduce that information somewhere."

3. "I agree only with the suggestions to remove the full time case load part, and beyond entry-level part. I think you should leave the supervise others"

6. "If you were dealing with pharmacy residents I believe those last three items are relevant but for a student it may be too much. You may have some dynamic students but probably not as a standard for evaluating all students."

Skill E – relevance question: Twelve (18%) panelists indicated the performance criteria were relevant, while 80% (53) panelists recommended merging Skills D and E

into a single skill. One (2%) panelist rejected the performance criteria (Table I-6). Comments suggested adding cultural competence and patient preference with the combined skill statement. Panelists suggested the rating scale for “Excellent” performance is too narrowly defined and would be applicable to a small minority of students. The following is a sample of comments.

1. “As a criteria for excellent, I would also the reviewer is exploring any non pharmacological measures that may be used or added to improve therapeutic response.”
2. “Unless a student has had several excellent APPE rotations, very few get enough of these experiences to achieve any level of competence at this. This is for residency or time (years) on the right job.”
3. “See previous comments regarding criterion in ‘Excellent’ and ‘Deficient’ Text of previous comment, “I don’t agree with the ‘beyond expected of entry-level practitioner’ for any of the excellent ratings for any of the skills. Very few students (less than 5%) graduate with skills at this level and most require time on the job or a residency to achieve ratings as you describe under the ‘excellent’ rating.”
4. “Psycho-social does not inherently include the concept of cultural competence, which should be considered for inclusion into this skill. Treatment regimens need to factor in for patient beliefs and expectations, their own healing traditions, and other cultural factors.”

Skill E - representative question: Thirty-one (47%) indicated the performance criteria were representative, while 50% (33) panelists recommended removing the last three items in the Excellent rating scale (i.e. full-time caseload, supervising others, beyond expected of entry-level practitioner). 3% (2) panelists rejected the checklist items and the rating scale behaviors as “not representative” (Table I-7). Panelists’ comments provided rationale for retaining or modifying the rating scale.

1. “I don’t think many students reach this level in today’s APPE experiences. Keeping this expectation is setting many up for failure.”

2. "I agree only with the suggestions to remove the full time case load part, and beyond entry-level part. I think you should leave the 'supervise others.'"
3. "Remove items 2 through 6, as they are now repetitive across this set of skills, and most have no linkage to the skill statement or the checklist!"
4. "Why remove the last three items? The student doing it for one patient in a day is no success. Them doing it for 15-20 in a day (i.e., full caseload) is excellent. . . period.'

Skill G – relevant question: Twenty-two (36%) panelists indicated the performance criteria were relevant and 64% (42) panelists recommended revisions (Table I-8). A number of panelists' comments suggested the need for students to learn how to create written documentation for patients and healthcare professionals. The following is a sample of comments.

1. "One thing I struggle with when working with students is when to use appropriate terminology (lay terms versus medical terminology) with patients and health care providers. I think they should be able to communicate with each group in a way that the patients and other health care providers will think they are knowledgeable, educated, and professional."
2. "If we teach them HOW to document, and then allow them TO document, and then at the end of the rotations we assess what they actually did with our teaching, then we're getting somewhere. The multidisciplinary approach isn't additional, it's foundational. I think adding "other healthcare professionals" should come right after adding that they should use a pencil or pen. . . basic precepting 101."
3. "I definitely feel that writing information for the patient and writing a true pharmacy care plan are two different things. We want the care plan to reflect the full drug therapy assessment completed by the student and to include all medical conditions being treated with medications. This is completely different from writing a patient information page for a patient, and I would not call that a care plan."

Skill G - representative question: Thirty-three (50%) indicated the performance criteria were representative, while 48% (32) panelists recommended revisions. Two

percent (1) panelist rejected the checklist items and the rating scale behaviors as “not representative” (Table I-9). Comments provided rationale for retaining or modifying the checklist and rating scale to include documentation for healthcare professionals. The following is a sample of panel comments.

1. “Also, again, excellent and competent have similar definitions”
2. “I agree it is critical for students to learn to write a note appropriate for inclusion in the medical record, i.e., communicates the plan to other healthcare professionals. However, if skill E is not achievable for students, I'm not sure this one will be either.”
3. “What is this removing thing? Should they go out into practice and tell a boss that they can do great. . . but only in one or two patients??? No, the excellent do it in their full caseload.”
4. “I note the recurrent theme of removal of full-time caseload/supervisors others/beyond expected entry-level practitioner and fully agree. I would never expect a student, or even a new graduate to carry a full-time case load, nor would I expect them to be able to supervise others. I do not allow my residents to supervise students until I am comfortable with their abilities. I concur that all 3 of these items would be more appropriate in the evaluation of a resident, but not a student.”

Complexity question: Forty (61%) panelists accepted the definition as written, while 38% (25) panels recommend examples. One panelist (2%) indicated the definition was not helpful (Table I-10). One panelist suggested that the degree of complexity should be assessed for the duration of the individual rotation rather than the course of all the clerkship rotations. The following is a sample of comments.

1. “Examples are usually constructive, as long as they are not assumed to be templates that must be followed.”
2. “Is it really possible to increase the level of complexity of patients/tasks/environment as a student moves through the APPEs? Our system is set up with a somewhat random match so it is possible a student may have a BMT rotation first and a community pharmacy rotation last - this would not necessarily allow for increased complexity as defined above. Instead, perhaps complexity should be over the course of a given

rotation - the student moves from simple to complex for each rotation with the goal of mastering increasing complexity over time.”

3. “I think examples provided will assist in making sure all preceptors are evaluating in a consistent manner.”

Deficient performance question: Twenty-two (33%) panelists accepted the definition as written and 65% (43) panelists recommended revisions to the description. and should include “ A student who requires close clinical supervision with constant monitoring” in the description (Table I-11).

1. “Deficient performance should also include elements of professional behavior not meeting standards.”

2. “. . . , I would strongly recommend getting rid of or modifying the statement about ‘performance reflects little or no experience’ to something about inability to improve over time and/or with coaching/feedback.”

3. “I like these revisions better than what is written but I would propose to keep the first sentence from the original (A student who requires close clinical supervision with constant monitoring and feedback, even with simple patients) then the last three sentences from the revision.”

4. “As stated earlier, I would change the word ‘Deficient’ to ‘Novice’ to avoid the negative connotations.”

5. “Recommended revisions are excellent. These statements are very well written and accurately describe what is deficient performance. Well worded and easy to comprehend.”

Excellent performance question: Twenty-four (36%) panelists accepted the definition as written and 64% (42) panelists recommended changes to performance description (Table I-12). The following is a sample of comments.

1. “I agree with the revisions as being more realistic of what a student can accomplish. I think too many of your excellent performance criteria are those of a resident, not a student.”

2. “We should be careful when using the term ‘no clinical supervision.’ I would be in favor of the revision.”

4.1.7 Summary of Delphi Panel Results

The first line of questions addressed domain specification. This is to gather expert agreement whether a particular competency skill is relevant in assessing Drug Therapy and Development competency. This first Delphi panel round gave strong results and agreement was greater than 86% for all competency skills, except for competency Skill F. The Delphi panel commented that the backup plan described in Skill F does not play a relevant role in the clinical setting.

The panel discussed how the activities described for skills D and E are intermixed and recommended combining these competency skills into a single skill for assessment. Panelists' comments described how activities, which consider optimal outcomes and drug performance characteristics (i.e. Skill D), and activities which consider patient preferences and social-economic factors (i.e. Skill E) tend to perform jointly.

Consistent with a clinical perspective of performance assessment, panel members inquired when and how the student would incorporate data obtained directly from the patient. That is, how will the student integrate patient-obtained data? Synthesizing medical information as described in Skill A is incomplete without including the role of patient-acquired data.

There was strong objection to some elements of the performance level descriptions. Using "maintain a full-time caseload" as an indicator of performance level was widely rejected. Similar objections were made about including "expansive view of the profession", and "capable of supervising others" in the performance level definitions. Panel members concluded that these elements were not relevant to a student's clinical duties in APPE rotations.

There was a strong theme across the Delphi panel of whether an item under evaluation was relevant in the clinical setting. This viewpoint reflects the unique perspective of clinical faculty in which performance assessment is intertwined with relevant clinical activity. Many panelists felt descriptions of many of the performance criteria were written in broad terms and that examples are needed. In other words, these training materials need to show the performance criteria in the context of the clinical setting. Video vignettes and case studies would help train preceptors in performance assessment.

4.2 Phase II Video Simulation

4.2.1 Expert Panel

An expert panel of preceptors was recruited from the Phase I Delphi panel. Out of the nine preceptors who agreed to join the expert panel, only seven completed all the tasks. Demographics of the expert panel are listed in Table 4-5. The participant ages ranged from 30 to 66 years with a majority (57%) reporting 40 years or younger. A majority (71%) of the participants was female and all (100%) held a doctor of pharmacy (PharmD) degree. All participants possessed advanced degrees or BPS certifications. The majority (86%) of participants reported that they had practiced 10 years or fewer. A majority (86%) reported precepting students for 10 years or fewer. The majority of participants (86%) indicated that their primary role with students was as a preceptor, followed by one participant (14%) who reported Clinical Coordinator as their primary role with students.

The site demographics are listed in Table 4-6. All the participants (100%) were affiliated with public institutions. The largest group (71%) worked in hospitals and the remainder worked in various settings, including Clinical Decision Support services,

representing organizations covering 2.5 million beneficiaries. The majority (86%) of participants reported working in an urban setting. Participants represented five states, and most respondents were from Alabama (29%) and Texas (29%).

4.2.2 Expert Panel Results

Using an internet based WEB site¹³⁸, members of the expert panel watched the video vignettes and evaluated the performances using the new rating scale. These evaluations took place two to three months after the last round of script edits. Based on these rating scale scores, performance levels were derived using a previously described rubric (see Appendix G).

4.2.2.1 Diabetes Case Study Results

The target performance levels for the Diabetes cases are presented Table K-1. Student pharmacist Mary in video Vignette #1 was intended to illustrate excellent performance for all three skills. Meanwhile, student pharmacist Thomas in video Vignette #2 the goal was to illustrate entry-level performance for all three skills. In video Vignette #3, student pharmacist Susan was scripted to illustrate deficient performance for all three skills. The assessments by the expert panel are in Table K-2. The bookends of targeted performances levels matched the expert panel rating scores. However, there were mixed assessments for Thomas' vignettes illustrating entry-level performance.

One panelist commented that the videos effectively illustrated the difference between excellent and entry-level performances, as compared to simply reading the scripts. Mary's (the excellent student's) performance was "... better than pharmacy residents and well above pharmacy students...". Another panelist commented how Thomas (the entry-level student) and Susan (the deficient student) performed "well

above average” compared to their pharmacy students. The following is a sample of participant comments.

1. [Mary vignette #1] “I rarely/never have students perform at this level. It is challenging to get residents to perform at this level. I sure would love it if students could perform this well, and residents too!”
2. [Thomas vignette #2] “I think you made the difference between excellent and entry clearer in the videos compared to the scripts. I remember that the entry students knew almost every answer. Having them not know some things is more entry level.”
3. [Thomas vignette #2] “Still well above average for students.”
4. [Susan vignette #3] “This is still better than about half the students we get, and I consider this level of performance to be lower than acceptable.”
5. [Susan vignette #3] “As with skill B, I think this was supposed to be ‘novice’ but in this area I felt like the student was not as good as Thomas in Video #2 - maybe he should have been more “excellent” in my ratings. I thought the student here just continued to demonstrate a lack of confidence and some lapses in judgment/information.”

4.2.2.2 Heart Failure Case Study

In the Diabetes cases, the target performance level for each skill was uniform across the students’ video vignettes. All the skills were excellent, entry-level, or deficient for each competency skill portrayed by the student actor. However, the Heart Failure vignettes were randomized and the target performance levels for the Heart Failure cases vary for each skill illustrated by a given student. The panel was blinded to this randomization. The expert panel assessments are in Table K-4 and show mixed results among all the Heart Failure video vignettes.

One panelist commented that they were able to recognize differences in performance levels between the three competency skills illustrated within a given student vignette. However, there were several comments about how panelists struggled

to make fair assessments based on a single case presentation. One expert panelist acknowledged how previous student performances influenced their rating scores of following students. The following is a sample of participant comments.

1. [Patricia #4] “Inconsistent level of understanding, she often asks for help, which indicates to me she is aware of deficiencies and is trying to correct them.”
2. [Patricia #4] “I had a harder time distinguishing this [Skill A], It seemed to be on the entry/novice line.”
3. [Patricia #4] “I think efficiency is difficult to judge sometimes from these limited interactions (one case presentation, however, preceptor should be able to use overall experiences to rate that too).”
4. [Joseph #5] “I think this was done very well. . . overall the only reason I hesitated with this was sometimes the student provided too much information which seemed to make the message of what he was saying be slower so I was struggling with whether that was ‘entry level’ instead of excellent.”
5. [Joseph #5] “The change in performance/knowledge between skills was recognizable.”
6. [Dorothy #6] “This student did not seem as proficient in this section. Again, as with the first section, I think it was in the way the student delivered the information that was made it seem entry level or even novice; however, I feel I chose entry level because the skills here were blurred some (in my mind) vs. the other sections so I gave the benefit of the doubt - this is what I actually do on rotations.”

4.2.2.3 Anticoagulation case study

Like the Heart Failure case study, the Anticoagulation case study vignettes were randomized and the target performance levels for the three competency skills vary. The performance targets are presented in Table K-5. The assessments by the expert panel are presented in Table K-6 and again show mixed results. As with the Heart Failure cases, there are several comments relating to the struggle experienced in making the

assessments. One panelist acknowledged referencing a copy of the scripts while making the rating scores. A sample of panelist's comments follows.

1. [Linda # 7] "Although not inefficient in this category, it was not as efficient as the presentation of the pt initially.[using the performance of Skill A to influence the assessment of Skills B and D/E]"
2. [Linda # 7] "I felt this was in between entry level and excellent - I felt at times it was "excellent" but maybe I leaned toward entry because of the performance in the previous skill."
3. [Barbara #9] "I'm not sure this was completely inefficient but maybe not as focused or fluid as it should have been so I thought it on the verge of entry/novice."

4.2.3 Summary Expert Panel Results

There were twelve out of twenty-seven competency skills illustrated in the video vignettes with unanimous agreement or had a single dissenting panelist. These twelve competency skills matched the intended performance level from the script. The following is a list by competency skill:

- Skill A: Mary (#1), Susan (#3), Dorothy (#6), Linda (#7), Barbara (#9)
- Skill B: Mary (#1), Susan (#3), Barbara (#9)
- Skill D/E: Mary (#1), Susan (#3), Patricia (#4), David (#8)

The assessments for all three-competency skills for Mary were unanimous. The assessments for Susan in the Diabetes cases were also unanimous. Mary's performance was excellent, while Susan's was deficient. Dorothy and Barbara's performance for competency Skill A was deficient. Linda's performance in the same competency skill was excellent. Barbara's performance for competency Skill B was deficient. Patricia and Barbara's performance for competency Skill D/E were both deficient.

Several factors may have contributed to different assessments within the expert panel and from the performance levels, the videos intended to portray. There were several comments about how panelists struggled to make fair assessments based on a single case presentation. These panelists were intimately involved with editing the scripts and devising the behaviors that represent each element in the rating scale. However, there were several months between the last round of script edits and assessment of the video performances. This time span would have diminished an intimate understanding of the performance criteria and rating scale. The video vignettes were long and played continuously, taking over 70 minutes to view. Panel members were able to view the videos at their convenience. Nevertheless, there were 27 individual competency skills for evaluation and this is a large number of video vignettes to watch and assess for a group of busy professionals.

Some panelists mentioned that they experienced difficulty recognizing when one competency skill ended and the next competency skill started. Other panel members commented how they were readily able to observe different levels of performances between competency skills. The principal researcher did not perform a debriefing to collect detailed information linking specific student behaviors illustrated in the video with rating scale items or recognition of individual competency skills.

There were several comments suggesting that performances of previous skills within the same vignette influenced rating scores. Each competency skill presented should have been scored based on its own performance criteria. In fact, one panelist used a copy of the script to aid in scoring the performances. However, the panel was blinded to the randomization and any panelist who used their copy of the script would

have given inaccurate rating scores. The rating scores for this panelist were included in analysis since it is not known how many other panelists used their copy of the scripts while scoring the video vignettes.

4.2.4 Preceptor Panel

Pharmacy preceptors participated in the study through an internet portal hosted by the Texas Pharmacy Association (TPA). The Accreditation Council of Pharmacy Education (ACPE) accredits TPA as a provider of continuing pharmacy education and this three hour program was accredited for 3.0 Contact Hours (0.3 CEUs). The CE course started with a 20-minute introduction to assessment concepts and the new performance criteria. Participants then watched nine video vignettes and scored the performances using the new performance criteria.

Participant demographics are listed Table 4-7. Forty-two participants completed the program out of the seventy-four who registered. Only data collected from participants who scored all the video vignettes were included for analysis. Participant ages ranged from 26 to 61 years with a majority (62%) reporting 40 years or younger. A majority (67%) of the participants was female and (71%) held a doctor of pharmacy (PharmD) degree. Several participants possessed advanced degrees and certificates including MSs and PhDs. The majority (53%) of participants reported that they had been practicing pharmacy for 10 years or fewer. A majority (84%) reported experience precepting students for no more than 10 years. The majority of participants (76%) indicated that their primary role with students was as a preceptor, followed by 5% who were Educational Coordinators.

Preceptor site demographics are listed in Table 4-8. Sixty nine percent of participants were affiliated with public institutions, while remaining participants were

affiliated with private institutions. The largest group (41%) practiced in hospitals and the remaining practiced in various settings. The majority (67%) of participants reported their practice resided in an urban setting. Participants represented five states and the majority of participants practiced in Florida (81%), followed by Texas (12%).

4.2.5 Preceptor Panel Results

Figure F-2 illustrates the data collection and analysis pathway previously described. Participants scored student video performances based on the new performance criteria. Using a previously described assessment rubric in appendix G, shows how student performance levels are calculated and assessment results are shown in Appendix L. Reliability and accuracy estimates comparing competent vs. non-competent assessments and competent vs. excellent assessments are evaluated. A comparison between the rating scale and a proxy for the supervision scale is evaluated.

4.2.5.1 Estimate of Reliability

The sample size calculation used in this study assumed ICC estimates would be equal to or greater than 0.5. Unfortunately, this assumption was not realized and the researcher was unable to estimate reliability of each of the three competency skills for each student video vignette as originally planned. However, reliability estimates of the three competency skills by case study offers an opportunity to observe important ICC trends.

The ICC estimates the degree of agreement among preceptor assessments. Higher ICC values reflect greater inter-rater reliability among preceptor assessments. The Landis & Koch¹⁵⁰ “Strength of Agreement” categories will be used to describe ICC results. However, these divisions are arbitrary and are used only for the purpose of discussion.

Table M-1 shows the competent vs. not-competent reliability estimates for each competency skill and is broken down by case study. The ICC point estimates range from 0.37, 0.31, and 0.30 for each competency skills A, B, and D/E respectively. According to previously described guidelines¹⁴⁹, these values indicate fair inter-rater reliability. Each point estimate has similarly wide 95% confidence intervals (CI) and ranges from 0.15, representing slight reliability, to 0.69, representing substantial reliability.

The ICC point estimates and 95% CIs of the excellent vs. entry-level comparison (Table M-2) are generally lower. The 0.24, 0.19, and 0.19 ICC point estimates indicated fair to slight inter-rater reliability for skills A, B, and D/E respectively. The 95% confidence intervals are wide and range from a low 0.09, indicating slight reliability, to 0.55, representing moderate reliability.

4.2.5.2 Influence of Randomization

The video vignettes were presented in the same order shown in the tables. Diabetes cases were first, followed by Heart Failure, and concluded with the anticoagulation cases. Although there may be some weariness due to the three-hour long program, the inter-rater reliability trend is inconsistent with participant fatigue. There were noticeable differences depending on the case study. The Diabetes cases have the highest inter-rater reliability estimates and slimmest (95%) CI ranges for all three-competency skills. The lowest reliability estimates were from the Heart Failure cases and some ICC results were not statistically significant. The ICC results increased slightly with the Anticoagulation cases. This trend is consistent for the competent vs. not-competent and the excellent vs. entry-level comparisons. The CI ranges are wide and none of the ICC results is significantly different from each other.

The results suggest preceptor panelists were able to assess performances in the non-random video vignettes (Diabetes) with a greater degree of reliability compared to the randomized vignettes (Heart Failure and Anticoagulation).

4.2.5.3 Rating Scale Items

The rating scale scores student performances on five domains of behavior (see Appendix G). Specific scores are associated with different levels of performance (e.g. excellent, entry-level, and deficient) by an assessment rubric. Tables M-3, M-4, and M-5 give the ICC values for each rating scale item for Skill A, Skill B, and Skill D/E respectively. The rating scale items for the Diabetes cases have the highest inter-rater reliability estimates and the slimmest (95%) CI ranges. The lowest reliability estimates were from the Heart Failure cases and some results were not statistically significant. The rating score reliability results increased for the Anticoagulation cases. In all cases, the 95% CI ranges were wide and none of the rating scale items was significantly different from each other. No particular rating scale item showed observable trends that crossed over the three competency skills.

4.2.5.4 Global Assessment

The principal researcher asked the preceptor panel to give their global impression. Specifically the level of performance (e.g. excellent, entry-level, and deficient) for each of the 27 competency skills they watched. Table M-6 shows the global assessment inter-rater reliability estimates comparing the competent vs. not-competent assessment. The inter-rater reliability was higher for the non-random Diabetes performances compared to the randomized cases. The previous reliability estimates also show this general trend. This similarity suggests that the rating scale, to some degree, reflects the preceptor's global assessment of the student's performance.

4.2.5.5 Estimating Accuracy

The Fisher Exact test is used to estimate accuracy and is used to compare the preceptor assessment with the expert panel.¹⁵¹ Low p-values indicate that the preceptor panel assessed the performances differently than the expert panel. The principal researcher selected p-values equal to or less than 0.10 as the significant threshold. Given that the 0.10 value is arbitrary, the study will note p-values up to the 0.30 level. Accuracy results are shown in Tables N-1, N-2, and N-3 for competency Skills A, B, and D/E respectively.

4.2.5.6 Competency

The expert and the preceptor panels appear to assess competency differently. Figures N-1, NO-2, and NO-3 show the expert panel rating scores resulted in tougher assessments compared to the preceptor panel. Forty-nine (49%) percent in Skill A, forty-eight (48%) percent in Skill B, and fifty-two (52%) percent in Skill D/E were assessed deficient by the expert panel. In contrast, more student performances were found competent by the preceptor panel. Rating scores from the preceptor panel resulted in, twenty-two (22%) percent in Skill A, twenty-nine (29%) percent in Skill B and thirty (30%) percent were assessed not-competent. The Fisher Exact test results for the competent vs. not-competent comparisons are shown in the middle column of the tables of Tables N-1, N-2, and N-3.

For competency skills A, B, and D/E, the p-values for Thomas range from 0.06 to 0.15. All seven (100%) members of the expert panel considered Thomas' performance competent for all three skills. This is in contrast to the mixed assessments from the preceptor panel. There were negative comments about the Metformin dose recommended in contradiction with the expert panel, which approved the script.

For competency in Skill D/E, the low p-value (0.03) for Patricia gives strong evidence that the preceptor panel's assessments are different from the expert panel. All seven (100%) members of the expert panel considered Patricia's performance deficient. On the other hand, only twenty-four (57%) preceptors out of forty-two gave the same negative assessment.

Unlike the reliability results, no trends across case studies were observed; nor were any trends across the three skills. Summary comments on the competency accuracy results are the same for the excellent accuracy results and are found at the end of the following section.

4.2.5.7 Excellence

Figures N-1, N-2, and N-3 show little overall difference between the expert and preceptor panel's ability to discriminate between excellent and entry-level performances. Thirty-one (31%) compared to thirty (30%) percent were assessed as excellent in Skill A, for the expert and preceptor panels respectively. For competency Skill B, twenty-six (26%) and twenty-seven (27%) percent were assessed as excellent, for the expert and preceptor panels respectively. For competency Skill D/E, twenty-three (23%) were assessed excellent A for both panels. The Fisher Exact test results for the excellent vs. entry-level comparison results are shown in the far left column of Tables N-1, N-2, and N-3.

For competency Skills A, B, and D/E, the p-values for Mary ranged from 0.14 to 0.18. All seven (100%) members of the expert panel considered Mary's performance excellent. This is in contrast to the mixed assessments from the preceptor panel. One preceptor related that the student's performance was excellent, however disagreed with

the exception of the drug therapy recommendation. Strongly suggesting differing assessments may be related to disagreements on technical aspects of the drug therapy.

For competency Skill A, Linda's p-value (0.10) gives evidence that the preceptor panel's assessments are different from the expert panel by chance of sampling alone. Six (86%) out of seven expert preceptors, determined Linda's performance was excellent. In contrast, only 25 (60%) preceptors gave the same outstanding assessment.

The low p-value (0.07) for Dorothy gives evidence that the preceptor panel's assessments are different from the expert panel for competency Skill B. The expert panel had difficulty assessing Dorothy's performance. One (14%) expert determined Dorothy's performance was excellent, one (14%) assessed entry-level, and the remaining four (57%) considered the performance deficient. Linda and Dorothy's assessments show a wide variation in the expert and preceptor panel assessments

For both the competency and excellent comparisons, many competency skills show a wide variation in the expert and preceptor panel assessments. This variation suggests the preceptor panel experienced the same difficulty assessing the performance in the video vignette as the expert panel. In some cases, the video vignettes may not effectively portray the intended performance level and limits interpreting results. On the other hand, some assessment variation appears related to different standards of medical care among the preceptors and with the expert panel. These comments present strong evidence of rater inconsistency.^{152, 153}

4.2.5.8 Rating Scale Comparison

The rating scale scores student performances on five domains of behavior. The SUCCESS instrument asked the preceptor to score the level of supervision required for

each competency skill. The principal researcher created a surrogate scale to approximate the SUCCESS instrument by collapsing the responses for levels of supervision from the rating scale. This proxy scale will be used to estimate the three supervision scale items currently used in SUCCESS. The Fisher Exact test will quantify differences between the two scales. Results for the competent vs. not-competent and the excellent vs. entry-level comparisons are shown in Tables O-1, O-2, and O -3 for competency skills A, B, and D/E respectively.

The Fisher Exact test shows that most of the competent vs. not-competent comparisons are significantly different. The two scales appear to assess competency differently. Figures O-1, O-2, and O-3 show the proxy scale was more lenient than the new rating scale. Seventy-eight (78%) percent in Skill A, sixty-six (66%) percent in Skill B and seventy-one (71%) percent were assessed competent with the supervision proxy scale. On the other hand, fewer student performances were found competent according to the new rating scale. With the new rating scale, fifty-six (56%) percent in Skill A, fifty-three (53%) percent in Skill B and fifty-six (56%) percent in Skill D/E were assessed as competent.

Most of the excellent vs. entry-level comparisons are not significantly different. The assessment results in figures O-1, O-2, and O-3 show little differences in the percentage of excellent assessments. However, there the Fisher Exact test and the graphs show that entry-level and deficient assessments are different between the two scales.

4.2.5.9 Preceptor Comments

Although the Diabetes vignettes may have more effectively illustrated student performance compared to the remaining randomized cases, there are other factors to

consider. In the Diabetes cases, the target performance levels were purposely selected. In each Diabetes vignette, all the competency skills were presented at the same level (e.g., excellent, entry-level, or deficient). On the other hand, the Heart Failure and Anticoagulation vignettes were randomized and preceptors could observe a student with an excellent performance for competency Skill A followed by a deficient performance for competency Skill B. Participants were verbally notified whether they were about to watch a video vignette with randomized performance levels. However, if participants were overwhelmed with a great deal of information presented in the webinar or simply misunderstood the directions, this could have lead to assessment errors.

The expert panel previously mentioned difficulty identifying when an activity for one competency skill was completed and an activity for the next competency skill started in a video. Although this is a real-life challenge faced by all preceptors, participants in this study may have nevertheless experienced similar difficulty untangling student behavior associated with a specific competency skill. To help mitigate this issue, the first line of dialog for each skill was given to the preceptor panel. Difficulty recognizing behavior associated with any specific competency skill would increase assessment variation. These errors would affect assessment results with the randomized cases to a greater degree than the non-randomized cases.

There were several comments suggesting that performances of previous skills within the same vignette influenced the process of rating the student. Comments mentioned like, “giving a break” and “[student name] was doing so well” in the previous performance, strongly suggest presence of the halo effect.^{153, 154} The halo effect has been described when a rater’s perception of one performance influences the

assessment on other performances, resulting in a failure to discriminate among different competency skills.^{153, 155, 156}

Rater inconsistency occurs when preceptor's application of the performance criteria is inconsistent with other preceptors.^{152, 153} There were contradictory comments between preceptors on student performance within the same competency skill. One preceptor chastised the student for "poor medication decisions", while another preceptor wrote for the same student "Made confident recommendations and follow-up plan." Another example of contradictory observations resulting in differing assessments can be seen with this pair of comments: "...didn't have all answers - seemed unsure of some things," while another preceptor gave different rating scores pointed out that the student was "Willing to say, 'I don't know' and check it out and not make false statements." These comments strongly suggest the presence of inter-rater inconsistency.

There were many comments indicating different standards of medical care among the preceptor panel. A preceptor reported that Mary's performance was excellent, however disagreed with the drug therapy recommendation. There were negative comments about the Metformin dose recommendation by Thomas. Both comments are in contradiction with the expert panel that approved the script. Another example of contradictory medical care can be seen with this pair of comments, "Good patient information gathering skills, very thorough analysis, demonstrates advanced knowledge of pharmaceutical care." This was followed with the comment "Student didn't feel as if SrCr was an issue or at least didn't mention it; also stated ASA was being used as anti-inflammatory when it's not." In another example, one preceptor complemented the thoroughness of the student (David) and specifically for recognizing the "... key role of

herbal interactions” with the Anticoagulation case. This was followed by another comment on the carelessness of the student (David) for “... missing DI of Coumadin/Amiodorone again.” Preceptor comments suggested there were disagreements on the technical aspects of the drug therapy shown in the vignette and strongly suggested different standards of medical care between preceptors and with the expert panel. This is an important source of inter-rater inconsistency.

Rater inconsistency occurs when preceptor’s application of the performance criteria is inconsistent with other preceptors.^{152, 153} Preceptors may have assigned high ratings to students who deserve low ratings and low ratings to those who deserve high ratings. This would increase assessment variation and reduce inter-rater reliability results, and limit interpretation of accuracy results.

4.2.6 Summary Preceptor Results

The ICC measures the degree of agreement between preceptors’ assessment of students scores. This test quantifies whether assessment reliability among preceptors is greater than by chance alone. In this study, the null hypothesis is $H_0: \rho = 0$ and the alternative hypothesis is $H_a: \rho \neq 0$. The competent vs. not-competent ICC point estimates range from 0.37, 0.31, and 0.30 for competency Skills A, B, and D/E respectively (Table M-1). Each competency skill has a similarly wide 95% confidence interval and ranges from 0.15 to 0.69. The ICC point estimates of the excellent vs. entry-level comparison (Table M-2) are lower and range from 0.24, 0.19, and 0.19 for skills A, B, and D/E respectively. The 95% confidence intervals are wide and range from a low 0.09 to a high 0.55. These results are greater than zero and indicate that assessment is greater than chance alone.

The global assessment ICC estimates were higher for the non-random Diabetes performances compared to the randomized cases. The ICC estimates with the rating scale also show this trend. This similarity suggests that the rating scale, to some degree, reflect the preceptor's global assessment of the student's performance. The Halo effect, rater inconsistency, large number of video vignettes, and the degree of familiarity with performance criteria would increase assessment variation. This would lower ICC point estimates and widen the 95% CI for all competency skills assessed. These factors diminished the ability of the preceptors to reliably differentiate between competent vs. not-competent student performances and between excellent vs. entry-level student performances.

One of the aims of this study was to establish benchmarks to facilitate comparison of alternate assessment instruments. Literature suggests measures of inter-rater reliability in the clinical setting are often low and estimates of 0.80 or above are rarely achieved.¹⁰⁶ Studies of medical residents^{157, 158} have shown inter-rater reliability estimates ranging from 0.79-0.87 and 0.29-0.42 for medical students.¹⁵⁹⁻¹⁶² Two physical therapy studies reported inter-rater reliability estimates ranging from 0.50 - 0.84.^{163, 164} The study reported single-measure reliability estimates with a two-way random effects ICC model. Study data is from summative assessments of cross-sectional snap shots of student performances. Examination of inter-rater reliability studies will illustrate a sample of challenges faced comparing reliability estimates among studies.

The Maxim and Dielman (1997)¹⁵⁷ study of third- and fourth-year medical students used the kappa-type statistics described by Landis and Kosh.¹⁵⁰ In another study of

medical students, Daelmans and colleagues (2005)¹⁵⁸ reported a one-facet design with preceptors nested within students to estimate variance. No details of the computational algorithms were included. Harber and Avin's (1994)¹⁵⁹ studied assessment of internal medicine residents and used a standardized mean agreement to estimate inter-rater reliability. The paper describes the standardized mean agreement is derived by calculating all variables without weights and by weighting each subject's contribution by the square root of the number of evaluations. These studies use fundamentally different computational methods and these differences limit the ability to draw meaningful comparisons.

Several medical resident studies use ICCs to estimate inter-rater reliability of preceptor assessments. Kwolek and colleagues (1997)¹⁶⁰ evaluated clinical performance of surgical residents. An average of seven preceptors rated 72 residents over one year. A different pattern of preceptors' evaluated residents each month. The inter-rater reliability of the mean overall performance was 0.82. During and colleagues (2003)¹⁶² evaluated reliability among preceptors who used the American Board of Internal Medicine's Monthly Evaluation Form. Preceptors evaluated 15 competency skills every month for one year. Reliability among preceptors for each competency skill assessed for a particular resident was evaluated and the study reported ICC results as high as 0.80. These two studies used repeated assessments of student performances. These results are difficult to compare since they do not describe the ICC models used, lack results at the individual competency level to facilitate for comparison with this study, and the repeated measures design limits the ability to make useful comparisons.

In 2002, the APTA published an evaluation of the student clinical performance instrument (CPI).¹⁶³ Two preceptors assessed student performances on a visual analog scale representing degree of supervision. The single-measure ICC was reported for each of the 22 competency skills. ICC point estimates ranged from 0.35 - 0.89 and confidence intervals were not reported. There are a number of competency skills similar to the three competency skills evaluated in this study. Student performances were observed over the four-week clerkship. A pair of preceptors gave a single summative assessment of the competency skills at the end of the clerkship. The number of students rated is included below.

- Applies principles of logic and scientific method to the practice of physical therapy (ICC=0.64, students rated=23)
- Screens patients using procedures to determine the effectiveness of and need for physical therapy services (ICC=0.41, students rated=17)
- Evaluates clinical findings to arrive at a physical therapy diagnosis (ICC=0.64, students rated=25)
- Designs a physical therapy plan of care that integrates goals, treatments, and discharge plan (ICC=0.61, students rated=25)

Meldrum and colleagues (2008)¹⁶⁴ evaluated the inter-rater results reliability of student assessments by physiotherapy preceptors in Ireland. Student clinical performances were assessed using a standardized assessment form. One summative assessment by two preceptors was completed for each of the 12 rotation periods. Scores were based on behaviorally anchored rating scales. Competencies covered three broad areas and patient management area contains many competency skills

similar to this study. The study reported an ICC of 0.75 for patient management but did not report the 95% CI. The two physical therapy studies used the single-measure reliability for a two-way random effects ICC model as used in this study. The inter-reliability results in this study are comparably lower.

The Fisher Exact test is used to estimate accuracy and gives the probability of whether assessments were unevenly distributed between the preceptor and expert panels.¹⁵¹ This test will show whether typical preceptors assess students, as compared to the expert panel, beyond chance alone. The null hypothesis was H_0 : the probability of observing no difference $\leq 10\%$ in the scores between preceptors and experts. The alternative hypothesis is H_a : the probability of observing a difference $> 10\%$ between preceptors and experts. Results are in Tables N-1, N-2, and N-3 for competency skills A, B, and D/E respectively. Unlike the reliability results, no discernable trends were observed. Interpreting the accuracy results is related to the degree of agreement among the expert panel themselves. Descriptive statistics show preceptor assessments were more lenient than the expert panel. However, the wide variation of expert panel assessments dilutes the use of these assessments as a standard for comparison. This limits the interpretation of the accuracy estimates of the preceptor panel assessments.

The study compared the rating scale with a proxy for the SUCCESS assessment strategy. Using the reposes from the same preceptors, this line of analysis gave the opportunity to compare the effect of the scale on assessment outcomes. The two scales appear to assess competency differently. The Fisher Exact tests showed the proxy scale was more lenient when assessing competency as compared to the new rating

scale. On the other hand, most of the excellent vs. entry-level comparisons were not significantly different.

Table 4-1. Delphi panel (round one) demographics

Age, years	Percent	Count
21-30	5%	1
31-40	36%	8
41-50	18%	4
50+	36%	8

Gender	Percent	Count
Male	52%	11
Female	48%	10

Degree		Count
BS Pharmacy	38%	12
PharmD	47%	15
MS	3%	1
PhD	3%	1
Other	9%	3

Years of Practice	Percent	Count
0-5	14%	3
6-10	18%	4
11-20	32%	7
21-30	9%	2
31+	23%	5

Years Precepting Students	Percent	Count
0-5	27%	6
6-10	27%	6
11-20	23%	5
21-30	14%	3
31+	5%	1

Primary role with students	Count	Count
Preceptor	73%	16
Clinical Coordinator	0%	0
Educational Coordinator	14%	3
Pharmacy Director or Manager	5%	1
Other	9%	2

Table 4-2. Delphi panel (round one) practice site characteristics

Institution Type	Percent	Count
Public	76%	16
Private	24%	5

Pharmacy Practice Setting	Percent	Count
Teaching Hospital	24%	5
Hospital	29%	6
Clinic	19%	4
Community Chain	10%	2
Community Independent	0%	0
Nursing Home	0%	0
Industry	0%	0
Other	19%	4

Population	Percent	Count
Rural (under 2,500)	28%	5
Small Town (2,500 - 9,999)	33%	6
Large Town (10,000-49,999)	22%	4
Urban (50,000 or more)	17%	3
International (non U.S.)	0%	0

State	Percent	Count
Florida	43%	9
Kansas	5%	1
New Mexico	5%	1
North Carolina	5%	1
Ohio	5%	1
Oregon	5%	1
Pennsylvania	24%	5
West Virginia	10%	2

Table 4-3. Delphi panel (round two) demographics

Age, years	Percent	Count
21-30	11%	7
31-40	41%	27
41-50	20%	13
50+	29%	19

Gender	Percent	Count
Male	36%	24
Female	64%	42

Degree	Percent	Count
BS Pharmacy	52%	34
PharmD	77%	51
MS	9%	6
PhD	2%	1
Other	12%	8

Years of Practice	Percent	Count
0-5	14%	9
6-10	32%	21
11-20	21%	14
21-30	15%	10
31+	18%	12

Years Precepting Students	Percent	Count
0-5	30%	20
6-10	30%	20
11-20	20%	13
21-30	18%	12
31+	2%	1

Primary role with students	Count	Count
Preceptor	56%	37
Clinical Coordinator	3%	2
Educational Coordinator	24%	16
Pharmacy Director or Manager	8%	5
Other	9%	6

Table 4-4. Delphi panel (round two) practice site characteristics

Institution Type	Percent	Count
Public	73%	48
Private	27%	18

Pharmacy Practice Setting	Percent	Count
Teaching Hospital	36%	24
Hospital	24%	16
Clinic	20%	13
Community Chain	0%	0
Community Independent	0%	0
Nursing Home	2%	1
Industry	0%	0
Other	18%	12

Population	Percent	Count
Rural (under 2,500)	3%	2
Small Town (2,500 - 9,999)	3%	2
Large Town (10,000-49,999)	17%	11
Urban (50,000 or more)	77%	51
International (non U.S.)	0%	0

State	Percent	Count
Alabama	6%	4
Arizona	2%	1
California	8%	5
Florida	15%	10
Georgia	2%	1
Iowa	2%	1
Massachusetts	9%	6
Michigan	8%	5
Minnesota	2%	1
Mississippi	2%	1
North Carolina	2%	1
Ohio	6%	4
Oregon	2%	1
Pennsylvania	9%	6
Texas	20%	13
Washington	2%	1
West Virginia	2%	1
Wyoming	6%	4

Table 4-5. Expert panel demographics

Age, years	Percent	Count
21-30	14%	1
31-40	43%	3
41-50	29%	2
50+	14%	1

Gender	Percent	Count
Male	29%	2
Female	71%	5

Degree	Percent	Count
BS Pharmacy	29%	2
PharmD	100%	7
MS	14%	1
PhD	0%	0
BPS Certification	71%	5
Other	29%	2

Years of Practice	Percent	Count
0-5	29%	2
6-10	57%	4
11-20	0%	0
21-30	0%	0
31+	14%	1

Years Precepting Students	Percent	Count
0-5	29%	2
6-10	57%	4
11-20	0%	0
21-30	0%	0
31+	14%	1

Primary role with students	Count	Count
Preceptor	86%	6
Clinical Coordinator	14%	1
Educational Coordinator	0%	0
Pharmacy Director or Manager	0%	0
Other	0%	0

Table 4-6. Expert panel - practice site characteristics

Institution Type	Percent	Count
Public	100%	7
Private	0%	0

Pharmacy Practice Setting	Percent	Count
Teaching Hospital	57%	4
Hospital	14%	1
Clinic	14%	1
Community Chain	0%	0
Community Independent	0%	0
Nursing Home	0%	0
Industry	0%	0
Other	14%	1

Population	Percent	Count
Rural (under 2,500)	0%	0
Small Town (2,500 - 9,999)	0%	0
Large Town (10,000-49,999)	14%	1
Urban (50,000 or more)	86%	6
International (non U.S.)	0%	0

State	Percent	Count
Alabama	29%	2
Florida	14%	1
Iowa	14%	1
Pennsylvania	14%	1
Texas	29%	2

Table 4-7. Preceptor panel demographics

Age, years	Percent	Count
21-30	29%	12
31-40	33%	14
41-50	12%	5
50+	26%	11

Gender	Percent	Count
Male	33%	14
Female	67%	28

Degree	Percent	Count
BS Pharmacy	43%	18
PharmD	71%	30
MS	2%	1
PhD	0%	0
Other	29%	12

Participants may have more than one degree and the percent column may be greater than 100%

Years of Practice	Percent	Count
0-5	29%	12
6-10	24%	10
11-20	12%	5
21-30	21%	9
31+	14%	6

Years Precepting Students	Percent	Count
0-5	55%	23
6-10	29%	12
11-20	10%	4
21-30	5%	2
31+	2%	1

Primary role with students	Count	Count
Preceptor	76%	32
Clinical Coordinator	5%	2
Educational Coordinator	5%	2
Pharmacy Director or Manager	7%	3
Other	7%	3

Table 4-8. Preceptor - practice site characteristics

Institution Type	Percent	Count
Public	69%	29
Private	31%	13

Pharmacy Practice Setting	Percent	Count
Teaching Hospital	24%	10
Hospital	17%	7
Clinic	10%	4
Community Chain	17%	7
Community Independent	2%	1
Nursing Home	5%	2
Industry	5%	2
Other	21%	9

State	Percent	Count
Conneticut	2%	1
Florida	81%	34
Maryland	2%	1
Oklahoma	2%	1
Texas	12%	5

CHAPTER 5 DISCUSSION

5.0 Overview

This study sought to show evidence supporting a proof-of-concept of a method to gather evidence demonstrating whether an APPE assessment program has content validity, and clinical faculty is able to make valid and reliable assessments. In this study of the content validity of these programs, nominated preceptors evaluated and refined the domain specification and performance criteria for competent performance in Drug Therapy and Development. This study successfully outlined a process to collect and analyze assessment data within a controlled environment. In the process of estimating reliability and accuracy, the study created performance benchmarks for future comparisons. The results demonstrated a need to continue review and refinement of relevant competency skills and performance criteria.

5.1 Phase I - Delphi Panel

A content validation study is the first step in accumulating evidence of validity and links the hypothetical construct of competency with observable student behaviors. Haynes, and colleagues (1995)³⁴ affirms the importance of content validation and argues, "A construct that is poorly defined, undifferentiated, and imprecisely partitioned will limit the content validity of the assessment instrument."

The Delphi panel conducted a content validation study and addressed two main subjects: 1) domain specification and 2) performance criteria. Specifically, the panel identified competency skills relevant for the assessment purpose. Then the panel evaluated and refined the performance criteria used to assess these competency skills (e.g. checklists, rating scales, and performance levels).

There were three prominent themes underlining many of the Delphi panel's comments. The first theme was evaluating whether each element of the domain specification and performance criteria was relevant in a clinical setting. The second theme was the role of the patient in providing clinical information, such that the pharmacist may integrate this information into an evaluation of the current and proposed medication regimens. Third, panelists felt descriptions of many of the performance criteria were written in broad terms. The panel recommended descriptions of performance criteria in the context of the clinical setting. This study showed that preceptors were able to evaluate and refine hypothetical constructs of competency and associated performance criteria.

5.1.1 Domain Specification

Domain specification is the first task in a content validation study.^{34, 100-103} The Delphi panel evaluated the relevancy of the seven competency skills in Drug Therapy and Development (see Appendix B). The panel recommended joining Skill D and Skill E together. However, the panel did not recommend Skill F. The Delphi panel commented that the backup plan described in Skill F does not play a relevant role in the clinical setting. There is no place in the patient's chart to describe backup plans, let alone a requirement to speculate about possible future problems arising from the primary therapeutic plan. Future problems are considered, but this process is an evaluation of benefits and risks towards the therapeutic objective. The panel asserted that alternative therapeutic plans tend to rely on important information that is usually unavailable during the development of the primary therapeutic plan. This is because symptoms and signs that account for the failure of the primary therapeutic plan would not be evident before

implementation of the primary plan in the first place. Theoretical backup plans and identification of likely problems may be more appropriate in the classroom setting.

The panel discussed how the activities described for skills D and E are intermixed and recommended combining these competency skills into a single skill for assessment. Panelists' comments described how activities, which consider optimal outcomes and drug performance characteristics (i.e. Skill D), and activities which consider patient preferences and social-economic factors (i.e. Skill E) tend to perform jointly. This viewpoint reflects the unique perspective of clinical faculty in which performance assessment is intertwined with clinical relevance. Studies by Bondy (1983)⁶⁶ and Lankshear (1993)⁸⁵ suggest that educators who develop assessment instruments conceptualize competent performances differently than practicing clinicians. This difference reflects different values. Educators and clinicians have similar concepts of passing students. However, clinicians distinguish relevancy by evaluation whether the competency skills or performance criteria are necessary in the clinical setting.

The performance criteria are meant to quantify performance of competencies outlined in the initial CAPE Educational Outcomes report published in 1998.¹⁶⁵ This document lists the required competencies of graduating pharmacy students from all schools of pharmacy. Leading figures in the pharmacy profession developed this document after nine years of diligent work. This report categorizes the practice of pharmacy into five general Professional Practice-Based Outcomes and seven General Ability-Based Outcomes. The report also included 34 skill statements describing the Professional Practice-Based Outcomes and the General Ability-Based Outcomes with an additional 84 specifications describing these skill statements. These competencies

formulate the attributes of a competent pharmacist. The large number of highly specified competency skills in the CAPE Educational Outcomes report is consistent with the large number of highly specified competency skills created by leaders in other healthcare professions.^{41, 166-168}

Huddle and Heudebert (2007)¹⁶⁹ disagree with the use of highly specified competency skills in clinical performance assessment, a practice that is common in the healthcare professions. The authors claim that highly specified skill statements tend to disregard the holistic connection between the practitioner and any particular clinical activity.¹⁶⁹ Huddle and Heudebert argue that healthcare practitioners are engaged in a process of responsiveness to a clinical situation in which perception may lead to the appropriate intervention or may lead to propositional reflection. This act of reflection draws on both propositional knowledge and the practitioner's judgment of the patient's problem. Consistent with this focus on clinical activity, ten Cate and Scheele (2007)¹⁷⁰ proposed that students demonstrate their degree of clinical competency as the preceptor feels the student is able to handle increasingly more demanding clinical duties.

The recommendations of the Delphi panel reduced the number of competency skills from seven to five. Pharmacy schools assess students for many more competency skills than evaluated in this study. This researcher claims that evaluating competency skill definitions according to a clinical activity viewpoint may reduce the total number of activities needed for APPE assessment. This clinical viewpoint is consistent with the stated mission of pharmacy, which is, "the provision of pharmaceutical care" according to the Commission to Implement Change in Pharmaceutical Education.^{13, 14}

5.1.2 Performance Criteria

The Delphi panel evaluated the performance criteria, which includes checklists, rating scales, and performance levels. Although there was spirited discussion of specific elements, the checklists and rating scale were generally accepted by the panel. Although panel members shared diverse recommendations, a number of important recommendations were suggested.

Consistent with a clinical perspective of performance assessment, panel members inquired when and how the student would incorporate data obtained directly from the patient. That is, how will the student integrate patient-obtained data into the medical information described into their knowledge of the patient history and examination results? Panel members pointed out the necessity for students to realize the limitations in medical documentation and recognize the need to incorporate information from the patient or proxy. One comment noted that home medication reconciliation is important in synthesizing medical information, and that medical documentation might have an incomplete home medical history. Therefore, a synthesis of medical information as described in Skill A is incomplete without including the role of patient-acquired data.

There was strong objection to some elements of the performance level descriptions. Using “maintain a full-time caseload” as an indicator of performance level was widely rejected. This performance criterion was judged as inappropriate for student pharmacists. Similar objections were made about including “expansive view of the profession”, and “capable of supervising others” in the performance level definitions. Panel members concluded that these elements were not relevant to student pharmacists. The panel asserted that these performance level descriptions were more appropriate for PGY1’s or specialty residents

Not a single panel member voiced an objection to using checklists and there were many helpful suggestions for improvement. These comments suggested a need for fundamental changes to the checklist format. As it is, the checklist describes a systematic process that a student should complete: a list of necessary steps the student must complete to demonstrate competency in a specific skill. Instead of outlining every activity, panel members suggested a benefit & risk approach to rational drug therapy, referencing appropriate evidence-based guidelines. These suggestions are consistent with the clinical viewpoint observed by this Delphi panel, wherein each activity for assessment is related to activities actually performed in the clinical setting.

Performance levels describe behavior that is less than expected, expected, or beyond the expectations of an entry-level pharmacist. Descriptions of rating scale items are written in broad terms and many panel members requested materials to help them understand the rating scale in the context of the clinical setting. Video vignettes and case studies describing differences in performance levels and associated rating scale scores would help train preceptors with student assessment.

5.1.3 The Delphi Panel

The Delphi method was used to facilitate agreement on a complex subject among a panel of experts.¹²⁴ There are no rigorous guidelines for inclusion criteria, panel size, and agreement criteria.^{116-120, 124-128} The principal researcher recruited 22 preceptors from the original list of 10 pharmacy schools in round one. The principal researcher was unable to increase panel size, never less the number of panel members was still within the 15-30 members planned. However, only eight panel members choose to participate in the second round. This number of participants was well below the recommended panel size. With these constraints in mind, the principal researcher took the unorthodox

step to recruit additional panel members directly into round two. The principal researcher contacted additional pharmacy schools to nominate additional expert preceptors and welcomed 58 new panel members directly into the second round. The panel for the second round had 66 members. Since the study used the reactive Delphi model, new members were readily incorporated into the existing panel. New and existing panel members were able to contribute valuable insights.

The principal researcher sought to recruit pharmacy preceptors with expertise in drug therapy and proficiency precepting students. The study sought to attract a mix of Board of Pharmaceutical Specialists (BPS) and non-BPS practitioners representing a variety of practice settings, regional locations, and educational institutions. Panel demographics and practice site characteristics were collected. Questions collecting information about gender, age, degree, years of practice, years precepting students, primary role with students, institution type, practice site characteristics, and population (of practice setting) were adopted from the 2008 AACP Annual National Survey of Volunteer Pharmacy Preceptors.²⁸

There were no significant demographic differences between round one panel members (Table 4-1) and preceptors who participated in the AACP survey.²⁸ On the other hand, there was one significant demographic difference with the second Delphi panel (Table 4-3), which had a greater proportion of female preceptors ($p=0.02$) compared to the AACP survey. Age or number of years typically measures clinical experience.¹⁷¹ The principal researcher expected panel members to be older, have greater clinical and preceptor experience, and hold higher degrees and certifications than the typical pharmacy preceptor. However, these measures were not significantly

different from AACP survey participants. The principal researcher did not seek to recruit the average preceptor and the impact of these demographic differences or lack of differences is difficult to measure. The low response rate of 26.5% of the AACP survey may limit the value of the comparisons made.²⁸

The principal researcher sought to recruit panel members representing a variety of practice settings, regional locations, and educational institutions. However, there were some differences in practice site characteristics compared with the 2008 AACP preceptor survey.²⁸ In the first Delphi panel, there was greater proportion of panel members practicing in public institutions ($p=0.02$) than private institutions. More panelists practiced in hospitals and clinics ($p=0.02$) and more of these practices were located in small towns and rural locations ($p=0.02$), see Table 4-2. There were two differences in the practice site characteristics of the second Delphi panel (Table 4-4) compared with the 2008 AACP preceptor survey.²⁸ This panel had a greater proportion of panelists practicing in public institutions ($p=0.05$) than private institutions and more practicing in hospitals and clinics ($p>0.00$). Panelists in the second round represented 18 states and most were from Texas (20%) followed by Florida (15%). Six percent of the panel members practiced in the states of California and Oregon. The low representation of west coast states is disappointing. These proportions represent a high level of participation by pharmacy schools in Texas and the direct marketing of preceptors in Florida by the principal researcher.

The quality of the Delphi panel's expertise impacts the quality of the panel's outcomes.¹²² The principal researcher sought to recruit pharmacy preceptors with outstanding qualities and relied on the judgment of experienced directors to nominate

participants with these qualities. Experienced directors have a demanding schedule and it is uncertain whether and to what extent these directors adhered to the selection criteria. Guion (1977)⁹⁷ suggests that panel members need to understand the boundaries of the competency skill and recognize when operationalized performance criteria are inside or outside those boundaries.⁹⁷ Therefore, training materials were used in order to familiarize the panel with conceptual foundations of performance assessment and descriptions of the operationalized performance criteria. However, whether these materials effectively prepared panel members is unknown and there was no follow-up testing of professional expertise in drug therapy or performance assessment knowledge. Independent evaluation may have ensured panelists met the desired qualifications. For example, the study could have restricted participation to recipients of preceptor awards. Zatas (1999)¹⁷² reported a self-administered questionnaire that identifies effective physician assistant preceptors. A similar instrument for pharmacy preceptors could be used to ensure preceptors with the desired expertise in drug therapy and performance assessment, an instrument that may improve this line of research in future studies.

The principal researcher provided materials outlining the competency skills and performance criteria. However, many panelists commented that these items were written in broad terms and requested examples of these skills as they occur in context of the clinical setting. Materials demonstrating examples of elements of the performance criteria in a typical clinical environment would be an effective training tool for preceptors. Video vignettes and case studies would help future participants understand how student behavior relates to the rating scale and performance levels.

Recruitment and retention was a major challenge to this study. Financial incentives and AACP support may have helped increase participation of preceptors with the desired expertise, as well as broaden regional representation. The nomination process, training, demographic makeup, and regional representation of the panel may limit the generalizability of the evidence collected. However, the panel met the objectives of the first study phase and permitted the study to continue with the next phase.

5.2 Phase II - Video Simulation

5.2.1 Video Production

The video vignette scripts were based on the recommendations of preceptors recruited from the Delphi panel. Although the expert panel received financial compensation, their professional commitment was outstanding. The principal researcher benefited from the many phone calls and email exchanges between the researcher and panel members. The enthusiasm for this research study was significant and appreciated.

The expert panel was recruited from members of the Delphi panel. The principal researcher sought to attract preceptors with considerable clinical expertise and preceptor experience. There were a number of demographic differences between the expert panel (Table 4-3) and preceptor demographics of the 2008 AACP survey.²⁸ There was a greater proportion of female preceptors ($p>0.00$). Although there was no significant differences in age, years of practice ($p>0.00$) and years precepting student ($p>0.00$) were less. The proportion of PharmD and advanced degrees was greater ($p>0.00$). The proportion of the expert panelists who reported their primary role as precepting students ($p>0.00$) was greater than reported in the AACP survey.

There were also differences in the practice site characteristics (Table 4-4) compared to the AACP survey.²⁸ There was a greater proportion of panel members practicing in public institutions ($p>0.00$) than private institutions. More panelists practiced in hospitals and clinics ($p>0.00$), and more of these practices were located in urban locations ($p>0.00$). Many members of the expert panel held key positions in teaching hospitals and pharmacy schools.

After selecting case studies to best illustrate student performances for the study, this expert panel developed a list of specific behaviors characterizing rating scale scores for specific performance levels. The objective was to simulate an authentic encounter between a preceptor and a student during APPE rotations. Third year pharmacy students portrayed the APPE students in the video vignettes based on these scripts.

The expert panel watched and scored the final video vignettes on-line. Expert panel assessments varied widely and panelists noted the difficulty of assessing the video portrayals that were based on scripts they actually wrote. Weeks and in some cases months between the last script edits and assessment of the videos took place. Familiarity with the performance criteria and the assessment instrument may have faded for some. On the other hand, one panel member commented on how the portrayals were more realistic than expected. After viewing the final video vignettes, one assessment expert, mentioned that the portrayal of these specific behaviors in the final video vignettes were subtle and might be difficult for an audience to recognize. Future studies should consider using professionals with experience in script development and

video production, as well as trained actors. These steps would help future studies produce video vignettes that more effectively portray the intended student behaviors.

With few exceptions, the diverse assessments among the panel suggest that many of the video vignettes did not effectively portray the student performances originally intended by the expert panel. This posed a serious measurement issue for the study. Video vignettes with higher assessment agreement among the expert panel would have set the stage for success in the next stage of the study. The use of video production professionals and trained actors may have increased the quality of student performances. Future studies will need to secure funding to allow re-filming of problem vignettes since they are a crucial element of this kind of study.

5.2.2 Preceptor Panel

The principal researcher sought to recruit typical preceptors to watch and score the video vignettes and expected demographic characteristics similar to preceptors who participated in the 2008 AACP survey.²⁸ However, there were a number of demographic differences (Table 4-7). The preceptor panel had a greater proportion of female preceptors ($p=0.01$) and they were younger ($p>0.00$). The panel had fewer years of practice ($p>0.00$) and fewer years precepting students ($P>0.00$). However, the proportion of PharmD and advanced degrees was greater ($p=0.02$) than PharmBS degrees. The proportion of the preceptors who reported their primary role was precepting students not significantly different. The members of this preceptor panel were younger, represented more women, and were more educated compared to the 2008 AACP survey. Age or number of years typically measures clinical experience and this panel does not represent demographics of the typical preceptor described in the 2008 AACP survey.¹⁷¹

There was no significant difference between public and private institutions or practicing in hospitals, clinics, or community settings compared to the AACP survey. More panelists practiced in large towns ($p=0.2$) rather than rural, small town, or urban locations. Practice site characteristics are shown in Table 4-8. Panelists represented five states and most were from Florida (81%) followed by Texas (12%). These proportions represent Texas Pharmacy Association's marketing efforts and the direct marketing of preceptors in Florida by the principal researcher.

5.2.3 Reliability Results

The ICC estimates the degree of agreement among preceptor assessments. Shrout and Fleiss (1979)¹⁴⁹ conceptualized the ICC as the ratio of between-groups variance to total variance. The study used a two-way random effects model and reported the single-measure reliability. Since the three performance levels (e.g. excellent, entry-level, and deficient) are hierarchical; and because this property is a violation of ICC assumptions, the study collapsed the scores into 2x2 tables for comparison. Competent students were compared to non-competent students and excellent students were compared to all other competent students as shown in Figure 3-4. Higher ICC estimates reflect increasing inter-rater reliability and perfect agreement gives an ICC result of one. Downing (2005) argues that inter-rater reliability is an essential component of validity evidence for all assessments using raters.¹⁰⁵

One of the central aims of this study is to establish benchmarks for comparison. However, the ICC is strongly dependent on the demographic variance of the population from which it is measured. The preceptors participating in this study were younger, less experienced, and more educated. There were also more women represented in this study than the 2008 AACP preceptor survey. Therefore, this panel does not represent

the demographics of the typical preceptor and the impact of this disparity on the validity of the results is uncertain. This problem may complicate any effort to generalize the results. An assessment instrument may be judged "reliable" or "unreliable," depending on the test population. Future studies should recruit participants that represent the demographic characteristics that are generalizable to the greater preceptor community.

Inter-rater reliability results from this study should be comparable to measures of reliability results in the clinical setting. Reported results are often low and estimates of 0.80 or above are rarely achieved.¹⁰⁶ Studies of medical residents¹⁵⁹⁻¹⁶² have reported inter-rater reliability estimates ranging from 0.79-0.87 and 0.14-0.42 for medical students.^{157, 159-162} Researchers explain that results are higher for residents since clinical faculty have greater opportunity to supervise students due to longer rotation schedules. In addition, residents treat patients and clinical faculty has a greater stake in supervising their performance.¹⁷³⁻¹⁷⁵ Physical therapists have reported reliability estimates ranging from 0.50-0.87^{163, 176, 177} and physical therapy students treat patients. These factors are important considerations when evaluating reliability results for pharmacy students in the clinical setting. However, comparison with other performance assessment studies can be challenging since there are different research designs and different computational methods to estimate inter-rater reliability.

Haber and Avins (1994)¹⁵⁹ questioned whether the American Board of Internal Medicine (ABIM) resident evaluation form can assess clinical competency. Faculty evaluations of 110 first year residents were analyzed in this multi-hospital study. The mean inter-rater agreement was a high 0.87; however, there was strong quantitative evidence of the halo effect. Despite the high inter-rater reliability, raters failed to

differentiate among the number of clinical care factors, and this revelation reduced the validity of assessments. This is also an example where the problem wherein professional organizations develop assessment instruments and fail to take any steps to test for reliability and accuracy. Kwolek and colleagues (1997)¹⁶⁰ studied surgical resident evaluations over a one-year period. The evaluation form contained 10 specific performance ratings and a global assessment. Inter-rater reliability of the overall performance rating was 0.82. Analysis gave strong evidence of the halo effect. The halo effect occurs when a rater's perception of a performance in one competency skill influences the assessment with other performances resulting in a failure to discriminate among different competency skills.^{153, 155, 156} Factor analysis indicated that faculty members were making a single global, undifferentiated judgment and that these ratings did not identify deficient performance skills.

Preceptor comments in this study provided graphic examples of the halo effect.^{153,}
¹⁵⁴ Literature describes three explanations for this effect. Linn and Gronlund (2000)¹⁵⁶ suggest the preceptor's general impression of the student influences rating individual performances. On the other hand, Robbins (1989)¹⁷⁸ mentions preceptor assessment may be influenced by their previous assessments. Saal, Downey, and Lahey (1980)¹⁵⁵ suggests there is a failure by the preceptor to discriminate among distinct facets of the student's performance. Preceptor comments reported assessments were influenced by previous assessment. The halo effect is a form of bias.

Noel and colleagues (1992)⁶⁸ evaluated the ratings skills of 203 medical faculty of which 96% were board certified and 74% served as clinical evaluation exercise (CEX) evaluators in the last 5 years. Faculty watched and scored two video vignettes. Each of

the 50-minute CEX simulations portrayed medical residents taking a history, physical examination, and counseling. Enough errors were included to merit a marginal rating. The results showed that half (50%) of the faculty rated each of the two simulations satisfactory or superior when both vignettes should have received a substandard rating.

Preceptor comments gave vivid accounts of rater inconsistency. This effect is present when raters inconsistently apply performance criteria compared to other raters. Longford (1994)¹⁷⁹ argues this type of rater error is a major factor in low estimates of reliability. Myford and Wolf (2004)¹⁵² proposes that rater inconsistency indicates a lack of understanding of the performance criteria. Prior to scoring the video vignettes, preceptors were given a twenty-minute introduction to assessment principles and new performance criteria. A majority of preceptors in the study practiced in Florida and should have recognized the competency skills. However, the performance criteria were new to all participants, specifically the rating scale and checklist. There was a relatively short time to become familiar with the new performance criteria. These observations suggest that assessments are unreliable when preceptors do not have a clear understanding of the assessment instructions.^{47, 180} Some of the participants in the previously described Noel (1992)⁶⁸ study watched a 15-minute instructional videotape on assessment; however, this failed to improve assessment quality. Literature strongly suggests healthcare preceptors should improve assessment skills with training.¹⁸¹⁻¹⁸⁵ Future studies could provide more substantial training prior to rating simulations. The training may mitigated some of the rater errors observed in this study. Training could focus on the judgment processes and improving detection and perception of the clinical performance. A major goal assessment training is to improve accuracy through by

decreasing common “rater errors”, like halo effect. Training will familiarize preceptors with the competency skills and performance criteria.

Preceptor comments in this study included several disagreements on aspects of drug therapy. These comments strongly recommend different standards of medical care between preceptors and with the expert panel. Findings show rater experience positively affects quality of assessments and raters who are judged as better professionals are better at rating the performance of others.⁷³ Findings in separate studies by Holmbre (2008)⁷¹ and Chapman (1998)⁸⁴ strongly suggest that clinical faculty with little experience or substandard clinical skills have more idiosyncratic assessment scores, thereby increasing score variation. This is a major source of inter-rater inconsistency since assessments are not grounded in professional standards.^{26, 27} Preceptors participating in this study were younger and less experienced, however they were more educated compared to the 2008 AACP survey.²⁸ Findings suggest that this lack of experience may increase score variation, however greater education would indicate increased clinical skills and decrease score variation.^{67, 71, 73, 84, 182}

The ICC combines two sources of rater disagreement. First, rater inconsistency and is associated whether preceptors understand the performance criteria in the same way. The second source of disagreement concerns rater bias. This is whether a preceptor's mean ratings are higher or lower compared to other preceptors. The ICC does not provide information that independently measures contributions of rater inconsistency and bias. The ICC results decrease in response to both lower correlation between raters and larger rater mean differences.

The halo effect and rater inconsistency are rater effects. These effects are due to failings in preceptor's rating skills and are not associated with measuring constructs targeted by the assessment.^{108, 186} Training has been shown to reduce rater effects and rater inconsistency. Rater effects are sources of construct irrelevant variance. These errors are important threats to the validity of inferences made from the assessment results.³³

The sample size calculation used in this study assumed that ICC estimates would be equal to or greater than 0.5. Unfortunately, this assumption was not realized and therefore the study could not measure the inter-rater reliability of individual competency skills for each student video. However, the non-randomized vignettes came closer to this target than the randomized vignettes. The results suggest the principal researcher underestimated the magnitude of the rater errors and the effectiveness of the video vignettes, impact on the ICC results.

Bonnett (2002)¹⁴⁵ determined that ICC results are a function of sample size, number of ratings per student, significance level, and confidence interval width. Since significance levels and confidence interval width are often predetermined, higher ICC results are achievable with higher agreement among preceptors, larger numbers of preceptors, or a combination of both factors. Future studies may achieve their ICC targets by increasing the number of preceptors and improving preceptor training.

The preceptor panel assessed every video vignette with the new performance criteria. Performance levels were calculated and results were collapsed into two tables for comparison. ICC estimates were made for a competent vs. non-competent comparison and for an excellent vs. entry-level comparison. However, not every

vignette gave the study an opportunity to make a distinction between excellent and entry-level performances. When all raters assess a student deficient (novice), there are no excellent or entry-level data points to evaluate let alone any rationale to make this evaluation. The lower ICC point estimates may reflect that fewer data points were available. This circumstance demonstrates how the rating per student tally influences ICC results.

5.2.4 Accuracy Results

The Fisher Exact test estimates accuracy by determining if typical preceptors assess performance in the same way as the expert panel beyond chance alone. The test gives the probability that assessments were unevenly distributed between the two panels under a hypergeometric function.¹⁵¹ Low p-values indicate that the preceptor panel assessed the performances differently than the expert panel.

The principal researcher encountered three central issues with interpreting accuracy results. First, interpreting the accuracy results is related to the degree of agreement among the expert panel. The study needed examples of high quality assessment to make useful comparisons. However, the low degree of agreement reduces the value of the expert panel's assessment as an example of high quality assessments. Twelve out of twenty-seven competency skills illustrated in the video vignettes produced unanimous agreement or had only a single dissenting panelist. These twelve competency skills matched the intended performance level from the script. These skills were bookends, in other words the extremes of the performance levels. All were either excellent or deficient (novice) performances. However, the Fisher Exact test indicated a significant difference between the expert and preceptor panels assessment ($p=0.3$). The test detected a difference Patricia's competency assessment

for her performance of Skill D/E. The preceptor panel gave mixed results, one rated excellent, seventeen (40%) rated entry-level, and 24 (57%) rated Patricia's performance deficient (novice). These assessments are very different to the expert panel. All seven (100%) assessed a deficient (novice) performance. Comments suggested rater inconsistency with the preceptor panel assessments.

Second, the Fisher Exact test assumes assessments are independent.¹⁵¹ This means that the value of a given assessment does not affect the value of another assessment. However, there is strong evidence of bias in both the expert panel and preceptor panels. Bias in the form of the halo effect interferes with assessment independence and reduces the veracity of the assessments collected.

Third, the expert panel members and the preceptors participating in this study were younger and less experienced, however they tended to be more educated compared to the 2008 AACP survey.²⁸ The impact of these demographics on assessment outcomes is difficult to measure. Findings suggest that lack of experience may increase assessment variation; however, greater education would indicate increased clinical skills and decrease variation.^{67, 71, 73, 84, 182} Any increase in assessment variation in both panels would dilute differences and increase p-values.

These issues work to reduce the value of the expert panels' results as examples of high quality assessments and reduce the veracity of the assessments collected. These issues dilute the test's capability to measure accuracy. The previously mentioned Noel (1992)⁶⁸ study underscores the distinction between reliability and accuracy, and the importance of measuring both reliability and accuracy. The study reported high inter-rater reliability estimates among a group of highly qualified clinicians with assessment

experience. However, these reliability results fail to demonstrate reasonable levels of accuracy and validity.

5.2.5 Rating Scale Comparison

Findings suggest that preceptors synthesize a number of student attributes in competency assessment and that these attributes are captured in a multifactorial rating scale.^{81-83, 187} These findings underscore the rationale to evaluate this rating scale for pharmacy APPE assessment. Florida's current APPE assessment system has a different assessment strategy in which preceptors assess student performances with a supervision rating scale. Specifically, preceptor's score the level of supervision a student's requires with a three-item scale. The supervision needs are assessed for each competency skill as shown in Figure 2-4.

Participation in the three-hour long program to collect assessment data for this study placed a heavy burden on busy professionals. However, it would be informative to compare differences between the two assessment systems. The principal researcher created a surrogate scale by collapsing the supervision domain of the multifactorial scale into a three item rating scale. This supervision proxy scale approximates SUCCESS's three level rating scale (see appendix O). The study compared assessment results of the multifactorial rating scale with the supervision proxy scale.

The Fisher Exact tests showed that most of the competent vs. not-competent comparisons. The supervision proxy scale appears to be more lenient than the multifactorial rating scale. These results are consistent with other research findings that suggest increasing scale categories has the potential to convey more information about the quality of student performance and discriminate more accurately between students.

⁶⁶ The measurement of discrimination is an essential factor for drawing meaningful

inferences. A scale with too few categories does not allow sufficient discrimination of student performances, whereas a scale with too many categories may be beyond the scale's ability to discriminate student performances, thereby increasing measurement errors.¹⁸⁸

Most of the excellent vs. entry-level comparisons were not significantly different, suggesting that the rating scale items representing excellent performances have similar degrees of discriminability between the two rating scales. Gardner (1960)¹⁸⁹ argues that validity is a function of the degree of discriminability inherent in the items being rated. Therefore, it becomes necessary to make the distinction between the maximum number of categories that a scale can discriminate and the number of categories that are meaningful in interpreting assessment scores.

The SUCCESS rating scale reflects a fundamentally different assessment strategy compared with the multifactorial rating scale. Rating supervision differs from other scale behaviors. Degree of supervision may be associated with behaviors already captured within the behavior domain of other rating scale items. Different rating scores for consistency, complexity, or efficiency may be linked to different supervision rating scores. The impact of multicollinearity on assessment results needs evaluation. On the other hand, the SUCCESS strategy lies solely with the degree of supervision. Limiting assessment to one domain of behavior may reduce the degree of discriminability. Whether one assessment strategy is superior to the other requires rigorous head-to-head testing.

5.3 Future Research

5.3.1 Content Validation

A robust assessment instrument is driven by competencies and an assessment process that make sense to its stakeholders.¹⁹⁰ A content validation study is the first essential step in accumulating evidence of validity and represents a link between the hypothetical construct of competency and measurable indicators. Specifically, a content validation study will evaluate and refine the domain specification and performance criteria that represent the standards of practice within the profession.

Healthcare organizations expend considerable time and resources creating competency statements, and the pharmacy profession is no exception.^{41, 42} The CAPE Educational Outcomes reports set the standards for accreditation, guides pharmacy education, and establishes competencies for graduating pharmacy students. CAPE published the first set of educational outcomes in 1994^{191, 192} and followed with revisions in 1998 and 2004.^{16, 165} Other pharmacy societies, including the American College of Clinical Pharmacy, American Society of Health-System Pharmacists, American Pharmacists Association, National Association of Chain Drug Stores, and National Community Pharmacists Association have published lists of competency skills for the benefit of the profession.¹⁹³⁻¹⁹⁶ These documents are long lists of general competencies that every pharmacist should possess upon graduation. These documents define pharmacy's scope of practice. Luminaries in the profession contribute considerable effort to develop these documents. Nevertheless, these documents still form a hypothetical construct of competency.

Various organizations within the organization readily describe general competencies and these competencies are broadly defined. Measuring a student's

performance of a general competency is difficult and the profession does not yet have standard guidelines to ensure valid and reliable assessments. However, healthcare educators are tasked with gathering evidence that their programs produce competent graduates. Therefore, educators prefer to use narrowly defined competency skills, which are observable and easier to assess. The drawback to this strategy is the long list of narrowly defined competency skills without a unifying set of principles to govern them. Professor ten Cate (2007)¹⁷⁰ argues these highly specified competency skills fail to reflect the original meaning of the general competency. This opinion is comparable to findings in this study. The Delphi panel recommended combining two narrowly defined competency skills together. This reconnected the newly defined competency skill with a relevant clinical activity. In another example, the Delphi panel recommended not to use one of the competency skills since it lacks a connection to the general competency in the clinical setting.

The initial steps taken by this study for a single competency skill should be expanded to evaluate all thirteen general competencies outlined in the SUCCESS instrument. This researcher recommends that any future content validation study focus on the difficult task of linking performance criteria with general competencies. The objective is to develop an APPE assessment instrument that lays the foundation to generalize the student's expected performance in similar clinical situations autonomously.⁶⁰ With adequate funding and support from pharmacy professional societies, a content validation study can be successfully completed.

The content validation study would benefit from a team member with a particular skill, and the following will explain the skill and benefit. Participants in a content

validation study for physical therapists appealed for a reduction in the number of clinical performance items. Clinical faculty in other professions has the same grievance.¹⁷⁰ The previous APTA assessment instrument contained 24 clinical performance skills. Redundant competencies were addressed and clinical performance skills were reduced to 18 items. These changes were well received by clinical faculty. However, no clinical performance behaviors were removed, but rather, the performance skills were rewritten and consolidated.²⁹ Writing well-developed and a widely accepted competency skills is a valuable art and usually requires training and experience.^{197, 198} Future content validation studies would benefit from including a team member with these skills.

5.3.2 Video Simulation - Strategy

Each video vignette in this study gave preceptors a single opportunity to observe the student performing for a specific competency skill. The use of video simulations in healthcare education is typically limited to cross sectional examples of a student's performance.^{68, 158, 159} Using a single observation per student limits the ability of the study to distinguish the impact of the quality of the case study and the quality of the video vignette on the assessment data collected.¹⁹⁹

Preceptors routinely make summative assessments based on multiple observations spanning the clinical rotation. Video vignettes illustrating multiple examples of a student performing the same competency skills with a mix of patients is one strategy. This presents multiple cross-sectional snap shots for assessment. An alternative strategy, would simulate multiple interactions between a student and preceptor for a number of competency skills revolving around consecutive exchanges with the same patient.²⁰⁰ This strategy gives a longitudinal perspective for assessment.

Either of these simulation strategies would increase the number of observations for assessment and more closely mirror the assessment challenges faced by preceptors.

Paradoxically, there is little evidence in the body of healthcare literature to guide summative assessment of multiple observations.^{182, 201} There is a need to research assessment guidelines to help preceptors give a single summative assessment for an individual competency skill based on multiple observations. A standard-setting study to establish guidelines to summarize individual observations and establish a cut-off score for summative assessment is beyond the scope of this study.^{202, 203}

5.3.3 Video Simulations - Production

After selecting case studies to best illustrate student performances for the study, this expert panel developed a list of specific behaviors characterizing rating scale scores for specific performance levels. However, after viewing the final video vignettes, several panelists mentioned that the portrayal of student behaviors in the final video vignettes were subtle and were difficult to recognize. With few exceptions, the wide assessments among the panel suggest that many of the video vignettes did not effectively portray the student performances originally intended by the expert panel. For effective illustration of important student behaviors, future studies should consider using professional video directors and scriptwriters.

The preceptor role in the video vignettes was scripted to ask the same questions to each student. This strategy was intended to minimize the influence of the preceptor's performance by emphasizing performance variation among student performances. However, the feedback revealed that many preceptors use the degree of prompting as strong indicator of performance.. The lack of this preceptor behavior may have affected

audience perception of performances and the assessments collected. Integrating this item into the scripts may improve the authenticity of the preceptor student encounter.

Third year pharmacy students portrayed the APPE students in the video vignettes. Fourth year students were not available since they were busy completing APPE clerkships. However, simulations would have benefited from student actors who have completed APPE assessments. APPE experienced student actors would benefit from training similar to simulated patients.²⁰⁴

The budget of this proof-of-concept study prevented re-filming any video vignettes with questionable quality. Future studies need to budget time and funding to re-film and reassess problem videos. The expert panel assessments were the standard of high quality assessment for comparison. This is a crucial element in measuring reliability and accuracy. As described in this study, high quality video vignettes can be used to collect valuable assessment data from pharmacy preceptors.

5.3.4 Training

The marginal assessment skills among healthcare faculty reported in the literature is consistent with the findings of this study. Cross (2000)⁴⁷ argues that assessment scores may be unreliable when preceptors do not have a clear understanding of the assessment instructions. Recently, the definitions of competency and methods of assessment have changed radically. These changes may challenge how preceptors conceptualize and assess competency.¹⁸⁰ However, there is strong evidence indicating that assessment skills improve with training.^{76, 181, 183-185, 205} Training programs are designed to familiarize the preceptors with a clear understanding of how to identify competency skills and the performance criteria used to score performances. Training to

reduce the halo effect and rater inconsistency are prominent themes since they are common issues that degrade the quality of assessment results.

To insure that physical therapy preceptors and students have a working understanding of the physical therapy assessment instrument, the APTA requires successful completion of an on-line training program prior to clerkship rotations. This requirement is based on strong evidence from an APTA-funded study measuring trainings positive influence on reliability and accuracy. Findings demonstrated preceptor assessments improved independent of previous assessment experience.¹⁸⁵

Pharmacy preceptors and students would benefit from a training program. However, comments from study participants strongly suggested that video depictions could be an effective tool to illustrate the assessment instrument in the clinical context. Brazeau and colleagues (2002)²⁰⁶ describe a teaching Objective Structured Clinical Examination (OSCE) program. Medical faculty retooled an OSCE and used it to provide medical students and clinical faculty interactive venue with clinical scenarios. Clinical faculty, guided by the assessment instrument, gave feedback as students progressed through the scenario. The researchers reported preceptors appreciated the opportunity to learn principles of assessment in a clinical context. Student responses were positive and valued the introduction to clinical performance assessment prior to clerkship rotations.

Validated video vignettes and scripts depicting APPE student performance could lay the foundation for a strong pharmacy preceptor and student training program. Video vignettes can give an introduction of the assessment instrument for preceptors. Validated scripts and supporting materials would provide a clinical scenario for teaching

students. This researcher recommends developing training material with this approach for clinical faculty and student training.

5.3.5 Rating Scale Analysis

The rating scale introduced in this study scored student performances on five domains of behavior. Preceptors were able to score from two to four rating points for each domain of behavior. Educators must decide how many points to use in a rating scale.²⁰⁷⁻²⁰⁹ According to Preston and Colman (2000)²⁰⁹ literature does not collectively outline a simple answer to the matter.

Findings suggest preceptors avoid rating students on either extreme of a rating scale and tend to rate students somewhere in the middle.¹⁵⁶ Iramaneerat and colleagues (2000)¹⁵⁶ explains this range restriction has two detrimental effects. First, this rater effect reflects preceptor's bias and not the actual student performance. Second, limiting rating variability limits reliable discrimination between competent and incompetent student performances. These detrimental effects may be mitigated with a rigorous rating scale development and testing process.

Findings suggest that test-retest and inter-item reliability increases as a function of the number of scale points. According to true score variance increases at a faster rate than error variance due to number of scale items.³⁵ This relationship contributes to the increasing reliability with increasing scale items.^{35, 210} Findings show strong evidence that increasing the number of scale categories has the potential to convey more information about the quality of student performance and discriminate more accurately among students.⁶⁶

Gardner (1960)¹⁸⁹ argues that validity is a function of the amount of discriminability inherent in the items being rated. The measurement of discrimination is an essential

factor to draw meaningful inferences. A scale with too few categories does not allow sufficient discrimination of student performances, whereas a scale with too many categories may be beyond the scales ability to discriminate student performances increasing measurement errors.¹⁸⁸ According to Komorita and Graham (1965)¹⁸⁸, the ultimate criterion for adoption of a particular number of scale points is its effect on validity. When the response scale does not correlate with a criterion, the validity of the scale may not be affected despite an increase in reliability.¹⁸⁸ The comparison of the new multifactorial scale with the supervision proxy scale demonstrates the influence of the rating scale makeup on assessment results. There is strong evidence that too few or too many rating items can negatively affect reliability, validity, and interpretation of assessment scores.^{73, 78}

Findings strongly suggest preceptors synthesize a number of student attributes in competency assessment and these attributes are captured in a multifactorial rating scale.^{81-83, 187} These findings underscore the rationale to evaluate a multifactorial rating scale for pharmacy APPE assessment. Further evaluation of rating scale items is warranted.

5.3.6 Performance Levels

This study adapted the three performance levels from Florida's SUCCESS instrument. Summative assessments fall into one of three performance levels as described in Figure 4-2. The three levels are: excellent, entry-level, and deficient (novice). However, other healthcare assessment instruments have used five or more performance levels and there is a call for the SUCCESS instrument to increase the number of performance levels.²¹¹

Finding from a study by Adams and colleagues (2008)²¹² evaluated the multifactorial rating scale and performance levels with the APTA national assessment instrument. The multifactorial rating scale in this study was adapted from the APTA rating scale. This study analyzed data from seven graduating classes and reported strong evidence preceptors were able to discriminate six level of student performance. The rational for adoption of the six performance levels was based on different minimally acceptable performance levels among the participating physical therapy schools. These results suggest the multifactorial rating scale adopted in this has the potential to discriminate six levels of student performance and merits investigation.

5.3.7 Beyond Graduation

The objective of APPE assessment instrument is to generalize about a student's expected performance in similar clinical situations autonomously.⁶⁰ Evaluating the performance of graduates would help schools measure whether the APPE assessment outcomes have any value in predicting future performance. Follow-up research would inform schools how assessment outcomes address aid in teaching and learning. Follow-up studies would provide the benchmark for setting meaningful performance standards.

5.3.8 National Validation Study

Schools of pharmacy have developed proprietary APPE assessment instruments or have worked within regional consortiums. The development and maintenance of assessment instruments consume substantial resources. There is the cost to develop the assessment instrument. Development requires expertise in performance assessment, the APPE learning environment and computer programming. Preceptor training requires expertise in developing and distributing educational materials. Maintenance of software application requires technical personnel, fees for internet

access if web-based, and maintenance of computers. Development costs for a national effort would reduce individual schools costs by distributing costs among the schools. Without valid and reliable national assessment instrument, the pharmacy profession is unable to ensure the competency of graduating students.

The 2008 AACP president and the ACCP Educational Affairs Committee have called for a standard APPE assessment instrument.^{21, 22} Healthcare education experts argue the need for reliable and valid student assessments.¹¹¹ Assessment instruments need to be defended with rigorous scientific methods. Evaluation needs to predict whether students are capable of performing in real world clinical settings.¹¹²⁻¹¹⁴ Despite this widely held position, the pharmacy profession has yet to establish a national policy outlining acceptable validation criteria for APPE assessment instruments in a meaningful way. Without rigorous scientific evaluation, there is no assurance that any of the proprietary APPE assessment instruments assess student performance in a meaningful way. Clearly, the public expects pharmacy schools to graduate competent practitioners, and it is only reasonable that pharmacy schools to be able to demonstrate graduates are competent. Sound assessment is a matter of public safety and trust. This study, in combination with the recommended studies in this section, outlines a sound validation method. This important effort warrants a national commitment.

5.4 Summary and Conclusions

APPEs account for 30% of the professional curriculum.²⁸ Recommendations to increase APPEs from one year to two years have been proposed.²¹³ This increase means APPEs could account for 50% of the curriculum. APPEs provide the only opportunity for students to refine clinical skills, under the guidance of an experienced pharmacist. Assessment during APPE rotations is not the only assessment strategy

used to examine students. However, they represent the only evaluation opportunity in the clinical setting. Students have a critical need for guidance.^{154, 214, 215} Findings from a study by Langendyk (2006)²¹⁵ underscore the need for guidance. Findings showed that low-achieving medical students gave inaccurate self-assessments and peer reviews. Paradoxically, high-achieving students were harsher on their performances than faculty. Efforts to train self-assessment skills were not always successful. Schools need to produce competent graduates and meaningful assessment relies on experienced practitioners.

The pharmacy profession is awash with a never-ending list of competencies. However, these lists describe hypothetical constructs of competency. Research question #1 in this study showed preceptors were able to evaluate and refine hypothetical constructs and associated performance criteria. Evaluations were grounded in relevancy to the clinical setting. Patient-centered practices were evident from recommendations to integrate patient data and preferences into the competency statements. In keeping with this clinical perspective, preceptors requested training materials to show competencies and the performance criteria in the clinical context.

The study demonstrated a systematic approach to illustrate student clinical performance with video vignettes. The principal researcher coordinated the input of an expert panel of preceptors to refine the pharmacy case studies, develop the format of a student reporting to a preceptor, and created preceptor questions. The panel developed specific student behaviors indicative to specific scores with the new multifactorial rating scale. These behaviors were building blocks to construct student behavior indicative of

performance at different levels. The step-by-step process demonstrated in this study can be used to develop new scripts and video vignettes.

Typically, pharmacy students mentor under a single preceptor during APPE assessment. The principal researcher needed to develop a controlled environment to collect assessment data that allowed preceptors to assess the same student performance. This study demonstrated the use of video vignettes to address research questions #2 and #3, specifically collecting evidence of assessment reliability and accuracy. Findings from this study showed low inter-rater reliability results. However, reliability results were higher for non-randomized compared to randomized simulations. Accuracy results showed preceptors more readily identified high and low student performances compared to average performing students. Findings in this study demonstrated preceptors were able to synthesize a number of student attributes with the new multifactorial rating scale and accepted the new checklists. The study demonstrated a sound method to collect assessment data for rigorous analyses.

The study noted the presence of the halo effect and rater inconsistency. Research has shown healthcare preceptors improve assessment skills with training.¹⁸¹⁻¹⁸⁵ The study noted differences in standards of care among the participating preceptors and the expert panel. Findings suggest clinical faculty with little experience or substandard clinical skills have more idiosyncratic assessment scores.^{71, 84} These examples describe a major source of rater inconsistency since assessments are not grounded in professional standards.^{26, 27} A study by Colthart and colleagues (2008)²¹⁶ suggested poorly performing clinicians are the least able to self-assess accurately.²¹⁶ This is consistent with studies suggesting the poorly performing clinicians are not necessarily

aware of their deficiencies despite some with high self-confidence.^{214, 217, 218} This study demonstrated pharmacy preceptors have the same vulnerabilities rating student performances as other healthcare preceptors, indicating the need for training in performance assessment. Video vignettes similar to the simulations created in this study can be used to build a robust preceptor training and testing program.

Never the less, pharmacy APPE rotations appear to train and assess students. Students are exposed to a number of clinical environments and preceptors during their APPE rotations. Findings from this study show preceptors have the highest accuracy with extreme performances. That is, preceptors noticed students with the highest and lowest performance levels. The degree of inter-rater reliability is dependent on the population of preceptors, student rating opportunities, and the instrument.³⁹ The greater variety of preceptors and the greater number of student rating opportunities increase inter-rater reliability results. The more time spent in APPE rotations increase reliability of assessments. It is hoped there are enough rating opportunities from proficient preceptors to identify poorly performing students. Findings from Brown (2000)¹⁸⁷ suggest preceptors may intervene with troubled students without regard to the assessment instrument. Lankshear (1990)²¹⁹ describes barriers preceptors face just to give accurate assessment of poorly performing students. These are examples where preceptors are left to their own devices without the support a robust assessment system. However, many preceptors simply pass students to the next rotation.^{85, 219} Pharmacy preceptors' need valid and reliable assessment instruments to document poorly performing students and to start the remediation process earlier rather than later.

There are applications for robust pharmacy assessment instruments beyond APPE clerkships. As the pharmacists transform their role in the healthcare and move from dispensing to clinical services, licensing examinations need to reflect this strategic change. Currently the United States medical licensing examination (USMLE) use simulated patients in the licensing process.²²⁰ The USMLE Step II Clinical Skills Exam's objective is to assess clinical competency. Like the medical licensure examination, a pharmacy version would need a valid and reliable assessment instrument applicable to newly graduated students nationwide. The validation process outlined in this study would address this requirement.

Taking a page from medical literature, portfolios are used to document professional development.²²¹ Medical residents routinely use their portfolios as evidence of clinical skills in employment interviews. PharmD students could use the results from a widely accepted instrument as evidence of outstanding clinical skills with employment interviews or applications for pharmacy residency programs. Employers and residency programs look for evidence of clinical skills in the selection process.

The residency and clerkship programs for most healthcare professions require a fixed time length. Many pharmacy residency programs range from one to two years and medical residency programs are often longer. During this time, residents must complete the training objectives outlined in the program. These residency programs are required for board certification. Medical institutions and licensing authorities ponder the change to a competency-based curriculum.^{112, 170, 222-225} Emerging research suggest acquisition of clinical skills can be accelerated within a competency-based program compared to the current standard time-based approach.^{112, 223, 224, 226, 227} The foundation to adopt a

competency-based program is grounded with a valid and reliable assessment instrument. Comments from the Delphi panel indicated many of the rating scale items and performance level descriptions were appropriate for pharmacy residents. The past decade has seen the rapid growth of BPS pharmacy specialties to meet the growing need for highly trained clinical pharmacists. The BPS program would benefit from valid and reliable assessment instrument to document clinical competency of residents and evaluate program effectiveness.

Healthcare education experts argue the need for reliable and valid assessment instruments.¹¹¹ In part, this study was inspired by a study by Dr. Hubbard and colleagues.²²⁸ In 1963, he examined a high stakes examination for US medical students. When a student failed this test, they did not progress to graduation. This was a bedside oral examination rated by a number of experienced physicians. Three years of results with over 10,000 examination outcomes showed weak inter-rater reliability. For decades, this test seemed reasonable to highly trained physicians. However, until someone took the time to analyze assessment data, no one realized the test did not assess student performance as expected. Nevertheless, it took almost ten years for the medical examination board to stop using this test. Dr. Hubbard's study demonstrated evaluation of assessment instruments need scientific evaluation. However, the pharmacy profession has yet to establish a national policy outlining validation criteria for APPE instruments. Evidence supporting or challenging assessment instruments needs grounding within rigorous scientific methods. The central purpose of this study is to demonstrate a rigorous method to evaluate the validity and reliability of APPE

assessment programs. This study is a step in establishing a culture of continuous assessment.

APPENDIX A SUCCESS COMPETENCIES

The SUCCESS instrument contains 96 skills statements categorized within 13 competencies.

1. Drug distribution systems
2. Disease State Knowledge
3. Drug therapy evaluation and development
4. Monitoring for Endpoints
5. Patient Case Presentations
6. Patient Interviews
7. Patient Education/Counseling
8. Drug Information
9. Formal Oral Presentations
10. Formal Written Presentations
11. Professional team interaction
12. Professionalism/Motivation
13. Cultural Sensitivity

APPENDIX B
DRUG THERAPY EVALUATION AND DEVELOPMENT

Skill A	Synthesizes complete patient history and laboratory and physical exam data to identify problems.
Skill B	Identifies and prioritizes both actual and potential drug related problem stating rationale.
Skill C	Identifies problems that require emergency medical attention
Skill D	Designs and evaluates treatment regimens for optimal outcomes using pharmacokinetic data and drug formulation data.
Skill E	Designs and evaluates treatment regimens for optimal outcomes using disease states and previous or current drug therapy as well as including psycho-social, ethical-legal, and financial data.
Skill F	Develops backup plans based on what problems are likely to occur from/with the primary plan.
Skill G	Provides written documentation of the pharmaceutical care plan that is clear, complete, and concise

Figure B-1. Seven competency skills

Skill A: Synthesizes complete patient history and laboratory and physical exam data to identify problems

Excellent: Independently synthesizes complete patient history, laboratory and physical exam data (collects this data if necessary) to identify most if not all problems.

Entry-level: With preceptor, guidance synthesizes complete patient history, laboratory and physical exam data (using incomplete data at times) to identify the most critical problems.

Deficient (Novice): Even with preceptor guidance, the student has difficulty synthesizing patient history, laboratory and physical exam data (makes no effort to fill in the gaps in information) to identify problems.

Skill B: Identifies and prioritizes both actual and potential drug related problem stating rationale (** critical skill)

Excellent: Independently identifies and prioritizes most if not all actual and potential drug related problems stating rationale for prioritization.

Entry-level: With guidance from the preceptor identifies and prioritizes the most critical actual and potential drug related problems stating rationale for prioritization when necessary.

Deficient (Novice): Even with preceptor guidance, the student has difficulty identifying and prioritizing both actual and potential drug related problems. Does not state rationale for prioritization. Assistance required preventing errors.

Skill C: Identifies problems that require emergency medical attention (** critical skill)

Excellent: Independently identifies any problems that require emergency medical attention and also identifies what steps should be taken to activate emergency procedures.

Entry-level: With preceptor, guidance identifies problems that require emergency medical attention and also identifies who to contact to determine what steps should be taken to activate emergency procedures with occasional assistance.

Deficient (Novice): Even with guidance from the preceptor, the student is not able to identify problems that require emergency medical attention and also does not know who to contact to determine what steps should be taken to activate emergency procedures. Preceptor intervention required to prevent errors.

Skill D: Designs and evaluates treatment regimens for optimal outcomes using pharmacokinetic data and drug formulation data. (** critical skill)

Excellent: Independently designs and evaluates most if not all treatment regimens for optimal outcomes using pharmacokinetic data and drug formulation data.

Entry-level: Designs and evaluates the most critical treatment regimens for optimal outcomes using pharmacokinetic data and drug formulation data. Requires preceptor's assistance for a more detailed evaluation.

Deficient (Novice): Even with preceptor's guidance, the student is not able to design or evaluate regimens for optimal outcomes using pharmacokinetic data and drug formulation data. Preceptor intervention required to prevent errors.

Skill E: Designs and evaluates treatment regimens for optimal outcomes using disease states and previous or current drug therapy as well as including psychosocial, ethical-legal, and financial data. (** critical skill)

Excellent: Independently designs and evaluates most if not all treatment regimens for optimal outcomes using disease states and previous or current drug therapy including psychosocial, ethical-legal, and financial data using documentation from a reliable source.

Entry-level: Designs and evaluates the most critical treatment regimens for optimal outcomes using disease states and previous or current drug therapy including psycho-social, ethical-legal, and financial data using documentation from a reliable source. Requires some assistance from the preceptor to produce more detail analysis.

Deficient (Novice): Even with guidance from the preceptor the student is not able to design and evaluate treatment regimens for optimal outcomes using disease states and previous or current drug therapy including psycho-social, ethical-legal, and financial data. Fails to use documentation from a reliable source. Preceptor intervention required to prevent errors.

Skill F: Develops backup plans based on what problems are likely to occur from/with the primary plan.

Excellent: Independently develops backup plans based on what problems are likely to occur from/with the primary plan for most if not all drug therapy problems.

Entry-level: Develops backup plans based on what problems are likely to occur from/with the primary plan for the most critical drug therapy problems. Requires some assistance for more detailed planning.

Deficient (Novice): Even with guidance from the preceptor, the student rarely develops backup plans based on what problems are likely to occur from/with the primary plan. Assistance required preventing errors.

Skill G: Provides written documentation of the pharmaceutical care plan that is clear, complete, and concise.

Excellent: Independently provides written documentation of the pharmaceutical care plan that is clear, complete, and concise.

Entry-level: Provides written documentation of the pharmaceutical care plan that is complete, but could be more concise and/or clear. Requires guidance to produce detail documentation.

Deficient (Novice): Either provides no written documentation of the pharmaceutical care plan or provides documentation that is not complete. Preceptor intervention required to prevent errors.

APPENDIX C
IRB DOCUMENTS



PO Box 112250
Gainesville, FL 32611-2250
352-392-0433 (Phone)
352-392-9234 (Fax)
irb2@ufl.edu

October 2, 2009

TO: Charles Douglas
7500 Kirby Drive, Apt. 131
Houston, TX 77030

FROM: Ira S. Fischler, PhD; Chair *isfled*
University of Florida
Institutional Review Board 02

SUBJECT: **Exemption of Protocol #2009-U-1000**
Pharmacy Preceptor Delphi Panel

SPONSOR: None

The Board has determined that your protocol is exempt from review. This exemption is issued because this protocol does not involve the use of human participants in research in accordance with 45 CFR 46. Human participants are defined by the Federal Regulations as living individual(s) about whom an investigator conducting research obtains (1) data through intervention or interaction with the individual; or (2) identifiable private information.

Should the nature of your study change or if you need to revise this protocol in any manner, please contact this office before implementing the changes.

IF:dl

An Equal Opportunity Institution

DATE: March 26, 2010

TO: Charles Douglas
7500 Kirby Drive Apt. 31
Houston, TX 77030

FROM: Ira S. Fischler, PhD; Chair *ISF*
University of Florida
Institutional Review Board 02

SUBJECT: **Approval of Protocol #2010-U-289**
An Operational Definition of Student Clinical Competency

SPONSOR: Pharmaceutical Outcomes and Policy UF

I am pleased to advise you that the University of Florida Institutional Review Board has recommended approval of this protocol. Based on its review, the UFIRB determined that this research presents no more than minimal risk to participants. Your protocol was approved as an expedited study under category 7: *Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.*

Given this status, it is essential that you obtain signed documentation of informed consent from each participant. Enclosed is the dated, IRB-approved informed consent to be used when recruiting participants for the research. If you wish to make any changes to this protocol, including the need to increase the number of participants authorized, you must disclose your plans before you implement them so that the Board can assess their impact on your protocol. In addition, you must report to the Board any unexpected complications that affect your participants.

It is essential that each of your participants sign a copy of your approved informed consent that bears the IRB approval stamp and expiration date.

Your approval is valid through March 22, 2011. If you have not completed the protocol by this date, please telephone our office (392-0433), and we will discuss the renewal process with you. It is important that you keep your Department Chair informed about the status of this research protocol.

ISF:d

October 14, 2010

TO: Charles Douglas
7500 Kirby Drive Apt. 31
Houston, TX 77030

FROM: Ira S. Fischler, PhD; Chair, *ISF*
University of Florida
Institutional Review Board

SUBJECT: Revision of Protocol #2010-U-0289
An Operational Definition of Student Clinical Competency

SPONSOR: Pharmaceutical Outcomes and Policy UF

The request to revise the above referenced protocol has been reviewed and approved. Approval of this study is valid through March 22, 2011.

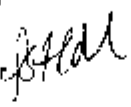
The Board must review any further revisions to this protocol, including the need to increase the number of participants authorized prior to implementation.

IF:dl

- revised paper-based consent form – changed estimated time to 3 hrs and indicated time of participation;
- revised paper-based consent form that includes compensation (\$100 gift card)
- created an on-line version of the consent

April 28, 2011

TO: Charles Douglas
5151 Edloe Street #11107
Houston, TX 77005

FROM: Ira S. Fischler, PhD; Chair, 
University of Florida
Institutional Review Board 02

SUBJECT: **Exemption of Protocol #2010-U-0289**
An Operational Definition of Student Clinical Competency

SPONSOR: Pharmaceutical Outcomes and Policy UF

Because the portion of this research that involved human participant contact has been completed and no further participants will be recruited, it is exempt from further review by this Board in accordance with 45 CFR 46 under category 8

(i) the research is permanently closed to the enrollment of new subjects; (ii) all subjects have completed all research-related interventions; and (iii) the research remains active only for long-term follow-up of subjects; or (c) where the remaining research activities are limited to data analysis.

UFIRB approval should no longer be necessary. Should the nature of your study change or if you need to revise this protocol in any manner, please contact this office before implementing the changes.

IF:d

**Attachment D
CE Course Informed Consent Letter**

Informed Consent Letter

Protocol Title: An Operational Definition of Student Clinical Competency

Please read this consent document carefully before you decide to participate in this study.

Purpose of the research study: The purpose of this study is to measure reliability and accuracy of pharmacy preceptor scores using a new clinical performance assessment scale. Pharmacy preceptors would use this new scale to assess student performance during clinical clerkships.

What you will be asked to do in the study: The study will describe how to use a new assessment scale of student performance during APPE clinical clerkships. The study will present nine (3-5 minute) video vignettes illustrating student performances and ask you to assess the performances with the new scale.

Time required: We estimate a total of 1 hour. Your participation will occur on a scheduled date between June 1st and July 31st.

Risks and Benefits: The study anticipates minimal risks associated with the CE class. The study anticipates minimal risk to your privacy due to improper release of data. You may benefit from exposure to the CE course.

Compensation: You will not be compensated for participating in this research study.

Confidentiality: Your identity will be kept confidential to the extent provided by law and your name will not be collected or used in any reports.

Voluntary participation and your right to withdraw from the study: Your participation in this study is completely voluntary. There is no penalty for not participating. You have the right to withdraw from the study at anytime without consequence.

Whom to contact if you have questions about the study: If you have any questions about this research protocol, please contact me at (352) 672-0595 or my faculty supervisor, L. Douglas Ried, at ried@cop.ufl.edu

Whom to contact about your rights as a research participant in the study: IRB02 Office, Box 112250, University of Florida, Gainesville, FL 32611-2250; phone (352) 392-0433.

Agreement: I have read the procedure described above. I voluntarily agree to participate in the procedure and I have received a copy of this description.

Participant: _____ Date: _____

Principal Investigator: Chris Taylor Date: 24 June 2010

Approved by
University of Florida
Institutional Review Board 02
Protocol # 2010-U-0289
For Use Through 03-22-2011

APPENDIX D DELPHI PANEL ROADMAP

Study Overview

Based on the Center for the Advancement of Pharmaceutical Education's (CAPE) competencies, the SUCCESS instrument is used to assess student performance during advanced pharmacy practice experiences (APPEs). SUCCESS is the result of collaboration between all four Florida schools of pharmacy. It addressed the burden of multiple assessment systems from different schools.

The purpose of this study is to collect validity evidence for the content of the *Drug Therapy Evaluation and Development* competency used in the SUCCESS instrument. Content validity is an essential step in validation, represents the link between targeted constructs with measurable indicators, and addresses the question: "Are we actually measuring what we intend to measure?" Competency in drug therapy is essential to the practice of pharmaceutical care, which is the mission of pharmacy education. *Drug Therapy Evaluation and Development* is one of thirteen competencies within SUCCESS.

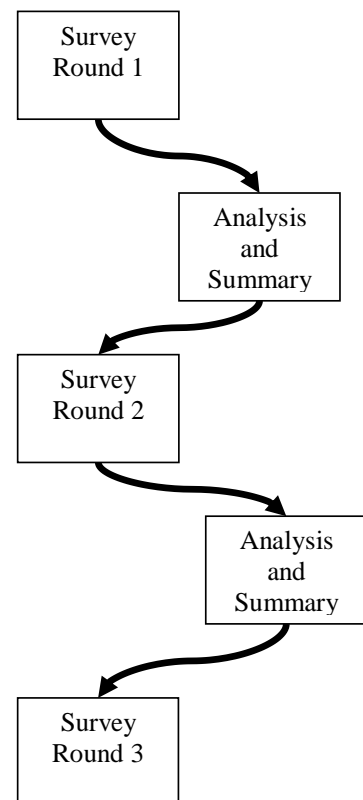
The study has two phases, and you have been asked to participate in the first phase exclusively.

In the first phase, a Delphi panel is charged with the critical examination of the operational definitions used to assess clinical performance. The panel will evaluate the performance criteria for relevancy and representativeness, as well as make recommendations, if warranted. In addition, the panel will also develop a glossary of standard terms. The Delphi method uses a series of survey rounds, along with anonymous feedback from the previous round, which allows an expert panel to quickly reach agreement on complex issues.

It is estimated that each Delphi survey will require 45-60 minutes to complete and the study will need 2 or 3 survey rounds to reach 80% agreement. The Delphi method has been widely used to generate expert consensus in health care education. The knowledge gained from this study will be useful for the validation of a national APPE assessment instrument and the development of national training materials for preceptors.

The Delphi process

The Delphi method uses a series of survey rounds with anonymous feedback. This allows an expert panel to freely share ideas and form a consensus quickly. In the first survey round, the panel will complete an on-line survey of



the existing and newly-developed operational descriptions. Completed surveys are automatically sent back to the study to summarize the responses and comments.

The same survey is used for the second round but will contain a summary of the previous round. The summary is both quantitative and qualitative. This will provide the panel with relevant data that can be easily interpreted. The percent response to each question will be provided. A thematic analysis of the comments from the panel members will be categorized by key words and compiled into a summary.

If agreement is not achieved by the end of the second round, a third and final round will be conducted. Panel agreement is defined at 80%. To reduce the burden on the panel's participation, items under discussion will be removed once the panel has reached agreement.

Survey Steps

You will receive an email with a link to the on-line survey. After reading the consent form on the first page of the survey, clicking on the "I accept" button will serve as your agreement to participate in study as described. This is followed by the instructions page.

Step 1) You will be asked if you think the seven skills included in the *Drug Therapy Evaluation and Development* competency are relevant and comprise a complete description of the competency. You will have the opportunity to add, remove or refine skill statements.

- Relevancy – Whether the individual skill statement is necessary for assessing the *Drug Therapy Evaluation and Development* competency.
- Comprehensive – Are there any aspects of pharmacy knowledge, attributes, or skills missing in any of the skill statements

Step 2) You will be asked to evaluate about the relevancy and representativeness of the performance criterion (e.g. checklist and rating scale) for each of the seven skills under *Drug Therapy Evaluation and Development* competency.

- Relevancy – Whether all the elements of the checklist and the rating scale behaviors are relevant for assessing *Drug Therapy Evaluation and Development* competency.
- Representativeness - Considering the universe of performance measures, are the elements of the performance criterion (e.g. checklist items and rating scale behaviors) representative of the skills, knowledge, and attitudes required to assess *Drug Therapy Evaluation and Development* competency.
- "Critical" skills - Next, you will be asked to evaluate if the skill is a "critical" skill. Some skills are labeled "critical" skills and students must show competency. A student performing a critical skill at an unsatisfactory level could potentially be harmful to a patient or to the practice site. Student remediation may be necessary and the preceptor would want to notify the APPE director.

What you will see:

Seven skill statements comprise the *Drug Therapy Evaluation and Development* competency. Each of the skills has a checklist and rating scale. In addition, some sub-competencies are labeled “critical.”

- The checklist is a description of activities the student needs to perform to competently complete an individual skill.
- The rating scale describes student behaviors for a particular performance level (e.g. excellent, competent, and deficient). You will be asked to evaluate the behavior descriptions in terms of the following six performance dimensions.
 - Supervision: refers to the level and extent of assistance.
 - Quality: refers to the degree of knowledge and skill proficiency demonstrated.
 - Complexity: refers to the number of elements that must be considered relative the expected performance an entry-level practitioner.
 - Reliability: refers to the frequency of occurrences of desired behaviors related to the performance criterion.
 - Efficiency: refers to the ability to perform in an effective and timely manner.

- Some skills are labeled a “critical” skill and students must show competency in these. A student performing a critical skill at an unsatisfactory level could potentially be harmful to a patient or to the practice site. Student remediation may be necessary and the preceptor would want to notify the APPE director.

Step 3) There are a few questions concerning a glossary of terms. The development of a standard glossary is intended to support consistent and clear interpretations of the performance criterion by preceptors.

Step 4) Finally, you will be asked for some basic demographic information. This will help study understand the composition of experts who participated in the Delphi panel.

Thank you for your participation and I look forward to your input.

Charles Douglas, EMBA
PhD student, UF College of Pharmacy
cadoug@ufl.edu

APPENDIX E
ANALYSIS PATH AND ASSESSMENT RUBRIC

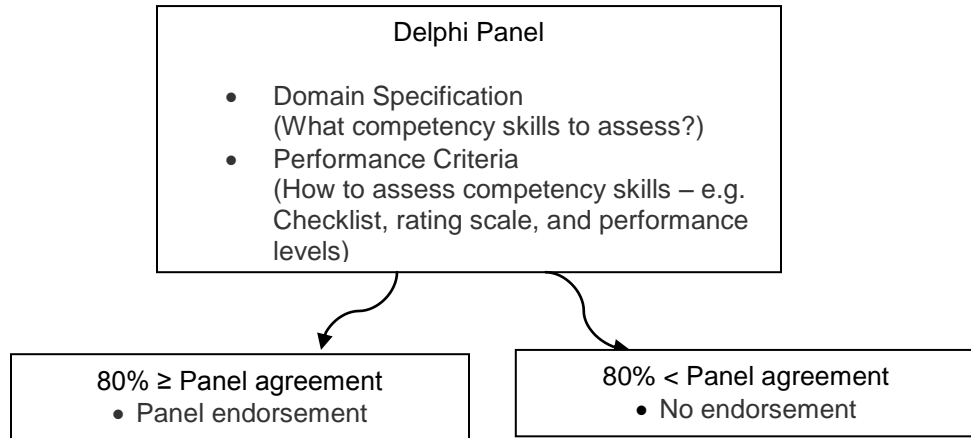


Figure E-1. Delphi panel endorsement criteria

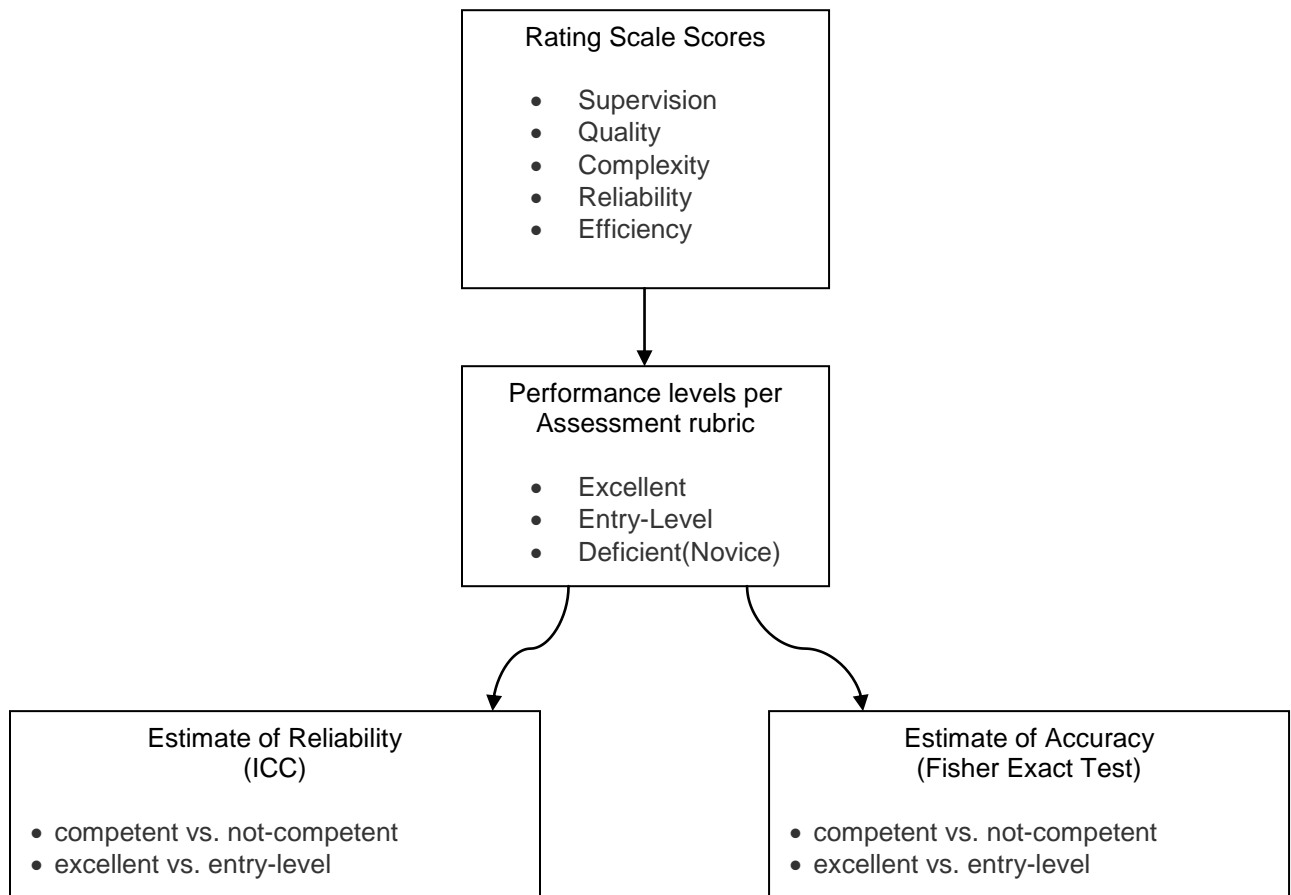


Figure E-2. Data collection and analysis pathway

Rating Scale

- Supervision: Close supervision (constant monitoring)
 More than standard level of supervision (for entry level)
 Standard level of supervision (for entry level)
 Minimal supervision (beyond entry level)
- Quality: Shows limited knowledge & skills
 Shows entry level knowledge and skills
 Shows knowledge and skills beyond entry level
- Complexity: Proficient with simple cases
 Proficient with complex cases (for entry level)
- Consistency: Inconsistent or poor performance
 Consistently proficient
- Efficiency: Inefficient/slow
 Effective/timely (for entry level)

Assessment rubric for “Excellent” performance

- Supervision: Close supervision (constant monitoring)
 More than standard level of supervision (for entry level)
 Standard level of supervision (for entry level)
 Minimal supervision (beyond entry level)
- Quality: Shows limited knowledge & skills
 Shows entry level knowledge and skills
 Shows knowledge and skills beyond entry level
- Complexity: Proficient with simple cases
 Proficient with complex cases (for entry level)
- Consistency: Inconsistent or poor performance
 Occasional lapses
 Consistently proficient
- Efficiency: Inefficient/slow
 Effective/timely (for entry level)

Assessment rubric for “Entry-level” performance

- Supervision: Close supervision (constant monitoring)
 More than standard level of supervision (for entry level)
 Standard level of supervision (for entry level)
 Minimal supervision (beyond entry level)
- Quality: Shows limited knowledge & skills
 Shows entry level knowledge and skills
 Shows knowledge and skills beyond entry level
- Complexity: Proficient with simple cases
 Proficient with complex cases (for entry level)
- Consistency: Inconsistent or poor performance
 Occasional lapses
 Consistently proficient
- Efficiency: Inefficient/slow
 Effective/timely (for entry level)

Assessment rubric for “Deficient (Novice)” performance:

- Supervision: Close supervision (constant monitoring)
 More than standard level of supervision (for entry level)
 Standard level of supervision (for entry level)
 Minimal supervision (beyond entry level)
- Quality: Shows limited knowledge & skills
 Shows entry level knowledge and skills
 Shows knowledge and skills beyond entry level
- Complexity: Proficient with simple cases
 Proficient with complex cases (for entry level)
- Consistency: Inconsistent or poor performance
 Occasional lapses
 Consistently proficient
- Efficiency: Inefficient/slow
 Effective/timely (for entry level)

APPENDIX F
SKILL STATEMENTS AND CHECKLISTS

Skill statement A:

Synthesizes complete patient history, laboratory, and physical exam data to identify problems

Checklist A:

1) Read the history and identify:

- Age, gender and race of the patient
- Symptoms and signs of disease
- Past medical history

2) Read the physical examination and look for signs of disease by system (e.g. cardiovascular or respiratory disease?). Read the laboratory and look for:

- Laboratories that reflect function of the system affected or associated with patient complain.
- Are the laboratories results normal or abnormal? What is the trend of the values (if previous laboratory results are available)?
- Look for the rest of laboratories (not associated with the specific symptom or system affected). Are they normal or abnormal? (interpretation of results)

Skill statement B:

Identifies and prioritizes both actual and potential drug related problem stating rationale.

Checklist B:

1) List the problems of this patient:

- Prioritize problems according to the impact on mortality and morbidity for this patient (e.g. treatment of hypertension has a greater impact decreasing mortality and morbidity than eczema)

2) List the current medications of this patient:

- From the current medication list, look for unnecessary medication
- From the current medication list look for medications that are contraindicated because can aggravate some of the symptoms or diseases of the patient
- From the current medication list, look for medications that can cause side effects similar or relevant to the patient's complaint.

Skill statement D/E:

Designs and evaluates treatment regimens for optimal outcomes using pharmacokinetic and drug formulation data. Design considers psycho-social, ethical legal and financial factors.

Checklist D/E:

- 1) Review in detail possible therapeutic options and if any medication could affect negatively (have side effects) the other affected systems in this specific patient.
- 2) Are there any better medications for this patient based on his/her:
 - Age (some medication are act different in children than in adults)
 - Gender (some medications are better according to gender)
 - Ethnicity (some medications are better for some ethnicities)
 - State of health (e.g. pregnancy or immediate post-surgical state)
 - Affected systems (the medication is good for the primary problem is treated but it is contraindicated in renal patients because it is excreted by urine)
- 3) Are there any better medications for this patient based on his/her:
 - Patient's life style (e.g. one dose vs. multiple doses)
 - Patient's financial state (e.g. use of generic vs. brand medications)
 - Patient's safety (e.g. anti-depressive medication increase risk of suicide in teenager)

APPENDIX G PERFORMANCE CRITERIA GLOSSARY

Excellent performance

A student, who requires minimal clinical supervision (beyond entry level) with simple to highly complex patients, is able to function in unfamiliar or ambiguous situations. At this level, the student's clinical reasoning is consistently proficient and shows knowledge and skills beyond entry level. The student's performance is timely and efficient. The student willingly assumes a leadership role for managing more difficult cases and is able to serve as a resource for others. Actively contributes to the enhancement of the pharmacy with an expansive view of the profession.

Entry-level performance

A student, who requires the expected degree of supervision for an entry level pharmacist and shows entry level knowledge and skills. Consults with others and resolves unfamiliar or ambiguous situations. At this level, the student's clinical reasoning is consistently proficient in simple and complex cases. The student performs in a timely and efficient manner.

Deficient (Novice) performance

A student who requires more than entry level clinical supervision even with simple patients. At this level, clinical performance is inconsistent and inefficient. Performance reflects little or no experience. Performance is slow and inefficient.

Complexity: The number of elements (e.g., simple or complex) that must be considered relative to the patient, task, and environment. As a student progresses through APPE's, the level of complexity of patients, tasks, and the environment should increase, with fewer elements under direct control by the preceptor.

Consistency: The frequency of occurrences of desired behaviors (e.g., infrequently, occasionally, and routinely). As a student progresses through APPE's, consistency of quality performance is expected to progress from infrequently to routinely.

Efficiency: The ability to perform in an effective and timely manner (e.g., inefficient/slow, efficient/timely). As the student progresses through APPE's, efficiency should progress from a high expenditure of time and effort to efficient and timely performance.

Performance criterion: A description of the behaviors (e.g. checklist and rating scale) that define the expected performance of students. When criteria are taken in aggregate, they describe the expected performance of the graduate upon entry into the practice of pharmacy.

Quality: The degree of knowledge and skill proficiency demonstrated. (e.g. limited skill, high skill). As a student progresses through clinical education experiences, quality should range from demonstration of limited skill to a skilled or highly skilled performance.

Supervision: The level and extent of assistance required by the student to achieve entry-level performance. As a student progresses through APPE's, the degree of supervision needed is expected to progress from close supervision to being capable of independent performance. This may vary with the complexity of the patient or environment.

APPENDIX H
DELPHI PANEL (ROUND ONE) RESULTS

Table H-1. Responses for Skill A

	Percent	Count
Yes (relevant)	100%	22
No (important but not relevant)	0%	0
No (not relevant)	0%	0

Table H-2. Responses for Skill B

	Percent	Count
Yes (relevant)	100%	22
No (important but not relevant)	0%	0
No (not relevant)	0%	0

Table H-3. Responses for Skill C

	Percent	Count
Yes (relevant)	86%	19
No (important but not relevant)	14%	3
No (not relevant)	0%	0

Table H-4. Responses for Skill D

	Percent	Count
Yes (relevant)	86%	19
No (important but not relevant)	9%	2
No (not relevant)	5%	1

Table H-5. Responses for Skill E

	Percent	Count
Yes (relevant)	86%	18
No (important but not relevant)	10%	2
No (not relevant)	5%	1

Table H-6. Responses for Skill F

	Percent	Count
Yes (relevant)	68%	15
No (important but not relevant)	32%	7
No (not relevant)	0%	0

Table H-7. Responses for Skill G

	Percent	Count
Yes (relevant)	91%	20
No (important but not relevant)	9%	2
No (not relevant)	0%	0

Table H-8. Responses to question 8

	Percent	Count
Yes	45%	10
No	55%	12

Table H-9. Responses to question A1

	Percent	Count
Yes (relevant)	100%	22
No (important but not relevant)	0%	0
No (not relevant)	0%	0

Table H-10. Responses to question A2

	Percent	Count
Yes (relevant)	100%	22
No (important but not relevant)	0%	0
No (not relevant)	0%	0

Table H-11. Responses to question B1

	Percent	Count
Yes (all are relevant)	81%	17
No (needs revisions)	19%	4
No (not relevant)	0%	0

Table H-12. Responses to question B2

	Percent	Count
Accept (representative)	67%	14
Reject (needs revisions)	33%	7
Reject (not representative)	0%	0

Table H.-13. Responses to question C1

	Percent	Count
Yes (all are relevant)	86%	18
No (needs revisions)	14%	3
No (not relevant)	0%	0

Table H-14. Responses to question C2

	Percent	Count
Accept (representative)	76%	16
Reject (needs revisions)	19%	4
Reject (not representative)	5%	1

Table H-15. Responses to question D1

	Percent	Count
Yes (all are relevant)	62%	13
No (needs revisions)	38%	8
No (not relevant)	0%	0

Table H-16. Responses to question D2

	Percent	Count
Accept (representative)	57%	12
Reject (needs revisions)	38%	8
Reject (not representative)	5%	1

Table H-17. Responses to question E1

	Percent	Count
Yes (all are relevant)	71%	15
No (needs revisions)	29%	6
No (not relevant)	0%	0

Table H-18. Responses to question E2

	Percent	Count
Accept (representative)	62%	13
Reject (needs revisions)	38%	8
Reject (not representative)	0%	0

Table H-19. Responses to question F1

	Percent	Count
Yes (all are relevant)	76%	16
No (needs revisions)	24%	5
No (not relevant)	0%	0

Table H-20. Responses to question F2

	Percent	Count
Accept (representative)	76%	16
Reject (needs revisions)	24%	5
Reject (not representative)	0%	0

Table H-21. Responses to question G1

	Percent	Count
Yes (all are relevant)	62%	13
No (needs revisions)	38%	8
No (not relevant)	0%	0

Table H-22. Responses to question G2

	Percent	Count
Accept (representative)	67%	14
Reject (needs revisions)	33%	7
Reject (not representative)	0%	0

Table H-23. Responses to complexity question

	Percent	Count
Accept as written	76%	16
Needs revision	24%	5
Not helpful	0%	0

Table H-24. Responses to reliability question

	Percent	Count
Accept as written	90%	19
Needs revision	10%	2
Not helpful	0%	0

Table H-25. Responses to deficient performance question

	Percent	Count
Accept as written	62%	13
Needs revision	38%	8
Not helpful	0%	0

Table H-26. Responses to efficiency question

	Percent	Count
Accept as written	95%	20
Needs revision	5%	1
Not helpful	0%	0

Table H-27. Responses to entry-level performance question

	Percent	Count
Accept as written	76%	16
Needs revision	19%	4
Not helpful	5%	1

Table H-28. Responses to excellent performance question

	Percent	Count
Accept as written	76%	16
Needs revision	24%	5
Not helpful	0%	0

Table H-29. Responses to performance criterion question

	Percent	Count
Accept as written	100%	21
Needs revision	0%	0
Not helpful	0%	0

Table H-30. Responses to quality question

	Percent	Count
Accept as written	81%	17
Needs revision	10%	2
Not helpful	10%	2

Table H-31. Responses to supervision question

	Percent	Count
Accept as written	95%	20
Needs revision	5%	1
Not helpful	0%	0

APPENDIX I
DELPHI PANEL (ROUND TWO) RESULTS

Table I-1. Responses to question A1

	Percent	Count
Do you accept as written (71% of the Delphi panel accepted the check list and performance scale as written)	66%	44
Do you recommend these revisions (19% of the Delphi panel recommended the rating scale reflect a greater emphasis on the quality of data collected from the student-patient encounter and recognition of the patient's chief complaint.)	33%	22
Do you reject (not representative)	1%	1

Table I-2. Response to question B1

	Percent	Count
Do you accept as written (71% of the Delphi panel accepted the check list and performance scale as written)	79%	53
Do you recommend revisions (19% of the Delphi panel recommended greater guidance on how to use determine appropriate caseload levels in the rating scale.)	19%	13
Do you reject (not representative)	1%	1

Table I-3. Responses to question D1

	Percent	Count
Do you accept as written (62% of the Delphi panel accepted the check list and performance scale as written)	23%	15
Do you recommend this revision (19% of the Delphi panel recommended merging skill D and E into a single skill for assessment. The new combined skill statement could read as: Designs and evaluates treatment regimens for optimal outcomes using pharmacokinetic and pharmacodynamic data. The drug regimen would consider disease states, previous or current drug therapy, drug formulation data, psycho-social, ethical-legal, and financial factors.)	55%	36
Do you also recommend this addition (5% of the Delphi panel recommended including "Severity of Illness" into the checklist.)	20%	13
Do you also recommend this addition (5% of the Delphi panel recommended including pharmacodynamic data into the skill statement.)	3%	2
Do you reject (not relevant)	0%	0

Table I-4. Collapsed responses to question D1

	Percent	Count
Do you accept as written (62% of the Delphi panel accepted the check list and performance scale as written)	23%	15
Do you recommend this revision (19% of the Delphi panel recommended merging skill D and E into a single skill for assessment.	77%	51

Table I-5. Responses to question D2

	Percent	Count
Do you accept as written (57% of the Delphi panel accepted the check list and performance scale as written)	47%	31
Do you recommend revisions (19% of the Delphi panel recommended removing the last three items in the Excellent rating scale {full-time caseload/supervising others/beyond expected of entry-level practitioner}.)	52%	34
Do you reject (not representative)	2%	1

Table I-6. Responses to question E1

	Percent	Count
Do you accept as written (71% of the Delphi panel accepted the check list and performance scale as written)	18%	12
Do you recommend this revision (19% of the Delphi panel recommended merging skill D and E into a single skill for assessment. The new combined skill statement could read as: Designs and evaluates treatment regimens for optimal outcomes using pharmacokinetic and pharmacodynamic data. The drug regimen would consider disease states, previous or current drug therapy, drug formulation data, psycho-social, ethical-legal, and financial factors.)	80%	53
Do you reject (not relevant)	2%	1

Table I-7. Responses to question E2

	Percent	Count
Do you accept as written (62% of the Delphi panel accepted the check list and performance scale as written)	47%	31
Do you recommend revisions (19% of the Delphi panel recommended removing the last three items in the Excellent rating scale {full-time caseload/supervising others/beyond expected of entry-level practitioner}.	50%	33
Do you reject (not representative)	3%	2

Table I-8. Responses to question G1

	Percent	Count
Do you accept as written (62% of the Delphi panel accepted the check list and performance scale as written)	36%	24
Do you recommend these revisions (33% of the Delphi panel recommended changing the checklist to include items appropriate for assessing student documentation with other healthcare professionals in addition to the patient.)	64%	42
Do you reject (not relevant)	0%	0

Table I-9. Responses to question G2

	Percent	Count
Do you accept as written (62% of the Delphi panel accepted the check list and performance scale as written)	50%	33
Do you recommend these revisions (19% of the Delphi panel recommended removing the last three items in the Excellent rating scale {full-time caseload/supervising others/beyond expected of entry-level practitioner}.)	48%	32
Do you reject (not representative)	2%	1

Table I-10. Responses to complexity question

	Percent	Count
Do you accept as written (76% of the Delphi panel accepted the check list and performance scale as written)	60%	40
Do you recommend revisions (24% of the Delphi panel recommended examples to help preceptors use complexity in assessment.) panel accepted the check list and performance scale as written)	39%	26
Not helpful and you do not recommend	1%	1

Table I-11. Responses to deficient performance question

	Percent	Count
Do you accept as written (62% of the Delphi panel accepted the check list and performance scale as written)	33%	22
Do you recommend these revisions (38% of the Delphi panel recommended these revisions to Deficient performance: A student who requires close clinical supervision with constant monitoring. At this level, clinical performance is inconsistent and inefficient. Performance reflects little or no experience. The student is not capable of carrying a reasonable caseload.)	65%	43
Not helpful and you do not recommend	2%	1

Table I-12. Responses to excellent performance question

	Percent	Count
Do you accept as written (76% of the Delphi panel accepted the check list and performance scale as written)	36%	24
Do you recommend these revisions (24% of the Delphi panel recommended these revisions: A student who requires minimal clinical supervision with simple to highly complex patients, and is able to function in unfamiliar or ambiguous situations. At this level, the student's clinical reasoning is consistently proficient and is capable of serving as a resource for others. The student is capable maintaining a reasonable caseload. The student willingly assumes a leadership role for managing more difficult cases. Actively contributes to the enhancement of the pharmacy with an expansive view of the profession.)	64%	42
Not helpful and you do not recommend	0%	0

APPENDIX J CASE STUDY SUMMARIES

Diabetes Case Study: Mrs. Davis is a 65-year-old retired nurse. She has a history of hypertension, occasional angina and type 2 diabetes mellitus. On examination her blood pressure is 135/80 mmHg, heart rate regular 65 beats/minute, weight 150 lbs, height 5 feet 6 inches, BMI 23.5, and waist circumference 30 inches.

Mrs. Davis is a non-smoker and consumes 2 to 3 small glasses of wine per week. She plays golf once a week and walks for one hour each day with her husband. Mrs. Davis reports that she is careful about the type and amount of food she eats and watches her intake by carbohydrate counting.

Over the last six months, Mrs. Davis' self-monitored random daily non-fasting blood glucose levels have been slowly increasing and over the last two weeks have been between 198 and 234 mg/dl. The ideal range for non-fasting blood glucose is 75 - 110 mg/dl. Six months ago, her glycated hemoglobin (HbA1c) measurement was 8.5% and 2 weeks ago, it was 9%. Three months ago a timed overnight urine collection demonstrated microalbuminuria at 72mg/day. The ideal range should be 30-300 mg/day. She has no evidence of retinopathy or neuropathy.

Her medication regimen has not been altered in the last 12 months. Mrs. Davis' current medications are: aspirin 81 mg daily, perindopril 4 mg daily, metoprolol tartrate 25 mg twice daily, sublingual nitrate prn (on average once per month), simvastatin 20 mg daily, metformin 850 mg three times a day, and glyburide 2.5 mg twice daily with food. She has no known allergies.

Heart Failure Case Study: Mr. Johnson is a 72-year old man with recently diagnosed heart failure. He has a history of ischemic heart disease and occasionally suffers from angina. His blood pressure readings consistently range between 132-142/86–94 mmHg. An echocardiogram eight weeks ago showed Mr. Johnson had a left ventricular ejection fraction of 30% and no valvular abnormalities. At that time, he was started on lisinopril 2.5 mg once daily and furosemide 40 mg once daily. His Primary Care Physician has been gradually increasing the lisinopril dose, aiming for a dose that will better manage his hypertension up to a maximal 40 mg /day dose.

Three weeks ago, his serum creatinine (SCr) and potassium (K) levels were normal and his physician increased his lisinopril from 5 mg to 10 mg once daily. Mr. Johnson's Chem-7, yesterday and two days ago, showed abnormal results: SCr 1.9 mg/dL and K 5.7 mEq/L. The normal SCr range for an adult is 0.7–1.3 mg/dL and the normal K range for an adult is 3.8–4.9 mEq/L. All other Chem-7 results were normal.

Mr. Johnson's other medications are aspirin 81 mg once daily and 0.4 mg sublingual nitrate PRN and uses it once every 4 to 8 weeks. Mr. Johnson feels well and is asymptomatic. He has no other significant medical history. Mr. Johnson lives by himself and is active and independent. He is not overweight, does not drink alcohol, and gave up smoking last year.

Anticoagulation Case Study: Mr. Williams is a 65-year-old man with atrial fibrillation, has been on warfarin for the past 12 months after he presented to the local emergency department with signs of a TIA. A head CT scan and trans-esophageal echocardiogram done at the time were normal. He has been well since. Mr. Williams reports that the most recent INR, measured this morning, was 4.6. Up until now, his INR results, which have been measured monthly, have been stable and in the range of 2.0-3.0.

Mr. Williams also has hypertension and osteoarthritis. He had a left total hip replacement 6 months ago. His current medications are atenolol 50 mg once daily, ramipril 10 mg once daily, amiodarone 200 mg daily, and warfarin 6 mg at night. Mr. Williams' primary care physician has admitted him to the hospital today, requested consultation with your anticoagulation service, and wants your recommendation on the next warfarin dose.

APPENDIX K
EXPERT PANEL SCRIPT TARGETS AND RESULTS

Table K-1. Diabetes performance targets from the script

<i>Skill</i>	<i>Mary</i>	<i>Thomas</i>	<i>Susan</i>
Skill A	Excellent	Entry-Level	Deficient
Skill B	Excellent	Entry-Level	Deficient
Skill D/E	Excellent	Entry-Level	Deficient

Table K-2. Diabetes assessments by expert panel

	<i>Excellent</i>	<i>Entry-Level</i>	<i>Deficient</i>
<i>Mary</i>			
Skill A	7 (100%)	0 (0%)	0 (0%)
Skill B	7 (100%)	0 (0%)	0 (0%)
Skill D/E	7 (100%)	0 (0%)	0 (0%)
<i>Thomas</i>			
Skill A	3 (43%)	4 (57%)	0 (0%)
Skill B	3 (43%)	4 (57%)	0 (0%)
Skill D/E	2 (29%)	5 (71%)	0 (0%)
<i>Susan</i>			
Skill A	0 (0%)	0 (0%)	7 (100%)
Skill B	0 (0%)	0 (0%)	7 (100%)
Skill D/E	0 (0%)	0 (0%)	7 (100%)

Table K-3. Heart Failure performance targets by the script

<i>Skill</i>	<i>Patricia</i>	<i>Joseph</i>	<i>Dorothy</i>
Skill A	Entry-Level	Excellent	Deficient
Skill B	Entry-Level	Deficient	Excellent
Skill D/E	Deficient	Excellent	Deficient

Table K-4. Heart Failure assessments by expert panel

	<i>Excellent</i>	<i>Entry-Level</i>	<i>Deficient</i>
<i>Patricia</i>			
Skill A	1 (14%)	2 (29%)	4 (57%)
Skill B	0 (0%)	3 (43%)	4 (57%)
Skill D/E	0 (0%)	0 (0%)	7 (100%)
<i>Joseph</i>			
Skill A	2 (29%)	3 (43%)	2 (29%)
Skill B	0 (0%)	2 (29%)	5 (71%)
Skill D/E	2 (29%)	3 (43%)	2 (29%)
<i>Dorothy</i>			
Skill A	0 (0%)	1 (14%)	6 (86%)
Skill B	2 (29%)	0 (0%)	5 (71%)
Skill D/E	1 (14%)	1 (14%)	5 (71%)

Table K-5. Anticoagulation performance targets by the script

<i>Skill</i>	<i>Linda</i>	<i>David</i>	<i>Barbara</i>
Skill A	Excellent	Entry-Level	Deficient
Skill B	Entry-Level	Excellent	Deficient
Skill D/E	Excellent	Deficient	Excellent

Table K-6. Anticoagulation assessments by expert panel

	<i>Excellent</i>	<i>Entry-Level</i>	<i>Deficient</i>
<i>Linda</i>			
Skill A	6 (86%)	0 (0%)	1 (14%)
Skill B	3 (43%)	2 (29%)	2 (29%)
Skill D/E	3 (43%)	2 (29%)	2 (29%)
<i>David</i>			
Skill A	1 (14%)	2 (29%)	4 (57%)
Skill B	2 (29%)	4 (57%)	1 (14%)
Skill D/E	0 (0%)	1 (14%)	6 (86%)
<i>Barbara</i>			
Skill A	0 (0%)	1 (14%)	6 (86%)
Skill B	0 (0%)	0 (0%)	7 (100%)
Skill D/E	0 (0%)	2 (29%)	5 (71%)

APPENDIX L
PRECEPTOR PANEL RESULTS

Table L-1. Diabetes assessments by preceptor panel

	<i>Excellent</i>	<i>Entry-Level</i>	<i>Deficient</i>
<i>Mary</i>			
Skill A	27 (64%)	10 (24%)	5 (12%)
Skill B	28 (67%)	9 (21%)	5 (12%)
Skill D/E	29 (69%)	9 (21%)	4 (10%)
<i>Thomas</i>			
Skill A	9 (21%)	22 (52%)	11 (26%)
Skill B	8 (64%)	23 (55%)	11 (26%)
Skill D/E	8 (19%)	19 (45%)	15 (36%)
<i>Susan</i>			
Skill A	0 (0%)	1 (02%)	41 (98%)
Skill B	0 (0%)	2 (05%)	40 (95%)
Skill D/E	0 (0%)	0 (0%)	42 (100%)

Table L-2. Heart Failure assessments by preceptor panel

	<i>Excellent</i>	<i>Entry-Level</i>	<i>Deficient</i>
<i>Patricia</i>			
Skill A	4 (10%)	15 (36%)	23 (55%)
Skill B	4 (10%)	11 (26%)	27 (64%)
Skill D/E	1 (02%)	17 (40%)	24 (57%)
<i>Joseph</i>			
Skill A	9 (21%)	24 (57%)	9 (21%)
Skill B	6 (14%)	15 (36%)	21 (50%)
Skill D/E	9 (21%)	21 (50%)	12 (29%)
<i>Dorothy</i>			
Skill A	2 (05%)	15 (36%)	25 (60%)
Skill B	4 (10%)	15 (36%)	23 (55%)
Skill D/E	1 (02%)	21 (50%)	20 (48%)

Table L-3. Anticoagulation assessments by preceptor panel

	<i>Excellent</i>	<i>Entry-Level</i>	<i>Deficient</i>
<i>Linda</i>			
Skill A	25 (60%)	13 (31%)	4 (10%)
Skill B	19 (45%)	18 (43%)	5 (12%)
Skill D/E	17 (40%)	18 (43%)	7 (17%)
<i>David</i>			
Skill A	11 (26%)	18 (43%)	13 (31%)
Skill B	11 (26%)	17 (40%)	14 (33%)
Skill D/E	5 (12%)	8 (19%)	29 (69%)
<i>Barbara</i>			
Skill A	0 (0%)	8 (19%)	34 (81%)
Skill B	2 (05%)	8 (19%)	32 (76%)
Skill D/E	2 (05%)	13 (31%)	27 (64%)

APPENDIX M
PRECEPTER RELIABILITY RESULTS

Table M-1. Reliability – competent vs. not-competent

<i>Skill</i> <i>Case per Skill</i>	<i>ICC</i>	<i>Lower</i> <i>(95% CI)</i>	<i>Upper</i> <i>(95% CI)</i>	<i>p-value</i>
Skill A	0.37	0.20	0.69	0.00 *
Diabetes	0.66	0.33	0.99	0.00 *
Heart Failure	0.15	0.03	0.88	0.00 *
Anticoagulation	0.46	0.17	0.97	0.00 *
Skill B	0.31	0.16	0.63	0.00 *
Diabetes	0.63	0.30	0.99	0.00 *
Heart Failure	0.00	-0.02	0.45	0.37
Anticoagulation	0.37	0.12	0.96	0.00 *
Skill D/E	0.30	0.15	0.62	0.00 *
Diabetes	0.67	0.34	0.99	0.00 *
Heart Failure	0.07	0.01	0.78	0.01 *
Anticoagulation	0.29	0.08	0.94	0.00 *

* Significant

Table M-2. Reliability – excellent vs. entry-level

<i>Skill</i> <i>Case per Skill</i>	<i>ICC</i>	<i>Lower</i> <i>(95% CI)</i>	<i>Upper</i> <i>(95% CI)</i>	<i>p-value</i>
Skill A	0.24	0.11	0.55	0.00 *
Diabetes	0.54	0.22	0.98	0.00 *
Heart Failure	0.10	0.01	0.84	0.00 *
Anticoagulation	0.26	0.07	0.93	0.00 *
Skill B	0.19	0.09	0.49	0.00 *
Diabetes	0.49	0.19	0.98	0.00 *
Heart Failure	0.00	-0.02	0.42	0.40
Anticoagulation	0.24	0.06	0.93	0.00 *
Skill D/E	0.19	0.09	0.48	0.00 *
Diabetes	0.52	0.21	0.98	0.00 *
Heart Failure	0.02	-0.01	0.61	0.12
Anticoagulation	0.18	0.04	0.90	0.00 *

* Significant

Table M-3. Reliability of scale items – Skill A

- Case - Rating Scale	ICC	Lower (95% CI)	Upper (95% CI)	p-value
Diabetes				
Supervision	0.65	0.33	0.99	0.00 *
Quality	0.76	0.46	0.92	0.00 *
Complexity	0.63	0.30	0.99	0.00 *
Consistency	0.62	0.29	0.99	0.00 *
Efficiency	0.67	0.34	0.98	0.00 *
Heart Failure				
Supervision	0.16	0.03	0.89	0.00 *
Quality	0.12	0.02	0.86	0.00 *
Complexity	0.11	0.02	0.85	0.00 *
Consistency	0.10	0.02	0.84	0.00 *
Efficiency	0.14	0.02	0.89	0.00 *
Anticoagulation				
Supervision	0.56	0.24	0.98	0.00 *
Quality	0.55	0.23	0.98	0.00 *
Complexity	0.45	0.16	0.97	0.00 *
Consistency	0.53	0.22	0.98	0.00 *
Efficiency	0.32	0.10	0.95	0.00 *

* Significant

Table M-4. Reliability of scale items – Skill B

- Case - Rating Scale	ICC	Lower (95% CI)	Upper (95% CI)	p-value
Diabetes				
Supervision	0.64	0.32	0.99	0.00 *
Quality	0.72	0.40	0.99	0.00 *
Complexity	0.65	0.32	0.99	0.00 *
Consistency	0.67	0.34	0.99	0.00 *
Efficiency	0.75	0.44	0.99	0.00 *
Heart Failure				
Supervision	0.03	-0.01	0.69	0.09
Quality	0.03	-0.01	0.68	0.09
Complexity	0.00	-0.01	0.47	0.32
Consistency	0.04	0.00	0.71	0.03 *
Efficiency	-0.01	-0.02	0.33	0.54
Anticoagulation				
Supervision	0.38	0.13	0.96	0.00 *
Quality	0.34	0.11	0.96	0.00 *
Complexity	0.36	0.11	0.96	0.00 *
Consistency	0.39	0.13	0.96	0.00 *
Efficiency	0.24	0.06	0.93	0.00 *

* Significant

Table M-5. Reliability of scale items – Skill D/E

- Case - Rating Scale	ICC	Lower (95% CI)	Upper (95% CI)	p-value
Diabetes				
Supervision	0.72	0.39	0.99	0.00 *
Quality	0.78	0.48	0.99	0.00 *
Complexity	0.71	0.39	0.99	0.00 *
Consistency	0.74	0.42	0.99	0.00 *
Efficiency	0.83	0.55	0.99	0.00 *
Heart Failure				
Supervision	0.15	0.03	0.88	0.09
Quality	0.24	0.06	0.93	0.09
Complexity	0.10	0.01	0.83	0.32
Consistency	0.19	0.04	0.91	0.03 *
Efficiency	0.12	0.02	0.86	0.54
Anticoagulation				
Supervision	0.35	0.11	0.96	0.00 *
Quality	0.32	0.10	0.95	0.00 *
Complexity	0.19	0.04	0.91	0.00 *
Consistency	0.28	0.08	0.94	0.00 *
Efficiency	0.18	0.04	0.90	0.00 *

* Significant

Table M-6. Reliability of global assessment – competent vs. not-competent

Skill Case per Skill	ICC	Lower (95% CI)	Upper (95% CI)	p-value
Skill A				
Diabetes	0.78	0.48	0.99	0.00 *
Heart Failure	0.25	0.07	0.93	0.00 *
Anticoagulation	0.57	0.24	0.98	0.00 *
Skill B				
Diabetes	0.65	0.32	0.99	0.00 *
Heart Failure	0.03	-0.01	0.68	0.13
Anticoagulation	0.35	0.11	0.96	0.00 *
Skill D/E				
Diabetes	0.77	0.47	0.99	0.00 *
Heart Failure	0.26	0.07	0.94	0.00 *
Anticoagulation	0.32	0.10	0.95	0.00 *

* Significant

APPENDIX N
PRECEPTOR ACCURACY RESULTS

Table N-1. Accuracy – Skill A

Skill A	Competent	Excellent
- Case study	vs.	vs.
- Student	Not-competent	Entry-level
	p-value	p-value
Diabetes cases		
Mary	0.45	0.14 *
Thomas	0.15 *	0.36
Susan	0.86	N/A ¹
Heart Failure cases		
Patricia	0.62	0.56
Joseph	0.50	0.46
Dorothy	0.19 *	0.89
Anticoagulation cases		
Linda	0.55	0.10 **
David	0.18 *	0.69
Barbara	0.62	0.18 *

** Significant up to 10%, * significant up to 30%

1) One variable is a constant and association not computed

Table N-2. Accuracy – Skill B

Skill B	Competent	Excellent
- Case study	vs.	vs.
- Student	Not-competent	Entry-level
	p-value	p-value
Diabetes cases		
Mary	0.45	0.18 *
Thomas	0.15 *	0.32
Susan	0.73	N/A ¹
Heart Failure cases		
Patricia	0.51	0.45
Joseph	0.26	0.54
Dorothy	0.44	0.07 **
Anticoagulation cases		
Linda	0.26 *	0.55
David	0.30 *	0.58
Barbara	0.18 *	N/A ¹

** Significant up to 10%, * significant up to 30%

1) One variable is a constant and association not computed

Table N-3. Accuracy – Skill D/E

Skill D/E	Competent vs. Not-competent p-value	Excellent vs. Entry-level p-value
Diabetes cases		
Mary	0.53	0.18 *
Thomas	0.06 **	0.67
Susan	N/A ¹	N/A ¹
Heart Failure cases		
Patricia	0.03 **	N/A ¹
Joseph	0.66	0.51
Dorothy	0.23 *	0.16 *
Anticoagulation cases		
Linda	0.38	0.50
David	0.34	0.64
Barbara	0.54	0.77

** Significant up to 10%, * significant up to 30%

1) One variable is a constant and association not computed

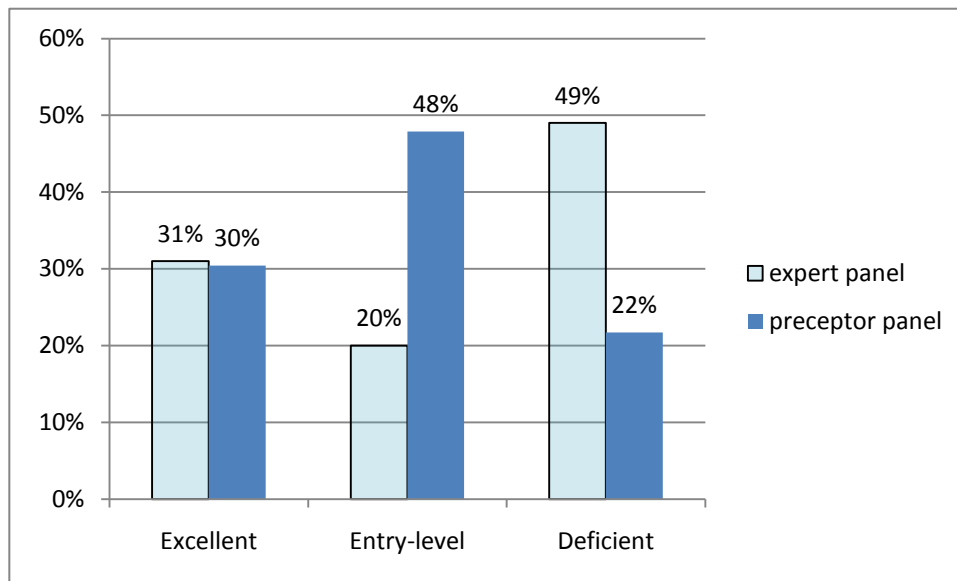


Figure N-1. Skill A – Comparing expert and preceptor panels

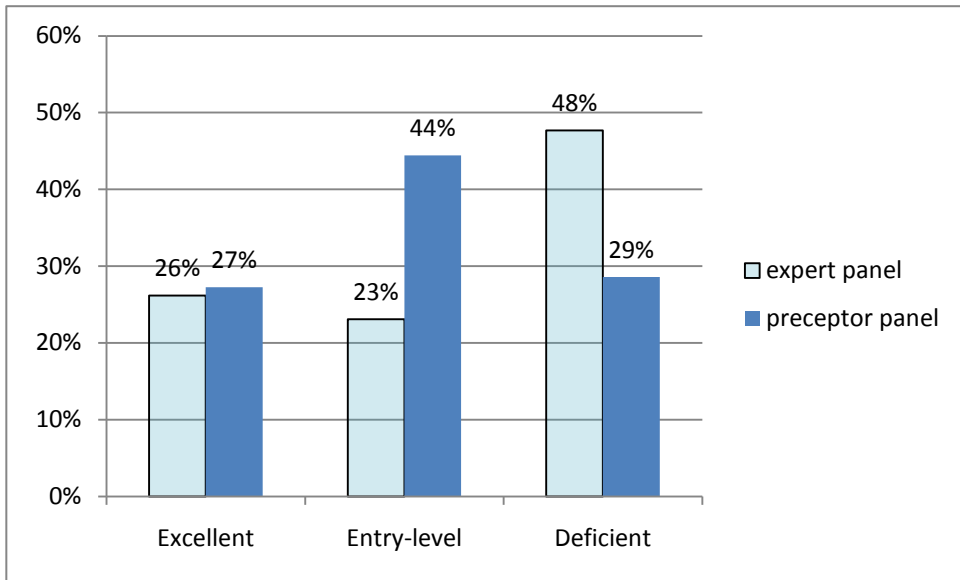


Figure N-2. Skill B – Comparing expert and preceptor panels

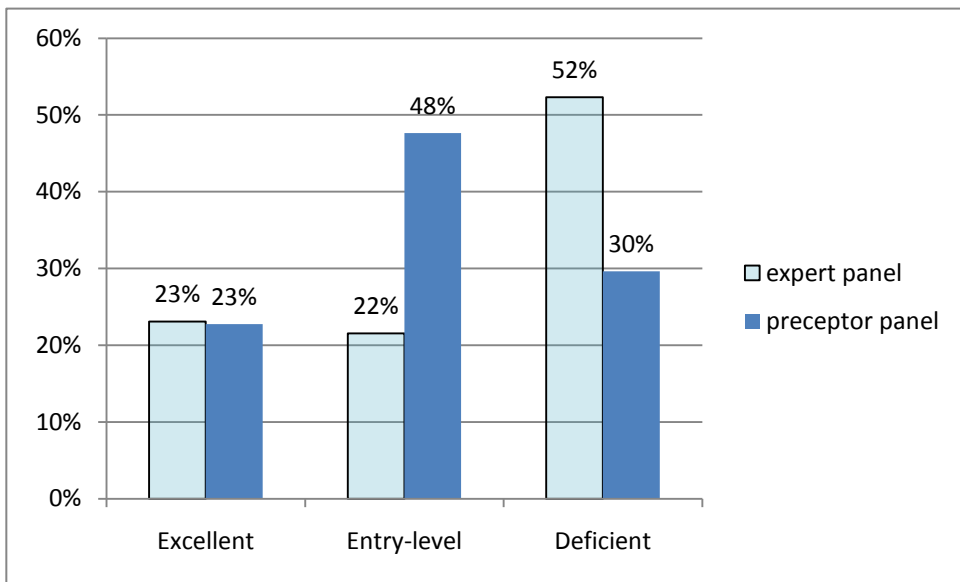


Figure N-3. Skill D/E- Comparing expert and preceptor panels

APPENDIX O
SUPERVISION PROXY RATING SCALE

The principal researcher created a surrogate scale by collapsing the supervision domain of the multifactorial scale into a three item rating scale. This supervision proxy scale approximates SUCCESS' three level rating scale.

1) Excellent performance:

- Supervision: Close supervision (constant monitoring)
 More than standard level of supervision (for entry level)
 Standard level of supervision (for entry level)
 Minimal supervision (beyond entry level)

2) Entry-level performance:

- Supervision: Close supervision (constant monitoring)
 More than standard level of supervision (for entry level)
 Standard level of supervision (for entry level)
 Minimal supervision (beyond entry level)

3) Deficient performance:

- Supervision: Close supervision (constant monitoring)
 More than standard level of supervision (for entry level)
 Standard level of supervision (for entry level)
 Minimal supervision (beyond entry level)

APPENDIX P
RATING SCALE COMPARISON

Table P-1. Skill A – Comparing rating scale with proxy scale

Skill A	Competent	Excellent
- Case study	vs.	vs.
- Student	Not-competent	Entry-level
	p-value	p-value
Diabetes cases		
Mary	0.10 **	0.50
Thomas	0.01 **	0.24 *
Susan	0.00 **	N/A ¹
Heart Failure cases		
Patricia	0.00 **	0.56
Joseph	0.00 **	0.35
Dorothy	0.01 **	0.64
Anticoagulation cases		
Linda	0.06 **	0.56
David	0.01 **	0.58
Barbara	0.01 **	N/A ¹

** Significant up to 10%, * significant up to 30%

1) One variable is a constant and association not computed

Table P-2. Skill B – Comparing rating scale with proxy scale

Skill B	Competent	Excellent
- Case study	vs.	vs.
- Student	Not-competent	Entry-level
	p-value	p-value
Diabetes cases		
Mary	0.22 *	0.58
Thomas	0.07 **	0.40
Susan	0.09 **	N/A ¹
Heart Failure cases		
Patricia	0.00 **	0.28 *
Joseph	0.01 **	0.36
Dorothy	0.01 **	0.49
Anticoagulation cases		
Linda	0.22 *	0.30 *
David	0.04 **	0.59
Barbara	0.02 **	0.41

** Significant up to 10%, * significant up to 30%

1) One variable is a constant and association not computed

Table P-3. Skill D/E – New rating scale vs. proxy supervision scale

Skill D/E	Competent	Excellent
- Case study	vs.	vs.
- Student	Not-competent	Entry-level
	p-value	p-value
Diabetes cases		
Mary	0.50	0.58
Thomas	0.26 *	0.00 **
Susan	0.12 *	N/A ¹
Heart Failure cases		
Patricia	0.04 **	0.27 *
Joseph	0.15 *	0.46
Dorothy	0.00 **	0.37
Anticoagulation cases		
Linda	0.03 **	0.42
David	0.01 **	0.22 *
Barbara	0.01 **	0.47

** Significant up to 10%, * significant up to 30%

1) One variable is a constant and association not computed

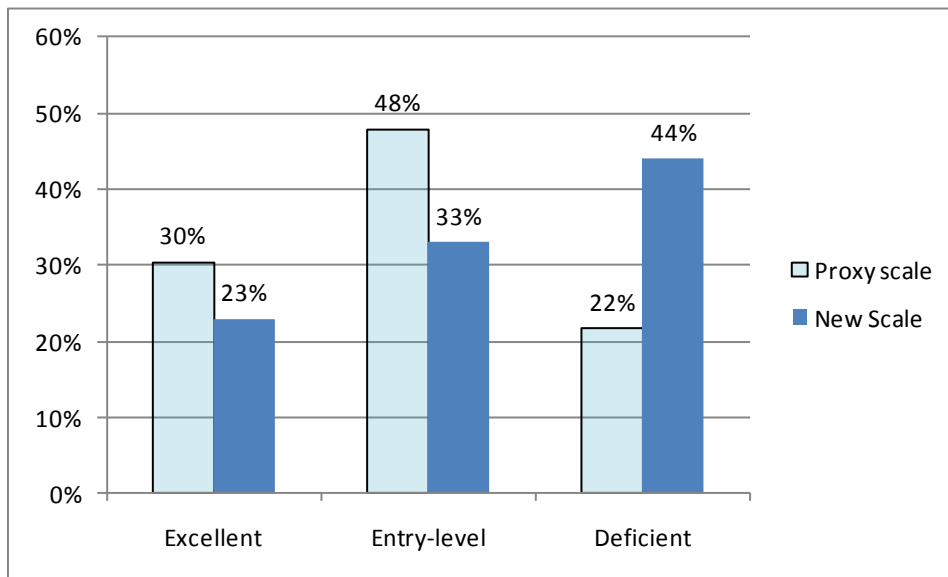


Figure P-1. Skill A – Comparing rating scale with proxy scale

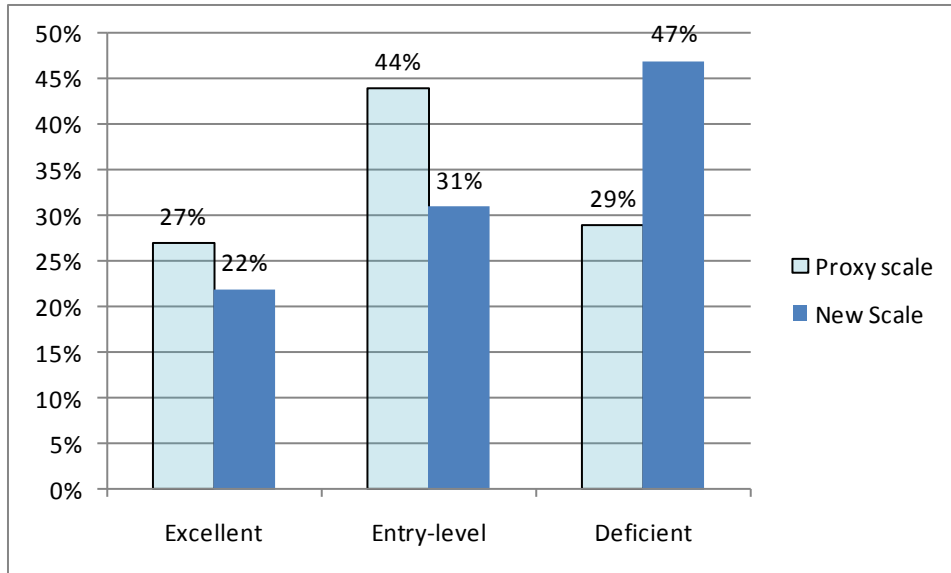


Figure P-2. Skill B – Comparing rating scale with proxy scale

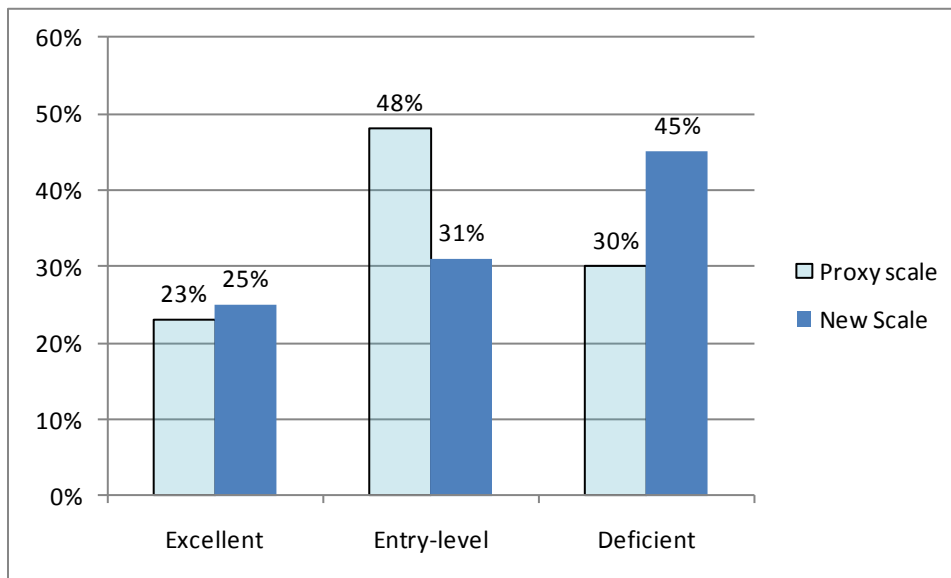


Figure P-3. Skill D/E – New rating scale vs. proxy supervision scale

LIST OF REFERENCES

1. Manasse J, Henri R., Speedie MK. Pharmacists, Pharmaceuticals, and Policy Issues Shaping the Work Force in Pharmacy. *Am J Health-Syst Pharm.* 2007;64(12):e30-e48.
2. IOM. *To Err is Human: Building a Safer Health System* Washington, DC: Institute of Medicine; 1999.
3. U.S. Bureau of Health Professions. *Report on Health Professional Accessibility.* Washington, DC 1996.
4. Knapp KK, Cultice JM. New pharmacist supply projections: lower separation rates and increased graduates boost supply estimates. *J Am Pharm Assoc (2003).* Jul-Aug 2007;47(4):463-470.
5. Bureau of Health Professions. *The Pharmacist Workforce: A Study of the Supply and Demand for Pharmacists.* 2000. Published Last Modified Date|. Accessed Dated Accessed|.
6. Leach DC. Building and Assessment Competence: The Potential for Evidence-based Graduate Medical Education. *Quality Management in Health Care.* 2002;11(1):39.
7. Lenburg C. The Framework, Concepts and Methods of the Competency Outcomes and Performance Assessment (COPA) Model. *Online Journal of Issues in Nursing.* 1999;4(2).
8. Richardson WC, ed. *Crossing the Quality Chasm: A New Health System for the 21st Century;* 2000.
9. Wennberg DE, Lucas FL, Birkmeyer JD, Bredenberg CE, Fisher ES. Variation in Carotid Endarterectomy Mortality in the Medicare Population: Trial Hospitals, Volume, and Patient Characteristics. *JAMA.* 1998;279(16):1278-1281.
10. O'Connor GT, Plume SK, Olmstead EM, et al. A regional intervention to improve the hospital mortality associated with coronary artery bypass graft surgery. The Northern New England Cardiovascular Disease Study Group. *JAMA.* March 20, 1996 1996;275(11):841-846.
11. Greiner AC, Knebel E, eds. *Health Professions Education: A Bridge to Quality;* 2003.
12. Harden RM. Developments in outcome-based education. *Medical Teacher,* 2002: 117-120.

13. Commission to Implement Change in Pharmaceutical Education. What is the mission of pharmaceutical education? Background paper I. . *American Journal of Pharmaceutical Education*. 1993;57:374-376.
14. Commission to Implement Change in Pharmaceutical Education. Entry-level education in pharmacy: Commitment to change Background Paper II. *American Journal of Pharmaceutical Education*. 1993;57:366-374.
15. Byrd G. Can the profession of pharmacy serve as a model for health informationist professionals? *Journal of Medical Library Association*. 2002;90(1):68-75.
16. CAPE Advisory Panel on Educational Outcomes. CAPE Educational Outcomes. http://www.aacp.org/Docs/MainNavigation/Resources/6075_CAPE2004.pdf. Accessed November 7, 2008, 2008.
17. Zlatic T. Redefining a profession: assessment in pharmacy education. In: Palomba CA BT, ed. *Assessing Student Competence in Accredited Disciplines: Pioneering Approaches to Assessment in Higher Education*. Sterling, VA: Stylus Publishing; 2001:49–70.
18. Schuwirth LWT, Southgate L, Page GG, et al. When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *MEDICAL EDUCATION*. 2002;36(10):925-930.
19. McAllister A. *Competency Based Assessment of Speech Pathology Students' Performance in the Workplace*. Sydney: School of Communication Sciences and Disorders, The University of Sydney; 2005.
20. Rethans JJ, Norcini JJ, Baro´n-Maldonado M, et al. The relationship between competence and performance: implications for assessing practice performance. 2002.
21. Raehl CL. AACP Pharmacy Education Assessment Services: Outcomes, Assessment, Accountability. *Journal of Pharmaceutical Education*. 2008;72(1).
22. CAPE Advisory Panel on Educational Outcomes. Utilization of the Center for the Advancement of Pharmaceutical Education Educational Outcomes, Revised Version 2004: Report of the 2005 American College of Clinical Pharmacy Educational Affairs Committee. 2006.
23. Gonczi A. Competency based assessment in the professions in Australia. *Assessment in Education: Principles, Policy & Practice*. 1994;1(1):27.
24. Kaslow NJ. Competencies in professional psychology. *American Psychologist*. 2004;59:774-781.

25. Marrelli T, & Hoge, 2005. Strategies for Developing Competency Models *Journal Administration and Policy in Mental Health and Mental Health Services Research* 2005;32(5-6).
26. Govaerts M, van der Vleuten C, Schuwirth L, Muijtjens A. Broadening Perspectives on Clinical Performance Assessment: Rethinking the Nature of In-training Assessment. *Advances in Health Sciences Education*. 2007/05/14/ 2007;12(2):239-260.
27. Clauser BE. Recurrent Issues and Recent Advances in Scoring Performance Assessments. *Applied Psychological Measurement*. December 1, 2000 2000;24(4):310-324.
28. Skrabal MZ, Jones RM, Nemire RE, et al. National Survey of Volunteer Pharmacy Preceptors. *American Journal of Pharmaceutical Education* 2008; 72 (5) Article 112. 2008.
29. Gandy JS. Personal communication with Director of Academic/Clinical Education Affairs at APTA; 2009.
30. Ried LD, Nemire, R., Doty, R., Brickler, M., Anderson, H., Frenzel-Shepherd, E., Larose-Pierre, M., and Dugan, D. An Automated Competency-based Student Performance Assessment Program for Advanced Pharmacy Practice Experiential Programs. *American Journal of Pharmaceutical Education*. 2007;71(6).
31. Messick S. Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice*. 1995;14(4):5-8.
32. Messick S. Validity of performance assessments. In: Phillips GW, ed. *Technical Issues in Large-Scale Performance Assessment*. Washington: National Centre for Education Statistics.; 1996:pp. 1-18.
33. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC 1999.
34. Haynes SN, Richard DCS, Kubany ES. Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. *Psychological Assessment*. 1995;7(3):238-247.
35. Nunnally JC. *Psychometric Theory*. 1978: McGraw-Hill Book Company; 1978.
36. Sireci SG. The construct of content validity. *Social Indicators Research*. Nov 1998;45(1-3):83-117.
37. Sireci SG. Gathering and analyzing content validity data. *Educational Assessment*. 1998;5(4):299-321.

38. Lynne S. Determining Content Validity of a Self-Report Instrument for Adolescents Using a Heterogeneous Expert Panel. *Nursing Research*. 2007;56(5):361-366.
39. Streiner DL, Norman GR. *Health Measurement Scales*. 2 ed. Oxford, England: Oxford University Press; 1995.
40. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *MEDICAL EDUCATION*. 2005;39(3):309-317.
41. Cross V, Hicks, C and Barwell, F. Comparing the Importance of Clinical Competence Criteria across Specialties Impact on undergraduate assessment. *Physiotherapy*. 2001;87(7):351-367.
42. Carr D. QUESTIONS OF COMPETENCE. *British Journal of Educational Studies*. 1993;41(3):253-271.
43. Dugan BD. Enhancing Community Pharmacy Through Advanced Pharmacy Practice Experiences. *American Journal of Pharmaceutical Education*. 2006;70(1):Article 21.
44. ACPE ACfPE. *Accreditation standards and guidelines for the professional program in pharmacy leading to the doctor of pharmacy degree 2009*.
45. Noyce P. Governance and the pharmaceutical workforce in England. *Res Social Adm Pharm*. Sep 2006;2(3):408-419.
46. Eraut M. *Developing professional knowledge and competence*. London: Falmer Press; 1994.
47. Cross V, Hicks C, Barwell F. Exploring the Gap Between Evidence and Judgement: using video vignettes for practice-based assessment of physiotherapy undergraduates. *Assessment & Evaluation in Higher Education*. 2001;26(3):189 - 212.
48. Hager P, Gonczi A. General issues about assessment of competence. *Assessment & Evaluation in Higher Education*. 1994/04// 1994;19(1):3.
49. Hager P, Gonczi A. What is competence? *Medical Teacher*. 1996/03// 1996;18(1):15.
50. Wolf A. *Competence-Based Assessment*. Buckingham: Open University Press.; 1995.
51. Thorndike RL, Hagen E. *Measurement and Evaluation in Psychology and Education*. 3 ed. New York: John Wiley and Sons, Inc; 1977.
52. Harris R, Guthrie H, Hobart B, Lundberg D. *Competency-Based Education and Training: Between a Rock and a Whirlpool*. Melbourne Macmillan Education; 1995.

53. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med.* Sep 1990;65(9 Suppl):S63-67.
54. Rethans JJ, Sturmans F, Drop R, van der Vleuten C. Assessment of the performance of general practitioners by the use of standardized (simulated) patients. *Br J Gen Pract.* Mar 1991;41(344):97-99.
55. Miller DM, Linn RL. Validation of Performance-Based Assessments. *Applied Psychological Measurement.* December 1, 2000 2000;24(4):367-378.
56. Andrich D. A framework relating outcomes based education and the taxonomy of educational objectives. *Studies In Educational Evaluation.* 2002;28(1):35-59.
57. Clark D. Learning Domains or Bloom's Taxonomy. 2007
<http://www.nwlink.com/~donclark/hrd/bloom.html>. Published Last Modified Date|. Accessed Dated Accessed|.
58. Krathwohl DR. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice.* 2002;41(4):212 - 218.
59. Epstein R, Hundert E. Defining and assessing professional competence. *JAMA: Journal of the American Medical Association.* 2002//09/2002 Jan 9 2002;287(2):226.
60. Kane MT. The Assessment of Professional Competence. *Eval Health Prof.* June 1, 1992 1992;15(2):163-182.
61. Worthen BR, Borg WR, White K. *Measurement and evaluation in the schools.* Longman Longman 1993.
62. Hays RB, Davies HA, Beard JD, et al. Selecting performance assessment methods for experienced physicians. *Med Educ.* Oct 2002;36(10):910-917.
63. Boud D. Sustainable Assessment: Rethinking Assessment for the Learning Society. *Studies in Continuing Education.* 2000;22(2).
64. Duke M. Clinical evaluation--difficulties experienced by sessional clinical teachers of nursing: a qualitative study. *Journal of Advanced Nursing.* 1996;23(2):408-414.
65. Sax G. *Principles of Education and Psychological Measurement and Evaluation (3rd ed.).* Belmont, CA: Wadsworth; 1997.
66. Bondy KN. Criterion-Referenced Definitions for Rating-Scales in Clinical Evaluation. *Journal of Nursing Education.* 1983;22(9):376-382.
67. Bondy KN. Clinical Evaluation of Student Performance: The Effects of Criteria on Accuracy and Reliability. *Research in Nursing & Health.* 1984;7(1):25-33.

68. Noel GL, Herbers JE, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How Well Do Internal Medicine Faculty Members Evaluate the Clinical Skills of Residents?. *Annals of Internal Medicine*. 1992;117(9):757.
69. Martin JA, Reznick RK, Rothman A, Tamblyn RM, Regehr G. Who should rate candidates in an objective structured clinical examination? *Acad Med*. Feb 1996;71(2):170-175.
70. Wray N, Friedland J. Detection and correction of house staff error in physical diagnosis. *JAMA*. 1983;249:1035-1037.
71. Holmboe ES, Hawkins RE. *Practical guide to the evaluation of clinical competence*. Philadelphia, PA: Mosby/Elsevier; 2008.
72. Norman G. Checklists vs. Ratings, the Illusion of Objectivity, the Demise of Skills and the Debasement of Evidence. *Advances in Health Sciences Education*. 2005;10(1):1-3.
73. Landy FJ, Farr JL. Performance rating. *Psychological Bulletin*. 1980;87(1):72-107.
74. Gomez-Mejia LR. Evaluating Employee Performance. *Journal of Organizational Behavior Management*. 1988;9(2):155 - 172.
75. Kingstrom PO, Bass AR. A CRITICAL ANALYSIS OF STUDIES COMPARING BEHAVIORALLY ANCHORED RATING SCALES (BARS) AND OTHER RATING FORMATS. *Personnel Psychology*. Summer81 1981;34(2):263-289.
76. Fay CH, Latham GP. Effects of training and rating scales on rating errors. *Personnel Psychology*. 1982;35:105-117.
77. Wolfe EW, Gitomer DH. The Influence of Changes in Assessment Design on the Psychometric Quality of Scores. *Applied Measurement in Education*. 2001;14(1):91-107.
78. Finn RH. Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement*. 1972;32:255-265.
79. Bendig AW. Reliability and number of rating scale categories. *Journal of Applied Psychology*. 1954;38:38-40.
80. Bendig AW. Reliability of short rating scales and the heterogeneity of the rated stimuli. *Journal of Applied Psychology*. 1954;38:167-170.
81. Jette DU, Bertoni A, Coots R, Johnson H, McLaughlin C, Weisbach C. Clinical Instructors' Perceptions of Behaviors That Comprise Entry-Level Clinical Performance in Physical Therapist Students: A Qualitative Study. *PHYS THER*. July 1, 2007 2007;87(7):833-843.

82. Cross V, Hicks C. What Do Clinical Educators Look for in Physiotherapy Students? *Physiotherapy*. 1997;83(5):249-260.
83. Alexander HA. Physiotherapy student clinical education: The influence of subjective judgements on observational. *Assessment & Evaluation in Higher Education*. 1996;21(4):357.
84. Chapman J. Agonising about assessment. In: Fish D, Coles C, eds. *Developing Professional Judgement in Health Care: Learning through the critical appreciation of practice*. Oxford: Butterworth-Heinemann; 1998:157-181.
85. Lankshear AJ. The use of focus groups in a study of attitudes to student nurse assessment.. *Journal of Advanced Nursing*. 1993/12// 1993;18(12):1986-1989.
86. Landy FJ, Guion RM. Development of scales for the measurement of work motivation. *Organizational Behavior and Human Performance*. 1970;5(1):95-103.
87. Davis MH, Harden RM. Competency-based assessment: making it a reality. *Medical Teacher*. 2003/11// 2003;25(6):565-568.
88. Wojtczak A. Medical education terminology. *Medical Teacher*, 2002: 357-357.
89. Gray. Global rating scales in residency education. *Acad Med*. 1996;71(1).
90. Halpern R, Lee MY, Boulter PR, Phillips RR. A synthesis of nine major reports on physicians competencies for the emerging practice environment. *Academic Medicine*. 2001;76(6):606-615.
91. Downing SM. Validity: on the meaningful interpretation of assessment data. *Medical Education*. 2003;37(9):830-837.
92. Clauser BE, Margolis MJ, Swanson DB. Issues of Validity and Reliability for Assessments in Medical Education. In: Holmboe ES, Hawkins RE, eds. *Practical guide to the evaluation of clinical competence*; 2008:10-23.
93. Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? A review of the published instruments. *J Gen Intern Med*. Sep 2004;19(9):971-977.
94. Messick S. In: Linn RL, ed. *Educational measurement*. Phoenix, Az: American Council on Education and Oryx Press; 1993:13-104.
95. Walsh WB, Bezt NE. *Tests and Assessment*. 4 ed. Upper Saddle River, NJ: Prentice Hall; 2000.
96. Ebel RL, Frisbie DA. *Essentials of Educational Measurement*. Englewood Cliffs, N.J.: Prentice Hall 1991.

97. Guion RM. Content Validity--The Source of My Discontent. *Applied Psychological Measurement*. 1977;1(1):1-10.
98. Lawshe CH. A QUANTITATIVE APPROACH TO CONTENT VALIDITY. *Personnel Psychology*. Winter75 1975;28(4):563-575.
99. Lynn M. Determination and quantification of content validity. *Nursing Research*. 1986;35(5).
100. Crocker L. Assessing Content Representativeness of Performance Assessment Exercises. *Applied Measurement in Education*. 1997;10(1):83 - 95.
101. Guion RM. CHANGING VIEW FOR PERSONNEL SELECTION RESEARCH. *Personnel Psychology*. Summer87 1987;40(2):199-213.
102. Stelly DJ, Goldstein HW. Application of Content Validation Methods to Broader Constructs. In: McPhail SM, ed. *Alternative validation strategies developing new and leveraging existing validity evidence*. San Francisco: Jossey-Bass; 2007.
103. Moss PA. Shifting Conceptions of Validity in Educational-Measurement - Implications for Performance Assessment. *Review of Educational Research*. Fal 1992;62(3):229-258.
104. Lennon RT. Assumptions Underlying the Use of Content Validity. *Educational and Psychological Measurement*. October 1, 1956 1956;16(3):294-304.
105. Downing SM. Threats to the validity of clinical teaching assessments: What about rater error? *Medical Education*. 2005;39(4):353-355.
106. Dauphinee WD. Assessing clinical performance: Where do we stand and what might we expect? . *The Journal of the American Medical Association*. 1995;274(9):741-743.
107. McDowell I. *The theoretical and technical foundations of health measurement* 1996.
108. Messick S, ed. *Validity*. 3 ed. New York: American Council on Education.; 1989. Linn RL, ed. *Educational measurement*.
109. Schuwirth LW, van der Vleuten CP. The use of clinical simulations in assessment. *Med Educ*. Nov 2003;37 Suppl 1:65-71.
110. Winters J, Hauck B, Riggs CJ, Clawson J, Collins J. Use of videotaping to assess competencies and course outcomes. *J Nurs Educ*. Oct 2003;42(10):472-476.
111. Crossley J, Humphris G, Jolly B. Assessing health professionals. *Medical Education*. 2002;36(9):800-804.

112. Carraccio C, Wolfsthal SD, Englander R, Ferentz K, Martin C. Shifting Paradigms: From Flexner to Competencies. *Academic Medicine*. 2002;77(5):361-367.
113. Schwabbauer M. But can they do it? Clinical competency assessment. *Clinical Laboratory Science*. 2000;13(1):47-52.
114. Whitcomb ME. Competency-based graduate medical education? Of course! But how should competency be assessed? *Acad Med*. May 2002;77(5):359-360.
115. SUCCESS. SUCCESS WEB site. <http://www.cop.ufl.edu/doty/success/help/>. Accessed December 15th, 2008, 2009.
116. Clayton MJ. DELPHI: A Technique to Harness Expert Opinion for Critical Decision-Making Tasks in Education. *Educational Psychology*. 1997/12// 1997;17(4):373.
117. Keeney S, Hasson F, McKenna HP. A critical review of the Delphi technique as a research methodology for nursing. *International Journal of Nursing Studies*. 2001/4 2001;38(2):195-200.
118. McKenna HP. The Delphi technique: a worthwhile research approach for nursing? *Journal of Advanced Nursing*. 1994;19(6):1221-1225.
119. Powell C. The Delphi technique: myths and realities. *Journal of Advanced Nursing*. 2003;41(4):376-382.
120. Williams PL, Webb C. The Delphi technique: a methodological discussion. *Journal of Advanced Nursing*. 1994/01// 1994;19(1):180-186.
121. Waltz CF, Strickland O, Lenz ER. *Measurement in Nursing Research*. Philadelphia: F.A. Davis Co; 1991.
122. Grant JS, Davis LL. Selection and use of content experts for instrument development. *Research in Nursing & Health*. 1997;20(3):269-274.
123. Dillman DA, Smyth JD, Christian LM. *Internet, Mail, and Mixed Mode Surveys The Tailored Design Method*. Hoboken, NY: John Wiley and Sons, Inc.; 2009.
124. Linstone H, Turoff M. *The Delphi method: techniques and applications*. Reading, MA: Addison-Wesley Publishing Company 1975.
125. Campbell S, Hann M, Roland M, Quayle JA, Shekelle P. The Effect of Panel Membership and Feedback on Ratings in a Two-Round Delphi Survey: Results of a Randomized Controlled Trial. *Medical Care*. 1999;37(9):964-968.
126. de Villiers MR, de Villiers PJT, Kent AP. The Delphi Technique in Health Sciences Education Research. *Medical Teacher*. 2005;27(7):639-643.

127. Duffield C. The Delphi Technique: a comparison of results obtained using two expert panels. *Int J Nurs Stud.* Jun 1993;30(3):227-237.
128. Erffmeyer RC, Erffmeyer ES, Lane IM. The Delphi Technique: An Empirical Evaluation of the Optimal Number of Rounds. *Group Organization Management.* 1986;11(1-2):120-128.
129. Defloor T, Van Hecke A, Verhaeghe S, Gobert M, Darras E, Grypdonck M. The clinical nursing competences and their complexity in Belgian general hospitals. *Journal of Advanced Nursing.* 2006;56(6):669-678.
130. Forrest FC, Taylor MA, Postlethwaite K, Aspinall R. Use of a high-fidelity simulator to develop testing of the technical performance of novice anaesthetists. *Br. J. Anaesth.* March 1, 2002 2002;88(3):338-344.
131. Garfunkel LC, Sidelinger DE, Rezet B, Blaschke GS, Risko W. Achieving Consensus on Competency in Community Pediatrics. *Pediatrics.* April 1, 2005 2005;115(4):1167-1171.
132. Hobgood C, Riviello R, Jouriles N, Hamilton G. Assessment of communication and interpersonal skills competencies. *Academic Emergency Medicine.* 2002;9(11):1257-1269.
133. Irvine F. Exploring district nursing competencies in health promotion: the use of the Delphi technique. *Journal of Clinical Nursing.* 2005;14(8):965-975.
134. Lindsay P, Schull M, Bronskill S, Anderson G. The Development of Indicators to Measure the Quality of Clinical Care in Emergency Departments Following a Modified-Delphi Approach. *Academic Emergency Medicine.* 2008;9(11).
135. Lofmark A, Thorell-Ekstrand I. An assessment form for clinical nursing education: a Delphi study. *Journal of Advanced Nursing.* 2004;48(3):291-298.
136. Pflieger D, McHattie L, Diack H, McCaig D, Stewart D. Developing consensus around the pharmaceutical public health competencies for community pharmacists in Scotland. *Pharmacy World & Science.* 2008/01/21/ 2008;30(1):111-119.
137. Polivka BJ, Stanley SAR, Gordon D, Taulbee K, Kieffer G, McCorkle SM. Public Health Nursing Competencies for Public Health Surge Events. *Public Health Nursing.* 2008;25(2):159-165.
138. *Survey Monkey* [computer program]. Version. Palo Alto, California, USA SurveyMonkey.com, LLC 2009.
139. Gibson F, Soanes L, Gibson F. The Development of Clinical Competencies for Use on a Paediatric Oncology Nursing Course Using a Nominal Group Technique. *Journal of Clinical Nursing.* 2000;9(3):459-469.

140. Walley T, Webb DJ. Developing a core curriculum in clinical pharmacology and therapeutics: a Delphi study. *British Journal of Clinical Pharmacology*. 1997;44(2):167-170.
141. Polit DF, Beck CT. The Content Validity Index: Are you sure you know what's being reported? *Research in Nursing & Health*. 2006;29(5):489-497.
142. Wynd CA, Schmidt B, Schaefer MA. Two Quantitative Approaches for Estimating Content Validity. *West J Nurs Res*. August 1, 2003 2003;25(5):508-518.
143. Beckstead JW. Content validity is naught. *International Journal of Nursing Studies*. 2009;46:1274–1283.
144. Doros G, Lew R. Design Based on Intra-Class Correlation Coefficients *American Journal of Biostatistics*. 2010;1(1):1-8.
145. Bonett DG. Sample Size Requirements for Estimating Intraclass Correlations with Desired Precision. *Statistics in Medicine*. 2002;21(9):1331-1335.
146. Berk RA. Generalizability of Behavioral Observations: A Clarification of Interobserver Agreement and Interobserver Reliability. *Am J Ment Defic*. Mar 1979;83(5):460-472.
147. McGraw KO, Wong SP. Forming Inferences About Some Intraclass Correlation Coefficients. *Psychological Methods*. 1996;1(1):30-46.
148. Shoukri MM, Asyali MH, Walter SD. Issues of Cost and Efficiency in the Design of Reliability Studies. *Biometrics*. 2003;59(4):1107-1112.
149. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. Mar 1979;86(2):420-428.
150. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics*. 1977;33(1):159-174.
151. Agresti A. *An Introduction to Categorical Data Analysis*. Second ed. Hoboken, NJ: John Wilry & Sons, Inc.; 2007.
152. Myford CM, Wolfe EW. Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *J Appl Meas*. 2004;5(2):189-227.
153. Iramaneerat C, Yudkowsky R. Rater Errors in a Clinical Skills Assessment of Medical Students. *Eval Health Prof*. September 1, 2007 2007;30(3):266-283.
154. Motycka CA, Rose RL, Ried LD, Brazeau G. Self-Assessment in Pharmacy and Health Science Education and Professional Practice. *American Journal of Pharmaceutical Education*. 2010;74(5).

155. Saal FE, Downey RG, Lahey MA. Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*. 1980;88(2):413-428.
156. Linn RL, Gronlund NE. *Measurement and assessment in teaching*. 8th ed. Upper Saddle River, NJ: Prentice Hall.; 2000.
157. Maxim BR, Dielman TE. Dimensionality, Internal Consistency and Interrater Reliability of Clinical Performance Ratings. *MEDICAL EDUCATION*. 1987;21(2):130-137.
158. Daelmans HEM, Hem-Stokroos HHvd, Hoogenboom RJI, Scherpbier AJJA, Stehouwer CDA, Vleuten CPMvd. Global clinical performance rating, reliability and validity in an undergraduate clerkship. *The Netherlands Journal of Medicine*. 2005;6 3(7).
159. Haber R, Avins A. Do ratings on the american board of internal medicine resident evaluation form detect differences in clinical competence? *Journal of General Internal Medicine*. 1994;9(3):140-145.
160. Kwolek CJ, Donnelly MB, Sloan DA, Birrell SN, Strodel WE, Schwartz RW. Ward Evaluations: Should They Be Abandoned? *Journal of Surgical Research*. 1997;69(1):1-6.
161. Davis JD. Comparison of Faculty, Peer, Self, and Nurse Assessment of Obstetrics and Gynecology Residents. *Obstetrics & Gynecology*. 2002;99(4):647-651.
162. Durning SJ, Cation LJ, Jackson JL. The reliability and validity of the American Board of Internal Medicine Monthly Evaluation Form. *Acad Med*. Nov 2003;78(11):1175-1182.
163. Roach K, Gandy J, Deusinger SS, et al. The Development and Testing of APTA Clinical Performance Instruments. *Physical Therapy*. 2002;82(4):329-353.
164. Meldrum D, Lydon A-M, Loughnane M, et al. Assessment of undergraduate physiotherapist clinical performance: investigation of educator inter-rater reliability. *Physiotherapy*. 2008;94(3):212-219.
165. CAPE Advisory Panel on Educational Outcomes. Professional and General Abilities Based Outcomes pdf] http://www.aacp.org/Docs/MainNavigation/ForDeans/5763_CAPEoutcomes.pdf. Accessed November 7,2008, 2008.
166. APTA Task Force. The development and testing of APTA clinical performance instruments. *Physical Therapy*. 2002/04// 2002;82(4):329-353.
167. (ACGME) ACfGME. Outcome Project. <http://www.acgme.org/outcome/project/proHome.asp>, 2008.

168. Kramer G, Neumann L. Validation of the National Board Dental Hygiene Examination. *Journal of Dental Hygiene*. 2007/06//2007 Summer 2007;81(3):63.
169. Huddle TS, Heudebert GR. Viewpoint: Taking Apart the Art: The Risk of Anatomizing Clinical Competence. *Academic Medicine*. 2007;82(6):536-541.
170. ten Cate O, Scheele F. Viewpoint: Competency-Based Postgraduate Training: Can We Bridge the Gap between Theory and Clinical Practice? *Academic Medicine*. 2007;82(6):542-547.
171. Elstad EA, Lutfey KE, Marceau LD, Campbell SM, von dem Knesebeck O, McKinlay JB. What do physicians gain (and lose) with experience? *Social Science & Medicine*. 2010;70(11):1728-1736.
172. Zayas T. Qualities of Effective Preceptors of Physician Assistant Students. *Perspectives on Physician Assistant Education*. 1999;10(1):7-11.
173. Remmen R, Denekens J, Scherpbier A, et al. An evaluation study of the didactic quality of clerkships. *MEDICAL EDUCATION*. 2000;34(6):460-464.
174. Busari JO, Scherpbier AJJA, van der Vleuten CPM, Essed GGM. The perceptions of attending doctors of the role of residents as teachers of undergraduate clinical students. *MEDICAL EDUCATION*. 2003;37(3):241-247.
175. Kachalia A, Studdert DM. Professional Liability Issues in Graduate Medical Education. *JAMA*. September 1, 2004 2004;292(9):1051-1056.
176. Meldrum D, Lydon AM, Loughnane M, et al. Assessment of undergraduate physiotherapist clinical performance: investigation of educator inter-rater reliability. *Physiotherapy*. Sep 2008;94(3):212-219.
177. Loomis J. Evaluating clinical competence of physical therapy students. Part 1: the development of an instrument. *Physiother Can*. Mar-Apr 1985;37(2):83-89.
178. Robbins SP. *Organizational behavior*. 4th ed. Englewood Cliffs, NJ: Prentice Hall; 1989.
179. Longford NT. Reliability of Essay Rating and Score Adjustment. *Journal of Educational and Behavioral Statistics*. September 21, 1994 1994;19(3):171-200.
180. Bargagliotti T, Luttrell M, Lenburg C. Reducing Threats to the Implementation of a Competency-Based Performance Assessment System. *Online Journal of Issues in Nursing*. Vol 4, No. 2. 1999;4(2).
181. Lievens F, Sanchez JI. Can training improve the quality of inferences made by raters in competency modeling? A quasi-experiment. *Journal of Applied Psychology*. 2007;92(3):812-819.

182. Beck DE, O'sullivan PS, Boh LE. INCREASING THE ACCURACY OF OBSERVER RATINGS BY ENHANCING COGNITIVE PROCESSING SKILLS. *American Journal of Pharmaceutical Education*. Fal 1995;59(3):228-235.
183. Holmboe E, Hawkins R, Huot S. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Annals of Internal Medicine*. 2004/06//2004 Jun 1 2004;140(11):874.
184. Lievens F. Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*. 2001;86(2):255-264.
185. Vendrely A, Carter R. The influence of training on the rating of physical therapist student performance in the clinical setting. *Journal of Allied Health*. 2004/03//2004 Spring 2004;33(1):62-69.
186. Scullen SE, Mount MK, Goff M. Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*. 2000;85(6):956-970.
187. Brown N. What are the criteria that mentors use to make judgements on the clinical performance of student mental health nurses? An exploratory study of the formal written communication at the end of clinical nursing practice modules. *Journal of Psychiatric & Mental Health Nursing*. 2000;7(5):407-416.
188. Komorita SS, Graham WK. Number of Scale Points and the Reliability of Scales. *Educational and Psychological Measurement*. December 1, 1965 1965;25(4):987-995.
189. Gardner PL. Scales and Statistics. *Review of Educational Research*. January 1, 1975 1975;45(1):43-57.
190. Sousa AC, Wagner DP, Henry RC, Mavis BE. *Better data for teachers, better data for learners, better patient care: college-wide assessment at Michigan State University's College of Human Medicine*; 2011.
191. Commission to Implement Change in Pharmaceutical Education. What is the mission of pharmaceutical education? Background paper I. . *American Journal of Pharmecutical Educucation*. 1994;57:374-376.
192. Commission to Implement Change in Pharmaceutical Education. Entry-level education in pharmacy: Commitment to change Background Paper II. *American Journal of Pharmaceutical Education*. 1994;57:366-374.
193. ASHP ASoH-SP. *ASHP Health-System Pharmacy 2015 Initiative. Revised March 2008*: American Society of Health-System Pharmacists; 2005.
194. American College of Clinical Pharmacy, Burke JM, Miller WA, et al. ACCP White Paper: Clinical Pharmacist Competencies. *Pharmacotherapy*. 2008;28(6).

195. ASHP. ASHP long-range vision for the pharmacy work force in hospitals and health systems. *Am J Health-Syst Pharm.* 2007;64:1320-1330.
196. NACDS, NCPA, APhA. *Project Destiny Executive Summary*. : National Association of Chain Drug Stores, National Community Pharmacists Association, American Pharmacists Association. Project Destiny Executive Summary. ; 2008.
197. Kelly KA, Coyle JD, McAuley JW, Wallace LJ, Buerki RA, Frank SG. EVALUATION, ASSESSMENT, AND OUTCOMES: THE AACP INSTITUTE SUPPLEMENT Writing PharmD Program-Level, Ability-Based Outcomes: Key Elements for Success. *American Journal of Pharmaceutical Education* 2008; 72 (5) Article 98. 2008.
198. Anderson M, Cohen J, Hallock J, Kassebaum D, Turnbull JW. *Med School Objectives Writing Group. Learning objectives for medical student education – guidelines for medical schools: Report I of the Medical School Objectives Project* Academic Medicine; 1999.
199. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of Rater Training on Reliability and Accuracy of Mini-CEX Scores: A Randomized, Controlled Trial. *J Gen Intern Med.* 2009;29(1):74–79.
200. Linssen T, Van Dalen J, Rethans J-J. Simulating the longitudinal doctor–patient relationship: experiences of simulated patients in successive consultations. *MEDICAL EDUCATION.* 2007;41(9):873-878.
201. McKinley DW, Boulet JR, Hambleton RK. A work-centered approach for setting passing scores on performance-based assessments. *Eval Health Prof.* Sep 2005;28(3):349-369.
202. Barman A. Standard Setting in Student Assessment: Is a Defensible Method Yet to Come? *Ann Acad Med Singapore.* 2008;37:957-963.
203. Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher.* 2000;22(2):120-130.
204. Wind LA, Van Dalen J, Muijtjens AMM, Rethans J-J. Assessing simulated patients in an educational setting: the MaSP (Maastricht Assessment of Simulated Patients). *MEDICAL EDUCATION.* 2004;38(1):39-44.
205. Beck DE, Boh LE, O 'Sullivan PS. Evaluating Student Performance in the Experiential Setting with Confidence. *American Journal of Pharmaceutical Education* 1995;59(3).
206. Brazeau C, Boyd L, Crosson J. Changing an Existing OSCE to a Teaching Tool: The Making of a Teaching OSCE. *Academic Medicine.* 2002;77(9).

207. Cicchetti DV, Shoinralter D, Tyrer PJ. The Effect of Number of Rating Scale Categories on Levels of Interrater Reliability: A Monte Carlo Investigation. *Applied Psychological Measurement*. 1985;9(1):31-36.
208. Cox EP. The Optimal Number of Response Alternatives for a Scale: A Review. *Journal of Marketing Research*. 1980;17(4):407-422.
209. Preston CC, Colman AM. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*. 2000;104(1):1-15.
210. Symonds PM. On the Loss of Reliability in Ratings Due to Coarseness of the Scale. *Journal of Experimental Psychology*. 1924;7(6):456
211. Ried LD. Professor and Chair College of Pharmacy University of South Florida; 2010.
212. Adams CL, Glavin K, Hutchins K, Lee T, Zimmerman C. An Evaluation of the Internal Reliability, Construct Validity, and Predictive Validity of the Physical Therapist Clinical Performance Instrument. *Journal of Physical Therapy Education*. 2008;22(2):42-50.
213. Jungnickel P, Kelley K, Marlowe K, Haines S, Hammer D. Addressing competencies for the future in the professional curriculum. Paper presented at: AACP Curricular Change Summit, 2009.
214. Eva KW, Cunnington JPW, Reiter HI, Keane DR, Norman GR. How Can I Know What I Don't Know? Poor Self Assessment in a Well-Defined Domain. *Advances in Health Sciences Education*. 2004;9(3):211-224.
215. Langendyk V. Not knowing that they do not know: self-assessment accuracy of third-year medical students. *MEDICAL EDUCATION*. 2006;40(2):173-179.
216. Colthart I, Bagnall G, Evans A, et al. The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice: BEME Guide no. 10. *Med Teach*. 2008;30(2):124-145.
217. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of Physician Self-assessment Compared With Observed Measures of Competence: A Systematic Review. *JAMA*. September 6, 2006 2006;296(9):1094-1102.
218. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ*. 2010;341:c5064.
219. Lankshear AJ. Failure to fail: the teacher's delemma. *Nursing Standard*. 1990;4(20).

- 220.** USLME. United States Medical Licensing Examination. <http://www.usmle.org/>, 2009.
- 221.** Tochel C, Haig A, Hesketh A, et al. The effectiveness of portfolios for post-graduate assessment and education: BEME Guide No 12. *Med Teach*. Apr 2009;31(4):299-318.
- 222.** Long DM. Competency-based residency training: the next advance in graduate medical education. *Acad Med*. Dec 2000;75(12):1178-1183.
- 223.** Reznick RK, MacRae H. Teaching Surgical Skills: Changes in the Wind. *New England Journal of Medicine*. 2006;355(25):2664-2669.
- 224.** Bhatti NI, Cummings CW. Competency in surgical residency training: defining and raising the bar. *Acad Med*. Jun 2007;82(6):569-573.
- 225.** Meyers FJ, Weinberger SE, Fitzgibbons JP, Glassroth J, Duffy FD, Clayton CP. Redesigning residency training in internal medicine: the consensus report of the Alliance for Academic Internal Medicine Education Redesign Task Force. *Acad Med*. Dec 2007;82(12):1211-1219.
- 226.** Brown AK, O'Connor PJ, Roberts TE, Wakefield RJ, Karim Z, Emery P. Ultrasonography for rheumatologists: the development of specific competency based educational outcomes. *Ann Rheum Dis*. May 2006;65(5):629-636.
- 227.** Dowson C, Hassell A, on behalf of the members of the Training Sub-committee of the West Midlands Rheumatology S, Training C. Competence-based assessment of specialist registrars: evaluation of a new assessment of out-patient consultations. *Rheumatology*. April 2006 2006;45(4):459-464.
- 228.** Hubbard JP, Levitt EJ, Schumacher CF. An objective evaluation of clinical competence. *N Engl J Med*. 1963;272:1321-1328.

BIOGRAPHICAL SKETCH

Charles Douglas was born and raised in Los Angeles California. He holds an Executive Masters of Business Administration from the Peter F. Drucker Graduate School of Management at the Claremont Graduate University. He is the recipient of a 2007 American Association of Colleges of Pharmacy/Wal-Mart scholarship, and the department's Graduate Teaching Assistant Award in 2008. He is interested in evaluation of pharmacy clinical services and pharmacy administration.