


January 2012

# The Relationship between Rating Scales used to Evaluate Tasks from Task Inventories for Licensure and Certification Examinations

Adrienne W. Cadle

University of South Florida, [adriennewcadle@gmail.com](mailto:adriennewcadle@gmail.com)

Follow this and additional works at: <http://scholarcommons.usf.edu/etd>

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), and the [Psychology Commons](#)

---

## Scholar Commons Citation

Cadle, Adrienne W., "The Relationship between Rating Scales used to Evaluate Tasks from Task Inventories for Licensure and Certification Examinations" (2012). *Graduate Theses and Dissertations*.  
<http://scholarcommons.usf.edu/etd/4296>

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact [scholarcommons@usf.edu](mailto:scholarcommons@usf.edu).

The Relationship between Rating Scales used to Evaluate Tasks from Task Inventories  
for Licensure and Certification Examinations

by

Adrienne Woodley Cadle

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
Department of Educational Measurement and Research  
College of Education  
University of South Florida

Major Professor: Jeffrey Kromrey, Ph.D.  
Robert Dedrick, Ph.D.  
John Ferron, Ph.D.  
Michael Brannick, Ph.D.

Date of Approval:  
October 18, 2012

Keywords: Job Analysis, Task Analysis, Survey Validation Study, Survey Validation  
Methodology, Task Rating Scales

Copyright © 2012, Adrienne Woodley Cadle

## DEDICATION

It is difficult to define the emotional and mental rollercoaster that one rides while completing a doctoral degree. There are moments when you feel like you are on top of the world, knowing everything there is to know about a topic, and moments when you stare at that blank screen on your computer think “there is no way I am going to be able to do this”. If it weren’t for the love and encouragement of my family, I’m not sure I would be where I am today.

To my mother, Cynthia Woodley, you are my foundation. You planted the seed in me from such a young age that I could do anything I wanted if I put my mind to it. You are a true inspiration to me, as an amazing, tough, hard-working and caring individual. I cannot quantify the number of hours you and David spent talking to me about every measurement topic under the sun; how much I have learned from you both along the way. And for all of that, I thank you.

To my husband, Drew Cadle, you have been my rock on this journey. In the five and a half years since I started working on my doctoral degree, you have never once doubted my ability to finish. You pushed me when I needed to be pushed and held me when I needed to be held. Because of your love and support, I was able to obtain my doctoral degree while having two incredible children, Roman and Berlin, and still pursue my career. There is no doubt in my mind that I would not have all of this if it weren’t for you. And for that, I am eternally grateful. Thank you, Drew, for giving me everything I could have ever wanted and so much more. I dedicate this work to you.

## ACKNOWLEDGMENTS

For every person who completes a doctoral degree, and goes through the arduous task of writing a dissertation, there are individuals along the way who provide guidance and support. For me, that was my major professor, Dr. Jeffrey Kromrey and a handful of professional colleagues. Dr. Kromrey has this indefinable calming effect. He has this ability to push you to try harder, to make you dig a little deeper, while providing unconditional support. He read and re-read countless versions of my dissertation, each time providing feedback and edits with such patience and grace. There is no way I would have been able to complete this dissertation if it weren't for him. For all of these reasons, and so many more, I would like to acknowledge and thank Dr. Kromrey.

In addition to the unwavering support from Dr. Kromrey, I would like to acknowledge all of the supportive colleagues at Professional Testing, Inc. Thank you to David Cox, Cynthia Woodley, Reed Castle, and Maria Gonzales for giving me time off to “finish this thing already”. Your belief in me was more motivating than you will ever know. To Fae Mellichamp, Lynn Webb, Christine Neiro, and Cynthia Woodley – you four have been my mentors for as long as I can remember. Thank you for allowing me to bounce ideas off of you on a seemingly daily basis, and for always taking time away from your hectic lives to guide me in this journey. To Melba Joiner, Isaac Li, and Lisa Everts, thank you for taking on extra responsibilities in my absence, and making me laugh when I wanted to pull my hair out. And a special thanks to my dear friend, work colleague, and fellow graduate student, Corina Owens, who was a sounding board for every aspect of my dissertation. You made this uphill journey so much more bearable.

## TABLE OF CONTENTS

LIST OF TABLES .....	iii
LIST OF FIGURES .....	vi
ABSTRACT .....	ix
CHAPTER ONE: INTRODUCTION.....	1
Problem Statement .....	3
Purpose of Study .....	4
Research Questions.....	5
Importance of Study.....	6
Definition of Terms.....	7
CHAPTER TWO: LITERATURE REVIEW.....	10
Licensure and Certification Testing.....	10
Job Analysis for Credentialing Exams.....	16
Critical Incident Technique.....	17
Functional Job Analysis.....	21
DACUM.....	23
Task Inventory Analysis .....	27
Survey Validation Studies for Job Analyses.....	32
Collecting Survey Respondent Demographic Information.....	33
Task Rating Scales.....	33
Relationships of Task Rating Scales.....	38
Survey Design Related to Scale Placement .....	43
Data Analysis and the Development of Examination Blueprints .....	46
The Use of Correlations in Meta-Analysis Research.....	47
CHAPTER THREE: METHODS.....	50
Purpose.....	50
Research Questions.....	50
Overview of Research Design .....	51
Sample Studies.....	52
Representativeness of Survey Respondents in Sample Studies.....	53
Coding of Sample Studies.....	64
Data Analysis .....	69
Missigness Analysis.....	69
Outlier Analysis .....	72

CHAPTER FOUR: RESULTS .....	79
Research Questions .....	79
Research Question One Results .....	80
Research Questions Two and Three Results .....	88
Research Questions Two and Three for All Correlations .....	88
Research Questions Two and Three for Correlations between Individual Scales .....	94
Research Questions Two and Three for Correlations between Composite Scales .....	100
Research Questions Two and Three for Correlations between Individual and Composite Scales .....	106
Research Question Four Results .....	112
CHAPTER FIVE: DISCUSSION .....	121
Summary of Individual and Composite Rating Scale Findings .....	122
Importance and Criticality Rating Scales .....	122
Frequency and Importance Rating Scales .....	123
Remaining Individual Rating Scales .....	124
Composite Rating Scales .....	125
Individual with Composite Rating Scales .....	127
Summary of Potentially Moderating Variables .....	127
Summary of Examination Blueprint Development Findings .....	128
Implications for Practice .....	131
Limitations and Implications for Future Research .....	133
REFERENCES .....	136
APPENDICES .....	142
Appendix A: Additional Detail Regarding Sample Studies .....	143
Appendix B: Correlation of Survey Variables .....	148
ABOUT THE AUTHOR .....	End Page

## LIST OF TABLES

Table 1. A Description of Rating Scales Used for Survey Validation Studies.....	37
Table 2. States in Which Respondents Reported Working.....	60
Table 3. Institutions in Which Respondents Reported Working .....	61
Table 4. Survey Respondents Highest Reported Education .....	63
Table 5. Reported Backgrounds of Physicians .....	63
Table 6. A Breakdown of The Number and Types of Tasks Used in Each Survey Validation Study.....	66
Table 7. A Breakdown of the 20 Sample Studies Based on Industry, Sample Size, Presentation Order, and Number of Tasks Rated .....	67
Table 8. Description of Outlier Correlations .....	73
Table 9. Distribution of Correlations Between 20 Studies .....	81
Table 10. Mean Weighted Correlation, CIs and PIs for All Combinations of Scales by Industry.....	89
Table 11. Fixed and Random Effects for Industries on Correlations .....	90
Table 12. Mean Weighted Correlation, CIs and PIs for All Combinations of Scales by Sample Size .....	91
Table 13. Fixed and Random Effects for Sample Size on Correlations .....	91
Table 14. Mean Weighted Correlation, CIs and PIs for All Combinations of Scales by Presentation Order.....	92
Table 15. Fixed And Random Effects for Presentation Order on Correlations .....	93
Table 16. Fixed and Random Effects for Number of Tasks on Correlations .....	93
Table 17. Mean Weighted Correlation, CIs and PIs for All Combinations of Scales by Number of Tasks .....	94
Table 18. Fixed And Random Effects for All Potential Moderator Variables in All Correlations on All Pairings of Scales .....	95

Table 19. Mean Weighted Correlation, CIs and PIs for All Pairings of Individual Scales by Industry.....	96
Table 20. Fixed and Random Effects for Industries on Correlations .....	97
Table 21. Mean Weighted Correlation, CIs, and PIs for All Pairings of Individual Rating Scales By Sample Size .....	97
Table 22. Fixed And Random Effects for Sample Size on Correlations .....	98
Table 23. Mean Weighted Correlation, CIs, and PIs for All Pairings of Individual Rating Scales by Presentation Order.....	99
Table 24. Fixed and Random Effects for Presentation Order on Correlations .....	99
Table 25. Mean Weighted Correlation, CIs, and PIs for All Pairings of Individual Rating Scales by Number of Tasks.....	99
Table 26. Fixed and Random Effects for Number of Tasks on Correlations .....	100
Table 27. Fixed and Random Effects for All Potential Moderator Variables on All Correlations of Pairings of Individual Scales.....	101
Table 28. Mean Weighted Correlation, CIs, and PIs for All Combinations of Composite Rating Scales by Industry.....	102
Table 29. Fixed and Random Effects for Industries on Correlations .....	103
Table 30. Mean Weighted Correlation, CIs, and PIs for All Combinations of Composite Rating Scales by Sample Size.....	104
Table 31. Fixed and Random Effects for Sample Size on Correlations .....	104
Table 32. Mean Weighted Correlation, CIs, and PIs for All Combinations of Composite Rating Scales by Presentation Order .....	105
Table 33. Fixed and Random Effects for Presentation Order on Correlations .....	105
Table 34. Mean Weighted Correlation, CIs, and PIs for All Combinations of Composite Rating Scales by Number of Tasks .....	105
Table 35. Mean Weighted Correlation, CIs, and PIs for Pairings of Individual and Composite Rating Scales by Industry .....	107
Table 36. Fixed and Random Effects for Industries on Correlations .....	108
Table 37. Mean Weighted Correlation, CIs, and PIs for Pairings of Individual and Composite Rating Scales by Sample Size.....	109



Table 38. Fixed and Random Effects for Sample Size on Correlations .....	109
Table 39. Mean Weighted Correlation, CIs, and PIs for Pairings of Individual and Composite Rating Scales by Presentation Order.....	109
Table 40. Fixed and Random Effects for Presentation Order on Correlations .....	110
Table 41. Mean Weighted Correlation, CIs, and PIs for Pairings of Individual and Composite Rating Scales by Number of Tasks.....	111
Table 42. Fixed and Random Effects for Number of Tasks on Correlations .....	111
Table 43. Fixed and Random Effects for All Potential Moderator Variables on All Correlations of Pairings of Individual and Composite Rating Scales.....	112
Table 44. Correlations Between Relative Rankings of Duty Areas on Derived Examination Blueprints and Actual Examination Blueprints .....	114
Table 45. Comparison Between Duty Weights on Actual and Derived Examination Blueprints for a Certification Exam .....	116
Table 46. Distribution of Absolute Differences Between the Weights of Derived Exam Blueprints and Actual Exam Blueprints.....	117
Table 47. Average Absolute Differences Between Duty Weights on Actual and Derived Examination Blueprints and Exam Blueprints in which All Duties are Equally Weighted .....	119
Table 48. Coding of Sample Studies.....	143
Table 49. Correlations of Sample Study Variables.....	148

## LIST OF FIGURES

Figure 1. Rating each task based on a single scale, then rating each task again based on additional scales.....	4
Figure 2. Rating one task at a time, based on multiple scales. ....	4
Figure 3. Illustration of a DACUM Chart.....	24
Figure 4. Scales presented one at a time, in which survey respondents rate all tasks on one scale before moving onto the next scale. ....	45
Figure 5. Scales presented together, in which survey respondents rate all tasks on one scale before moving onto the next task.....	45
Figure 6. Illustration of single scale used in survey validation studies. ....	53
Figure 7. Other credentials obtained by survey respondents in one of the studies included in this analysis.....	56
Figure 8. Plumbing codes followed by survey respondents in one of the studies included in this analysis.....	56
Figure 9. Highest level of education reported by survey respondents in one of the studies included in this analysis.....	57
Figure 10. Reported age of survey respondents in one of the studies included in this analysis.....	58
Figure 11. Reported gender of survey respondents in one of the studies included in this analysis.....	59
Figure 12. The number of physicians reported as working in the Phlebology portion of the practice in one of the studies included in this analysis. ....	62
Figure 13. The number of vascular technologists reported as working in the Phlebology portion of their practice in one of the studies included in this analysis.....	63
Figure 14. The number of years respondents reported working in Phlebology in one of the studies included in this analysis.....	64
Figure 15. Rating one task at a time, based on multiple scales. ....	68

Figure 16. Number of survey respondents who completed the journeyman plumber validation survey. ....	70
Figure 17. Number of survey respondents who completed the Phlebology validation survey. ....	71
Figure 18. Distribution correlations between task ratings and missingness. N=478,079. ....	72
Figure 19. Distribution correlations between the amount of missingness across all task rating scales. N=222,064. ....	73
Figure 20. Illustration of data analysis method to answer research question 1A. ....	74
Figure 21. Illustration of how the summary correlation are be derived for each of the pairs of rating scales and/or composites using Fisher’s $z_r$ -transformation. ....	76
Figure 22. Distribution of all obtained correlations. N=129. ....	82
Figure 23. Distribution correlations between two individual scales. N=34. ....	83
Figure 24. Distribution correlations between two composite scales. N=15. ....	83
Figure 25. Distribution correlations between a composite and individual scale. N=80. ....	84
Figure 26. Distribution of all obtained correlations by all combinations of scales. N=129. ....	85
Figure 27. Distribution of all obtained correlations for composites with composites, individual scales with composite scales, and individual scales with individual scales. N=129. ....	86
Figure 28. Distribution of correlations between individual scales. N=34. ....	87
Figure 29. Distribution of all correlations between composite scales. N=15. ....	87
Figure 30. Distribution of all correlations between individual and composite scales. N=80. ....	88
Figure 31. Distribution of all obtained correlations by industry. N=129. ....	90
Figure 32. Distribution of all obtained correlations by sample size. N=129. ....	91
Figure 33. Distribution of all obtained correlations by presentation order. N=129. ....	92
Figure 34. Distribution of all obtained correlations by number of tasks. N=129. ....	94
Figure 35. Distribution of correlations between individual rating scales by industry. N=34. ....	96

Figure 36. Distribution of correlations between individual rating scales by sample size. N=34. ....	97
Figure 37. Distribution of correlations between individual rating scales by presentation order. N=34. ....	98
Figure 38. Distribution of correlations between individual rating scales by number of tasks. N=34. ....	100
Figure 39. Distribution of correlations between pairings of composite rating scales by industry. N=15. ....	102
Figure 40. Distribution of correlations between pairings of composite rating scales by sample size. N=15. ....	103
Figure 41. Distribution of correlations between pairings of composite rating scales by presentation order. N=15. ....	104
Figure 42. Distribution of correlations between pairings of composite rating scales by number of tasks. N=15. ....	106
Figure 43. Distribution of correlations between individual and composite rating scales by industry. N=80. ....	107
Figure 44. Distribution of correlations between individual and composite rating scales by sample size. N=80. ....	108
Figure 45. Distribution of correlations between individual and composite rating scales by presentation order. N=80. ....	110
Figure 46. Distribution of correlations between individual and composite rating scales by number of tasks. N=80. ....	111
Figure 47. Sample examination blueprint. ....	113
Figure 48. Distribution of correlations between relative ranks of examination blueprints derived from individual and composite scale, as well as the actual examination blueprints used on the licensure or certification exam. N=30. ....	115
Figure 49. Absolute differences between duty areas on actual and derived examination blueprints and duty areas on examination blueprints in which all duty areas are equal. ....	120
Figure 50. Pairings of Composite scales used in data analysis. ....	125

## ABSTRACT

The first step in developing or updating a licensure or certification examination is to conduct a job or task analysis. Following completion of the job analysis, a survey validation study is performed to validate the results of the job analysis and to obtain task ratings so that an examination blueprint may be created. Psychometricians and job analysts have spent years arguing over the choice of scales that should be used to evaluate job tasks, as well as how those scales should be combined to create an examination blueprint. The purpose of this study was to determine the relationship between individual and composite rating scales, examine how that relationship varied across industries, sample sizes, task presentation order, and number of tasks rated, and evaluate whether examination blueprint weightings would differ based on the choice of scales or composites of scales used. Findings from this study should be used to guide psychometricians and job analysts in their choice of rating scales, choice of composites of rating scales, and how to create examination blueprints based upon individual and/or composite rating scales.

A secondary data analysis was performed to help answer some of these questions. As part of the secondary data analysis, data from 20 survey validation studies performed during a five year period were analyzed. Correlations were computed between 29 pairings of individual and composite rating scales to see if there were redundancies in task ratings. Meta-analytic techniques were used to evaluate the relationship between each pairing of rating scales and to determine if the relationship between pairings of rating scales was impacted by several factors. Lastly, sample examination blueprints

were created from several individual and composite rating scales to determine if the rating scales that were used to create the examination blueprints would ultimately impact the weighting of the examination blueprint.

The results of this study suggest that there is a high degree of redundancy between certain pairs of scales (i.e., the Importance and Criticality rating scale are highly related), and a somewhat lower degree of redundancy between other rating scales; but that the same relationship between rating scales is observed across many variables, including the industry for which the job analysis was being performed. The results also suggest the choice of rating scales used to create examination blueprints does not have a large effect on the finalized examination blueprint. This finding is especially true if a composite rating scale is used to create the weighting on the examination blueprint.

## CHAPTER ONE: INTRODUCTION

The face of licensure and certification testing has changed dramatically over the past sixty years. What was once a group of men sitting in a room deciding what to put on a credentialing exam is now a systematic process for exam development. This systematic process has evolved over time based on organizations' desires to credential people and on the growing number of lawsuits related to credentialing exams. As more organizations seek to develop credentialing exams, it is imperative that each component of the exam development process be detailed and agreed upon prior to development. The steps for developing a licensure or certification exam involve conducting a job or task analysis, performing a survey validation study, developing an examination blueprint, writing items, assembling an exam form, reviewing the initial exam form, conducting an initial pilot test of the exam, and setting a passing score.

Upon completion of a job or task analysis, a survey is administered to validate the resulting task list. This process is called a survey validation study, the purpose of which is twofold: to confirm the results of the task analysis and to help develop an examination blueprint. The survey validation study involves asking job incumbents to rate each job task on one or more rating scales. Some examples of the types of scales used are listed below:

- Consequence or Criticality of Error – if the task is performed incorrectly, or not at all, what is the risk of an adverse consequence?

- Difficulty of Learning – how difficult is it to learn how to perform this task?
- Need at Entry – is the task required of entry-level professionals?
- Task Frequency – how often is each task performed?
- Task Importance – how important is it to know how to perform each task?
- Time Spent – how much time is spent performing this task?

Each of the sample scales listed above is a rating scale ranging from three- to five-points, depending on the scale. For example, task frequency is typically used as a five point scale using either absolute frequencies (Daily=4, Weekly=3, Monthly=2, Annually=1, Never=0) or relative frequencies (Very often=4, Fairly often=3, Occasionally=2, Seldom=1, Never=0).

After a task analysis is complete, two decisions must be made before the creation and administration of the survey validation study. First, the job analyst must decide which rating scales should be used to evaluate the task list. Second, if more than one rating scale is used, the job analyst must decide if the rating scales will be combined.

Unfortunately, there is little agreement in the field as to which rating scales should be used in the survey validation process. Friedman (1990) argues that time-spent and importance scales are redundant and that the job analyst should choose either one scale or the other, but not use both scales. Sanchez and Fraser (1992) found that task criticality and task importance rating scales were highly redundant and that job analysts should choose one scale or the other, but should not use both scales at the same time. Sanchez and Fraser also found high correlations between overall task importance rating scales and composites that included task importance, indicating that overall importance



ratings may provide similar results with composite ratings, so overall task importance ratings should be used alone.

In addition to conflicting arguments about the choice of scales, there is disagreement about whether to use one scale or a composite of several scales. And if several scales are used, there is disagreement about how to combine those scales. Sanchez and Levine (1989) found that the composite of criticality and difficulty of learning rating scales provided more reliable task ratings than a single overall importance rating, and that in general, composite ratings of simple rating scales would provide more reliable task ratings than highly complex single rating scales. This finding is contrary to Sanchez and Fraser's (1992) findings that overall task importance ratings are just as reliable as composites that include task importance. Kane, Kingsbury, Colton, and Estes (1989) recommend that a multiplicative model combining criticality and frequency be used. Raymond (2005) recommends combining any two or more unidimensional rating scales into an overall composite rating, rather than using a single rating scale. Lastly, Spray and Huang (2000) recommended a composite of scales, but only after using IRT to transform ordinal rating scales into interval scales.

### **Problem Statement**

As illustrated above, there is disagreement in the field as to 1) whether or not one rating scale or a composite of rating scales should be used to rate job tasks, 2) if one scale is used, the overall scale that should be used, and 3) if a composite of scales is used, how the scales are combined. While this study could not answer all of these questions, it is a step in the direction towards eventually answering these questions. There are limitations in each of the studies mentioned above. The studies were often conducted in one

industry. Of the studies conducted in multiple industries, there was a small number of survey respondents included in the analysis (significantly less than 100). Each study evaluated the relationships of task ratings in one of two task presentation orders. Some of the studies compared task ratings in which each participant rated all tasks on one scale and then all tasks on a different scale (as illustrated in Figure 1), while others looked at task ratings in which each participant rated one task at a time, but looked at multiple scales for each task (as illustrated in Figure 2). None of the aforementioned studies examined the relationship between scales using both types of presentation orders.

Scale 1	
Task 1	Rating 1
Task 2	Rating 2
Task 3	Rating 3
Scale 2	
Task 1	Rating 4
Task 2	Rating 5
Task 3	Rating 6
Scale 3	
Task 1	Rating 7
Task 2	Rating 8
Task 3	Rating 9

*Figure 1.* Rating each task based on a single scale, then rating each task again based on additional scales.

	Scale 1	Scale 2	Scale 3
Task 1	Rating 1	Rating 2	Rating 3
Task 2	Rating 4	Rating 5	Rating 6
Task 3	Rating 7	Rating 8	Rating 9

*Figure 2.* Rating one task at a time, based on multiple scales.

### **Purpose of Study**

The purpose of this study is to determine the relationship between individual and composite rating scales, examine how that relationship varies across industries, sample

sizes, task presentation order, and number of tasks rated, and evaluate whether examination blueprint weightings would differ based on the choice of scales or composites of scales used. The individual rating scales included in this study are task frequency, task importance, criticality or consequence of error, and need at entry. There are four composite scales included in this study:

- Composite 1 = 2\*Importance + Frequency,
- Composite 2 = Criticality\*Frequency,
- Composite 3 = 2\*Importance + 2\*Criticality + Frequency, and
- Composite 4 = 2\*Importance + Frequency + Need at Entry.

A secondary data analysis was performed using task analysis data from multiple industries. The number of respondents in the survey validation studies varied from less than 100 survey respondents to over than 1,000 respondents. The relationship between individual and composite task ratings was compared when the scales were rated one scale at a time, as well as when scales were rated all at once, one task at a time (presentation order). The relationships between individual and composite task ratings was compared for small task lists (50 tasks or less), medium task lists (51-100 tasks rated), and large task lists (more than 100 tasks rated). Lastly, sample examination blueprint weights were generated based upon the varying choice of scales to determine if the examination blueprint weighting would differ based on the scale or scales used to create the blueprint.

### **Research Questions**

There are four overarching research questions in this study:

1. What is the relationship between the different types of individual and composite rating scales?

- a. What is the relationship between different types of individual rating scales?
  - b. What is the relationship between different types of composite rating scales?
  - c. What is the relationship between different types of individual rating scales and different types of composite scales?
2. To what extent do the relationships of individual and composite rating scales vary across industries?
  3. To what extent do the relationships of individual and composite rating scales vary across survey design factors?
    - a. To what extent do the relationships of individual and composite rating scales vary across varying numbers of survey respondents?
    - b. To what extent do the relationships of individual and composite rating scales vary across scale presentation order?
    - c. To what extent do the relationships of individual and composite rating scales vary across the number of tasks rated?
  4. To what extent are examination blueprint weightings different based on the choice of scale composites used in the survey validation study?

### **Importance of Study**

This study is important to those who perform job analyses with survey validation studies, as both paper-and-pencil and online surveys are expensive and time consuming. If there is a strong relationship between two or three individual rating scales then a job analyst might decide to use only one (or two) of the scales rather than all of the scales, as

it is less time consuming to ask SMEs to evaluate tasks on one or two set(s) of rating scales rather than two or three. Similarly, if there is a strong relationship between an individual rating scale and a composite of rating scales then it might not be worth using the composite rating scale in future survey validation studies. Lastly, if all of the composites produce comparable examination blueprints then job analysts can stop arguing over which composite should be used to create examination blueprints.

### **Definition of Terms**

*Certification.* “The process by which a governmental or nongovernmental agency grants recognition to an individual who has met certain predetermined qualifications set by a credentialing agency” (Shimberg, 1981).

*Credential.* For the purpose of this study, both licenses and certifications are collectively referred to as credentials. Licensure exams and certification exams are collectively referred to as “credentialing exams”.

*Enablers.* “Enablers are essential items that enable workers to perform their duties and tasks but that are not duties or tasks themselves” (DACUM Handbook, p. D-17). Enablers include general knowledge, skills, tools, equipment, resources, and worker behaviors.

*Job Analysis.* “A general term referring to the investigation of positions or job classes to obtain descriptive information about job duties and tasks, responsibilities, necessary worker characteristics (e.g. knowledge, skills, and abilities), working conditions, and/or other aspects of the work” (AERA, APA, NCME, 1999, p. 177). For the purposes of this study, job analysis, practice analysis, and task analysis are used interchangeably.

*Licensure.* “A process by which an agency of government grants permission to an individual to engage in a given occupation upon finding that the application has attained the minimal degree of competency required to ensure that the public health, safety, and welfare will be reasonably well protected” (Schimberg, 1981, p. 1138).

*Meta-analysis.* A statistical tool for combining the effect size of a number of studies to determine if general patterns occur in the data. (Goodwin, 2005).

*Practice Analysis.* “A general term referring to the investigation of a certain work position or profession, to obtain descriptive information about the activities and responsibilities of the position and about the knowledge, skills, and abilities needed to engage in the work of the position. The concept is essentially the same as a job analysis but is generally preferred for professional occupations involving a great deal of individual decision making” (AERA, APA, NCME, 1999, p. 179). While the term “practice analysis” is more commonly used in licensure and certification testing, “job analysis” is a more common term in general, and so for the purpose of this study, the term job analysis includes practice analysis.

*Subject Matter Expert (SME).* “SME, as the term is used by job analysts, refers to a job incumbent, a supervisor of a specific job, or to any person who is intimately familiar with the target job(s)” (Gael, 1988, p.432).

*Task Analysis.* “A systematic method of accounting for all of the behavioral interactions between one or more individuals and a system, together with the conditions that must be satisfied if those interactions are to occur effectively” (Van Cott & Paramore, 1988, p. 651).

*Validity.* “The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test” (AERA, APA, NCME, 1999, p. 184).

## CHAPTER TWO: LITERATURE REVIEW

This literature review is divided into four sections. First, an overview of licensure and certification testing will be presented, along with a description of exam development for licensure and certification tests. Second, the process of performing a job analysis will be discussed, including the varying job analysis methods that are traditionally used in licensure and certification testing. Third, a description of survey validation studies and the rating scales used in the survey validation studies will be presented. Lastly, a brief explanation regarding the use of correlations in meta-analytic research will be provided.

### **Licensure and Certification Testing**

Although the distinction between licensure and certification testing has become blurred, there are differences between the two (Downing, 2006). Licensure is required to perform a job, while certification is often voluntary. Licensure implies minimal competence, whereas certification implies something higher than minimal competence. Licensing is mandated by regulatory bodies or government agencies, while certifications are offered by credentialing bodies or professional organizations. A licensed individual has provided evidence (typically by passing a licensing exam) that he or she knows how to, or is able to, perform a job without harming the health, safety, or well-being of the general public. A certified individual has also provided evidence (through passing a certification exam) that he or she has some knowledge, skills, or abilities, but in certification testing the certified individual has illustrated that he or she has some advanced knowledge or skills above and beyond protecting the health, safety, and welfare



of the general public. For example, a dentist must have a license in order to practice dentistry implying that he or she has the minimal competence necessary to practice safely. The same dentist may desire to later become a board certified general dentist, which would indicate to the public that he or she may practice dentistry at a higher level of proficiency.

Both licenses and certifications fall under the broader heading of “credentials” and in both cases an examination (in conjunction with other requirements) is typically used to determine whether or not a credential should be awarded. Whether or not an individual is granted the credential is often based on the individual meeting some form of eligibility criteria and successfully passing an examination. Credentialing organizations (regardless of whether they are the regulatory bodies that grant licenses or the public or private organizations that grant certifications) are required to follow a set of standards and guidelines that outline how exams should be developed, administered, and scored. Credentialing exams are high-stakes exams because without the credential, an individual is either not allowed to practice (licensure) or is unable to practice at a desired level (certification). As such, it is crucial for credentialing organizations to follow standards in order to provide assessments that are both fair to candidates and legally defensible.

There are a number of guidelines and standards that illustrate how tests or exams used for selection purposes should be developed and maintained, including *The Standards for Educational and Psychological Testing* (AERA, NCME, APA, 1999), *Standards for the Accreditation of Certification Programs* (ICE, 2004), *Code of Fair Testing Practices in Education* (JCTP, 2004), *Principles of Fairness: An Examination Guide for Credentialing Boards* (CLEAR, 1992) and *ISO/IEC 17024* (ISO/IEC, 2003). In

all cases, the most important component in the development of an exam is that the credentialing organization that develops the exam provides evidence of validity. More specifically, the relationship between the examination used to credential an individual and the job in which the individual is being credentialed (the predictor-criterion relationship) must be demonstrated.

The Equal Employment Opportunity Commissions (EEOC) explicitly expresses the need for evidence of the predictor-criterion relationship. As part of the EEOCs Enforcement Guidance's and Related Documents, the Employment Test and Selection Procedures section states:

Employers should ensure that employment tests and other selection procedures are properly validated for the positions and purposes for which they are used. The test or selection procedure must be job-related and its results appropriate for the employer's purpose. (Employer Best Practices for Testing and Selection, Bullet 2)

The EEOC Guidelines, first established in 1966, clearly illustrate the need for credentialing organizations to document the relationship between the exam being used for credentialing purposes, and the job for which one is being credentialed.

Failure to illustrate the relationship between the job for which someone is being credentialed and the examination that is used to determine whether or not a person should be credentialed (the validity of the examination) has dire consequences. In the first groundbreaking lawsuit related to the validity of selection exams, *Griggs v. Duke Power Co.* (1971), a group of 13 African American men sued the Duke Power Company in Draper, North Carolina. The prosecutors argued that Duke Power Company was using

selection tests (both a high school diploma and two aptitude tests) for the purposes of both hiring and promotion and that those tests were not related to the job for which they were being used to hire or promote. The case was eventually tried in front of the Supreme Court with the Court ruling that “neither the high school completion requirement nor the general intelligence test is shown to bear a demonstrable relationship to successful performance of the jobs for which it was used” (para. 12). Furthermore, the Supreme Court ruled that “employees who have not completed high school or taken the tests have continued to perform satisfactorily, and make progress in departments for which the high school and test criteria are now used” (para. 13). The Supreme Court ruled against Duke Power Co. because Duke Power Co. failed to provide validity evidence for their selection tests (or they failed to illustrate the predictor-criterion relationship).

Almost 40 years later (2009), a group of 17 firefighters in New Haven, Connecticut sued the city of New Haven, New Haven’s Mayor, and five other officials based on the New Haven Civil Service Board’s decision to throw-out the results of a selection test used for the promotion of lieutenant and captain positions in the fire department (*Ricci et al. v. DeStefano et al.*). The city hired an outside testing consultant to develop and administer an exam for both the lieutenant and captain positions. The testing consultant, Industrial/Organizational Solutions Inc. (IOS), began the development process by “performing job analyses to identify the tasks, knowledge, skills, and abilities that are essential for the lieutenant and captain positions” (Syllabus, p. 4). As part of the job analysis, IOS conducted interviews of job incumbents and performed observations of job incumbents prior to administering a validation survey of the results of the job analysis

(*Ricci et al. v. DeStefano et al.*). IOS used the results of the job analysis and subsequent survey validation study to develop both written and oral examinations for selection of both lieutenant and captain positions. The process IOS used to develop the selection exams provided evidence of the predictor-criterion relationship, thus providing validity evidence for the exams.

After administering the selection tests, New Haven city officials found that the exam adversely impacted two minority groups (African Americans and Hispanics), and that between the two exams, only 17 white and 2 Hispanic candidates would be eligible for promotion. To *avoid* a lawsuit under Title VII of the Civil Rights Act and EEOC Guidelines the City of New Haven decided to throw-out the results of the selection exams. The resulting lawsuit by the 17 firefighters was eventually tried in the Supreme Court, and in 2009 the Supreme Court sided with the firefighters. The Supreme court determined that “the City chose not to certify the examination results because of the statistical disparity based on race-i.e., how minority candidates had performed when compared to white candidates” (Opinion of the Court, p. 19). In trying to avoid a lawsuit, the City of New Haven ended up getting sued.

The results of the aforementioned lawsuits help to support the need for a set of best practices or standards for exam development. The most commonly used set of standards are *The Standards for Educational and Psychological Testing* (American Educational Research Association, National Council on Measurement in Education, & American Psychological Association, 1999). These standards provide a set of best practices or recommendations for many aspects of both commercial and educational testing (some of the topics covered include exam development, administration, scoring

and equating, score reporting, and fairness to candidates). Like the EEOC Guidelines, *The Standards* (AERA, NCME, & APA, 1999) state that when assessments are going to be used to make decisions about individuals, one must provide evidence of the validity of those decisions. “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of test scores” (AERA, NCME, & APA, 1999, p. 9). In licensure and certification testing, when an examination is used to make decisions about whether or not an individual is competent to perform a job, the relationship between the content covered on the exam and the activities performed on that job or the knowledge necessary to perform those activities (the predictor-criterion relationship) must be documented.

One of the ways in which organizations can provide validity evidence for the decision is to demonstrate the relationship between the examinations used to license or certify the individual and the job in which the individual is licensed or certified (to illustrate the predictor-criterion relationship). This relationship is often documented through the use of a job analysis (Tannenbaum & Wesley, 1993; Kane, 1982; Shimberg, 1981; Smith & Hambleton, 1990). The credentialing organization will begin by conducting a job analysis of the job. Next, the organization will validate the results of the job analysis through the use of a large-scale validation study. The organization will develop an examination blueprint based on the results of the validation study. Finally, an examination will be developed based on the examination blueprint. In developing a credentialing exam in this manner, one is able to provide evidence that the pass/fail decision (to issue or withhold a credential) is valid.

## **Job Analysis for Credentialing Exams**

Job analysis is a process or procedure for analyzing the tasks performed by individuals in an occupation, as well as the knowledge, skills, and abilities required to perform those tasks. Specifically, job analysis can be defined as “any systematic procedure for collecting and analyzing job-related information to meet a particular purpose” (Raymond, 2001, p. 372). Job analysis can be used for multiple purposes including, but not limited to, job description, job classification, job evaluation, performance appraisal, selection, training, worker mobility, workforce planning, efficiency, safety, and legal and quasi-legal requirements (Brannick, Levine, & Morgeson, 2007).

A job analysis is a foundational requirement for any valid credentialing program and helps to identify the core knowledge areas, critical work functions, and/or skills that are common across a representative sampling of current practitioners or job incumbent workers. Empirical results from the job analysis provide examinees and the public the basis of a valid, reliable, fair and realistic assessment that reflects the skills, knowledge, and abilities required for competent job performance.

Within the field of Industrial/Organizational Psychology, there are a number of job analysis methods that are considered quite useful including critical incident technique (Flanagan, 1954), functional job analysis (Fine & Getkate, 1995), the job element method (Primoff, 1988), Position Analysis Questionnaire (McCormick, 1976), and Fleishman’s Ability Requirements Scales (Fleishman, 1988) (Levine, Ash, Hall, & Sistrunk, 1983; Raymond, 2001). However, some these job analysis methods are not applicable for the development of a credentialing examination.

Specifically, the job element method, Position Analysis Questionnaire, and Fleishman's Ability Requirement Scales are not applicable for the development of a credentialing examination. Both the job elements method and Fleishman Scales are avoided because they focus solely on worker attributes (i.e., creativity, attention to detail) rather than the tasks performed on a job, and therefore cannot illustrate the predictor-criterion relationship as well as other job analysis methods. The Position Analysis Questionnaire is a questionnaire used to evaluate jobs on 194 job elements, and is typically avoided because its resulting job description is too general to be used to develop an examination blueprint.

The most common methods of job analysis for the development of a licensure or certification program are the critical incident technique (CIT), Functional Job Analysis (FJA), DACUM (Developing A CurriculUM), and Task Inventory Analysis (Knapp & Knapp, 1995; Raymond, 2001). Although a brief description will be provided for each of the aforementioned job analysis processes, for the purposes of the proposed study only job analyses in which either the Task Inventory Analysis or the DACUM method was implemented will be used in the secondary data analysis.

#### *Critical Incident Technique*

The Critical Incident Technique (CIT) is a job analysis method popularized by Flanagan (Flanagan, 1954). The CIT procedure involves observing and interviewing incumbent workers and developing a task list based on the observations and interviews. Flanagan described the CIT as consisting of "a set of procedures for collecting direct observations of human behavior in such a way as to facilitate their potential usefulness in solving practical problems and developing broad psychological principles" (Flanagan,

1954, p. 327). The goal of CIT is to identify specific incidents of worker behaviors that were particularly effective or ineffective. Through the process of a group interview or questionnaire, a collection of critical incidents is obtained. Those incidents are used to identifying the underlying worker behaviors critical for successful performance on a particular job.

The process of performing the CIT is less formal than other job analysis methods and should be thought of as a set of guidelines rather than a specific structure. The CIT is performed one of two ways. Either a job analyst interviews job incumbents and supervisors, or those same job incumbents and supervisors complete a set of questionnaires developed by job analysts. The incidents that are obtained during the process should include 1) an overall description of the event, 2) the effective or ineffective behavior that was displayed during the event, and 3) the consequences associated with the individual's behavior. The job analyst performing the CIT interview should be familiar with the CIT process, however there is no formal training required of the job analyst.

The interviewer begins by explaining the purpose of the CIT interview. The job analyst should be careful in his or her explanation of the process, and should choose terms carefully. For example, it is sometimes helpful to describe the incidents in terms of "worker behaviors" rather than "critical incidents", as there can be a negative connotation with the term "critical incidents". Again, the analyst directs the incumbent workers and supervisors to describe the incidents in terms of 1) the context or setting in which the incident occurred, including the behavior that led up to the incident; 2) the



specific behavior exhibited by the incumbent worker, and 3) the positive or negative consequences that occurred as a result of the behavior.

Often the interviewees (job incumbents or supervisors) will focus their attention on incidents or worker behaviors that are ineffective rather than those that are effective, as ineffective behaviors are often easier to think of. While this is acceptable, it is also important for the job analyst to ask the participants to describe what the effective behavior would be, had the individual being described performed the job effectively.

Please see the example below from a CIT interview:

*A school librarian found a pair of glasses in his library. One of the students stated that the glasses were hers, and so the teacher gave the pair of glasses to the student claiming that the glasses belonged to her without further questioning. A few days later a parent contacted that school librarian indicating that her son had lost his glasses. The school librarian realized that he had mistakenly given the missing glasses to the wrong student, but couldn't remember which student he had given the glasses to. As a result, the school librarian was forced to pay for a new pair of glasses out of pocket.*

While this incident has the right level of detail and describes a “critical incident”, it is imperative that the school librarian also describe what the effective behavior would be had he performed the job effectively. The job analyst would have no way of knowing the correct behavior without the school librarian providing that information.

A typical CIT interview will generate hundreds of critical incidents (Brannick et al., 2007; Knapp & Knapp, 1995), therefore the next step in the process is to analyze the incidents and organize them in terms of the worker behaviors described during the

process. The job analyst performs a content analysis of the incidents, identifying all of the general behavioral dimensions (i.e., demonstrating a high tolerance for ambiguity) discussed during the job analysis. On average, the incidents can be broken down into five to twelve general behavioral dimensions. Those behavioral dimensions can be used in conjunction with another job analysis to develop an examination blueprint for a credentialing exam.

The CIT is typically used in conjunction with other job analysis methods because its focus is on describing or defining a job in terms of the most “critical” job elements, rather than describing a job in its entirety. As SMEs tend to describe jobs in terms of the job tasks that are most frequently performed instead of focusing on job tasks that are most critical, CIT is useful in obtaining critical job tasks and the underlying worker behaviors that may be missed by other, more holistic job analysis methods. The list of behavioral dimensions and job tasks derived from the CIT may not be a complete picture of the job as most jobs require many worker behaviors for job tasks that are routinely performed, but not considered “critical”.

A potential downside to CIT is that it may be highly labor intensive. It may take many observations and interviews to produce enough incidents to fully describe all of the “critical” behaviors. And, it is possible to miss mundane tasks using critical incidents. However, CIT is a useful addition to any holistic job analysis as it may identify those tasks and underlying worker behaviors that are rarely performed but critical to a job. In many instances, it is those underlying worker behaviors that are most “critical” that are used for the development of a credentialing exam.

### *Functional Job Analysis*

Functional Job Analysis (FJA, Fine & Cronshaw, 1999; Fine & Getkate, 1995) is another popular job analysis process used in the development of credentialing exams. FJA was first introduced by the United States Employment Service and Department of Labor. It was used by these government agencies to classify jobs into categories using a standardized format, resulting in the *Dictionary of Occupational Titles*. Sidney Fine has published several books and articles (Fine, 1988; Fine & Cronshaw, 1999; Fine & Getkate, 1995) describing an updated version of FJA. The FJA process that is described in this chapter is based on Fine's description of FJA, rather than the Department of Labor's description.

FJA begins with the job analyst gathering information about the job in order to determine the purpose and goal of the job. The job analyst should use multiple sources to gain information about the job so that the analyst has a clear understanding of the job prior to beginning the second stage, the interview process. The job analyst must have a very clear understanding of the job because unlike with other job analysis methods, the job analyst will be generating the task statements (in many cases the SMEs generate task statements themselves).

Next, the job analyst collects data about the job from the job incumbents. Typically, data are collected by seating a panel of SMEs or job incumbents and asking them to describe the tasks that they perform on the job. Although Fine and Cronshaw (1999) argued that data should be collected during these focus group meetings, data can also be obtained through observations and interviews of job incumbents in addition to, or

in place of, a focus group meeting. The role of the job analyst is to turn the descriptions provided by the group of SMEs into task statements.

Like many job analyses, FJA requires a very specific structure for formulating task statements. Each task statement should contain the following five elements: 1) the action performed; 2) the object or person on which the action is performed; 3) the purpose or product of the action; 4) the tools and equipment required to complete the action; and 5) whether the task is prescribed or at the discretion of the worker (Raymond, 2001). For example, a sample task statement for a cosmetologist might be “Apply premixed hair color to client’s hair using color applicator to obtain clients’ desired final color”. The task statements generated during FJA are longer than the task statements generated by other job analysis processes (i.e., DACUM and Task Inventory Analysis).

Once the job analyst has created the set of task statements, the SMEs review and rate the task statements. The task statements created by the job analyst are evaluated for level of complexity in terms of how they function related to three entities: people, data, and things. In FJA, *people* are exactly what we would normally think of as people, but also include animals. *Data* are numbers, symbols, and other narrative information. Finally, *things* refer to tangible objects that one interacts with on the job. Thinking about the cosmetologist example, the task “Apply premixed hair color to client’s hair using color applicator to obtain clients’ desired final color” may have a high rating with “people” and “things”, but a very low rating with “data”.

In addition to levels of complexity for data, people, and things, FJA provides worker-oriented descriptors as well. Other characteristics include language development,

mathematics development, and reasoning development (Brannick et al., 2007; Raymond, 2001). The physical strength associated with each task may also be evaluated.

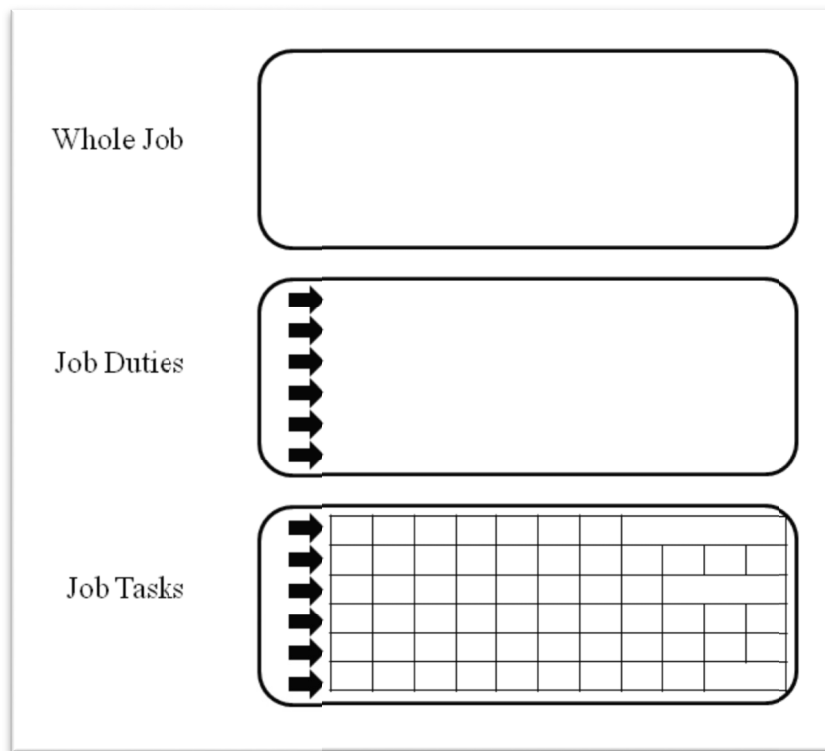
Like all job analysis methods, FJA has its strengths and weaknesses. A significant strength *and* weakness of FJA is the specific way in which task statements are structured. The structure provides an extremely clear and concise description of a task – what the worker does, how it is done, and for what purpose. However, it is not easy to write proper task statements according to the FJA structure (Fine speculated as much as six months of supervised experience is needed for proficiency). Also, the cost associated with hiring a job analyst who has an extensive background in FJA may be a deterrent for some organizations. Another weakness of FJA is that it may be overly complex and detailed for the use in developing a credentialing exam (Knapp & Knapp, 1995; Raymond, 2001).

### *DACUM*

DACUM is a systematic, group consensus job analysis method used to generate task lists associated with an occupation or job (Norton, 2008, Rayner & Hermann, 1988). DACUM is an acronym for Developing A CurriculUM, and is based on three principles. The first principle is that job incumbents know their job better than anyone else, and therefore they are the best at describing what it is that they do. Many job analysis methods use both job incumbents and supervisors (e.g., functional job analysis, critical incident technique), but the DACUM process uses only job incumbents. Second, the best way to define a job is by describing the specific tasks that are performed on the job. Third, all tasks performed on a job require the use of knowledge, skills, and abilities (KSAs) that enable successful performance of those tasks. Unlike other job analysis

methods, DACUM clearly documents the relationship between each task and the underlying KSAs.

In its most basic form, the DACUM process consists of a workshop or focus group meeting facilitated by a trained DACUM facilitator leading five to twelve job incumbents or subject matter experts (SMEs). The primary outcome of the workshop is a DACUM chart, which is a detailed graphic representation of the job, as illustrated in Figure 3. The DACUM chart divides the whole job into duties, and then further divides the duties into tasks. Each task is associated with one or more KSAs.



*Figure 3.* Illustration of a DACUM Chart.

The DACUM process begins with the selection of the focus group panel. A working definition of the job or occupation to be analyzed is created or obtained, and that definition is used to aid in choosing panel members. The panel members should be full-time employees representative of those who work in the job or occupation. Whenever

possible, SMEs selected to participate in the DACUM process should be effective communicators, team players, open-minded, demographically representative, and willing to devote their full commitment to the process (Norton, 2008). SMEs who are not able to participate in the entire process from start to finish should not be included in the DACUM panel, as building consensus among all of the panel members is critical to the DACUM process.

Following selection of the DACUM panel, the actual workshop is typically a two-day focus group meeting. The workshop begins with an orientation to the DACUM process during which time the facilitator provides a description of the process. Upon completion of the orientation, the facilitator leads the group in the development of the DACUM chart. The SMEs are first asked to describe their job overall, and then break their job down into overarching areas of work or “duties”. Duties are general statements of work, representing a cluster of related job tasks. Duties can usually stand alone – they are meaningful without reference to the job itself. The reader should be able to understand the duty clearly without additional reference. For example, *Prepare Family Meals* may be a duty for the job of a homemaker.

Once all of the job duties have been identified, each duty is further divided into tasks. Tasks represent the smallest unit of activity with a meaningful outcome. They are assignable units of work, and can be observed or measured by another person. Job tasks have a defined beginning and end, and can be performed during a short period of time. They often result in a product, service, or decision. All tasks have two or more steps associated with them, so in defining job tasks, if the SMEs are not able to identify at least two steps for each task, then it is likely that the task in question is not really a task at all,

but rather a step in another task. Lastly, job tasks are not dependent on the duty or on other tasks. Thinking about the previous example, *Bake Chocolate Cake*, *Cook Breakfast*, and *Make Lunch* may all be tasks that fall within the duty of *Preparing Family Meals*. Each of these tasks have two or more steps in them (*Bake Chocolate Cake* may require one to *Preheat the Oven*, *Obtain the Ingredients*, *Mix the Ingredients*, *Grease Cake Pan*, and *Set Oven Timer*). And each of the tasks listed can be performed independently of the other tasks in the overall duty area – one does not need to *Bake Chocolate Cake* in order to *Cook Breakfast*.

Finally, the associated KSAs are described for each task. In addition to the KSAs required for successful performance of each task, a list of tools, equipment, supplies, and materials is also created for each of the tasks. The facilitator proceeds through each of the tasks one-by-one, asking the panel what *enablers* are required for successful performance of the task. There should be a direct relationship between the task and the enablers so that each task has an associated set of enablers. Such a procedure is intended to document KSAs that are required for each task rather than those that are beneficial, but not required.

Upon completion of the workshop, the job analyst or facilitator drafts the DACUM chart and distributes the draft to a group of stakeholders for additional feedback. The group that reviews the DACUM chart is comprised of the initial group of SMEs who participated in the focus group meeting, as well as any additional stakeholders. Finally, the DACUM chart is converted into a survey in which the tasks outlined during the focus group meeting are rated based on one or more rating scales.



This last step is called a survey validation study, in which the survey is administered to a larger group of SMEs. More detail about this process is described in the next section.

The DACUM process is different from CIT in that it strives to define *all* of the duties, tasks, and KSAs associated with a specific job, and it relies upon a trained facilitator. It is similar to FJA in that both utilize trained facilitators, and both have specific rules for developing task statements.

One criticism of the DACUM method is that time is spent defining duties, tasks, and KSAs that one would never use in the development of a credentialing exam. For example, to be licensed electrician, one is required to obtain continuing education credits throughout ones career. Because completing continuing education is a required component of the job, the task of *Obtaining Continuing Education Credit* would be identified along with the KSAs required to perform the task successfully. The task and the KSAs associated with it would be included in the job analysis because it is part of the job, and again, the DACUM process describes *all* of the job. However, it seems unlikely one would include anything related to continuing education credits on a credentialing exam.

#### *Task Inventory Analysis*

The final method discussed in this section and often used for the development of credentialing exams is the Task Inventory Analysis, sometimes referred to as “task inventories”. The United States Air Force (USAF) and other branches of the military formalized the task inventory analysis methodology in the 1950s and 1960s (Christal & Weismuller, 1988). Task inventories have been used extensively for the development of licensure and certification examinations (Gael, 1983; Raymond, 2002; Raymond &

Neustel, 2006). Task inventories can be thought of as a four step process: 1) identifying the tasks performed on a job; 2) preparing a questionnaire including scales selected for the purpose of the analysis; 3) obtaining task ratings through a survey or questionnaire, and 4) analyzing and interpreting survey data.

Like functional job analysis, task inventories begin with a job analyst developing a list of tasks based on multiple sources of information. Sources of information include observations and interviews of job incumbents and supervisors (SMEs), small focus groups with job incumbents and supervisors (SMEs), and any written descriptions of the job. Also like FJA and DACUM, the task statements used in task inventories follow a specific format.

The format for writing a task statement begins with a verb or action, followed by the object on which the action is being performed. For example, a task statement might be to “bake cookies”, whereby “bake” is the verb or action and “cookies” is the object on which the action is being performed. Task statements often include a qualifier to describe extra information essential to the task, however task inventories do not require the use of a qualifier. Thinking about the previous example, one might update the task statement to “bake chocolate chip cookies”. In this case, the type of cookie is a qualifier. It describes extra information essential to the task. Baking a chocolate chip cookie has a different set of steps than baking a peanut butter cookie.

Compared to FJA, the task statements in task inventory analysis are shorter and more succinct. Such tasks tend to be narrower in scope than in FJA. For this reason, there tend to be many more tasks in the task inventory approach than in functional job

analysis. A typical task inventory process will produce between 100 and 250 tasks (Brannick et al., 2007; Raymond, 2002).

The level of specificity with which task statements are developed can be hard to define. General, overarching task statements should be avoided. Only those tasks with a defined beginning, middle, and end should be included. An example of a task statement that is too broad and overarching for a nurse would be *Provide Patient Care*. While nurses do provide patient care, the task statement is too general, and does not have a defined beginning, middle, and end. On the other hand, task statements that describe discrete physical movements are overly specific. Thinking again about the nurse, a sample task may be *Review the Physician's Order*. The task may further be broken down into picking up the patient's chart and looking at what the physician has ordered, but these steps are too specific as they start to describe the physical movement of the nurse.

As part of the task inventory process, a survey or questionnaire is developed based on the tasks identified during the analysis – this is often referred to as a survey validation study. The survey can be broken into two parts. The first part of the survey asks the respondents to rate each of the tasks based on one or more scales. (A discussion about choice of scales can be found in the next section on survey validation studies.) The second part of the survey is the demographic section. It is important that those who respond to the survey or questionnaire are representative of those who currently perform the job or those who would like to perform the job. Ideally, the survey should include all job incumbents, as the more people that respond to the survey, the more confident one can be in the results. At the end of the survey, most job analysts typically ask survey respondents to report any tasks identified as “missing” from the task list.

The last step in the task inventory analysis process is to analyze the survey data. The job analyst should verify that a representative sample of job incumbents was obtained. If a sub-group of job incumbents is missing, then the survey should be administered again using quota sampling to ensure that the missing sub-group is included in the second administration. For example, if one of the demographic questions assessed the number of years respondents had been working in the industry, and we found that all of the survey respondents had been working in the industry for a long period of time (20-30 years), we would want to re-administer the survey and target a specific population – in this case, those who have worked in the industry for a shorter period of time.

Once a representative sample of job incumbents has responded to the survey, the task ratings should be analyzed. Typically, means and standard deviations are calculated. Those tasks that received low ratings on one or more of the scales should be reviewed further by the job analyst and a group of SMEs. It is possible that those tasks that received low ratings do not belong on the final job analysis. In addition to reviewing those tasks that received low ratings, tasks that had a high standard deviation should be reviewed. It is possible that job incumbents with specific demographics perform tasks differently than those with other demographics. For example, job incumbents who have been performing a job for 20 years may skip over some tasks that new job incumbents perform often. Or those that are new to the job may not have a good grasp of which tasks are more or less important than others which again lead to variability in task ratings. For these reasons, all tasks that have high standard deviations should be further reviewed by a group of SMEs.

There are two main limitations of task inventories. First, the KSAs required to perform each task are not identified. Job analysts trying to describe jobs that are highly analytical and less vocational will be at a disadvantage when using task inventory analysis. For example, it may be very difficult to ask a playwright to describe his or her job in terms of the specific, observable tasks that he or she may perform. The second limitation to using task inventories is that the rating scales used to evaluate the task statements may be misinterpreted or ambiguous. If survey participants do not have a clear understanding of the rating scales then the resulting survey data analysis will be problematic.

There are two main benefits to using task inventories over other job analysis methods. First, task inventories can be much more efficient in terms of time and cost than other job analysis methods if there are large numbers of incumbents, particularly when the incumbents are geographically dispersed. The job analyst can create the initial list of tasks in a reasonably short period of time, especially considering the simplicity with which the task statements are structured. Then, the time and cost associated with administering and analyzing a survey is relatively small. The entire job analysis process can be completed in a shorter period of time than it might take the same job analyst to perform some other type of job analysis.

The second benefit to using a task inventory analysis over other job analysis methods is that the results lend themselves to the development of an examination blueprint. The quantitative task ratings may be easily converted to test weights. Those tasks that are rated the highest may receive the highest overall weighting on the examination blueprint, whereas those tasks that received low ratings or high standard

deviations may receive little or no weighting on an exam. This is discussed in more detail in the next section.

### **Survey Validation Studies for Job Analyses**

When conducting job analyses utilizing either the DACUM or task inventory analysis method, the last step in the process is to conduct a survey validation study and perform an analysis of the results of the survey (Nelson, 1994; Raymond, 2005). The purpose of the survey validation study is to *validate* the results of the job task analysis. In a DACUM job analysis, the task list is generated from a single focus group meeting – with a small group of SMEs. In a task inventory analysis, the task list is generated from literature reviews, a focus group, observations, and/or small group interviews. In both cases, the task list is derived from the opinions of a small group of people. To be more confident in the results from the job analysis, the task list is converted into a large-scale survey called a survey validation study.

Survey validation studies are typically administered via computer using a web-based survey tool, but can also be administered via paper-and-pencil. The survey is administered to a larger group of people – usually those who either have the credential for which the job analysis is being performed, or those who might seek to obtain the credential for which the job analysis is being performed. For example, if the original purpose of the job analysis is to develop a new credential, then there won't be anyone who is currently credentialed in the field. In that case, the survey would be administered to everyone who has the *potential* of obtaining the credential. If, on the other hand, the purpose of the job analysis is to revalidate an already existing credential, then the target population for the survey validation study would be those that are currently credentialed.

### *Collecting Survey Respondent Demographic Information*

Regardless of who the survey is administered to, like in a task inventory analysis, a survey validation study can be thought of as having two components. First, there is a demographic section in which demographic information about the survey respondents is collected. This section often includes between 10 and 20 questions (Raymond, 2005). Some examples of demographic questions are listed below:

1. In which location do you work?
2. How long have you worked in your field/profession?
3. What is your highest level of education?
4. Within your profession, in which specific area do you work?
5. How much experience do you have in specific work areas?

These questions are asked of the survey respondents to ensure that a representative sample of participants respond to the survey. In generating demographic questions, the job analyst typically asks SMEs to identify all of the demographic areas that might lead to high amounts of variability in task ratings. For example, in some jobs those that work in one region of the United States might describe their job differently than those that work in a different region of the United States causing survey respondents in one region to rate a task differently than survey respondents in the other region. In this case, it is imperative that survey respondents from all regions in the United States respond to the survey and provide ratings for each of the tasks initially identified.

### *Task Rating Scales*

The second component of the survey validation study is rating actual task statements. Survey participants are asked to rate each task on one or more scales

(Raymond, 2005). Determining the rating scales that should be used in a survey validation study is the most important step in the development of the survey. As Raymond illustrates, “developing task inventory questionnaires is mostly about determining the questions to be asked and designing rating scales for eliciting responses to those questions” (2005, p.30). The individual rating scales that are chosen and the way those rating scales are explained are critical decisions for the job analyst. The more rating scales that are included in a survey, the greater the time required of the survey respondent, and the greater the cost to the survey administrator.

A list of the most common rating scales for survey validation studies, along with definitions of those rating scales, and whether they are absolute or relative scales is presented in Table 1 (Knapp & Knapp, 1995; Manson, Levine, & Brannick, 2000; Raymond, 2001; Raymond, 2005; Sanchez & Fraser, 1992; Sanchez & Levine, 1989).

Although all of the rating scales listed in Table 1 are commonly used in survey validation studies, some of the rating scales are preferred over others when the purpose of the job analysis and survey validation study is to develop a credentialing exam. The most frequently used rating scales when the purpose is to develop a credentialing exam are task importance, task frequency, criticality or consequence of error, and need at entry (Newman, Slaughter, & Taranath, 1999; Manson, Levine, & Brannick, 2000; Raymond, 2002; Raymond, 2005; Knapp & Knapp, 1995).

Task importance has continuously been considered a crucial scale for inclusion on survey validation studies. The *Standards* state “the content domain to be covered by a credentialing test should be defined clearly and justified in terms of the importance of the content for credential-worthy performance in an occupation or profession” (AERA et al.,



1999, p. 161). Kane (1982) also argued that more emphasis should be placed on tasks that are considered most important. He uses the example of a physician's licensing exam, in which more emphasis should be placed on treating concussions than the common cold. While a physician may treat more patients with the common cold than those with concussions, the consequences of not treating a concussion correctly are more dire than the consequences of not treating the common cold correctly, thus knowing how to treat concussions is of greater importance. Lastly, Tannenbaum and Wesley (1993) discuss task importance as being the single scale that should be included in a survey validation study, expressing that "elements of the content domain confirmed to be important are considered eligible for inclusion in the development of the licensure test" (p. 975).

Task frequency is also considered an important scale for inclusion on survey validation studies, as those tasks that are performed most frequently should be included in a job analysis and should have a higher weight on a resulting examination blueprint (Newman, et al., 1999; Raymond, 2001; Raymond, 2005). Kane (1982) provided another example related to task frequency and physician's licensing exams. In his example, Kane argues that a greater emphasis should be placed on heart disease, diabetes, and cancer, than should be placed on tropical diseases. Even though tropical diseases can be as deadly as heart disease, there is a much lower incidence of tropical diseases in the United States. A greater emphasis should be placed on treating those diseases that are encountered more frequently, over those diseases that are encountered less frequently.

From a legal standpoint, task criticality or consequence of error may be the most crucial rating scale to include when the purpose of the job analysis is to develop a licensure or certification exam. This point is best illustrated below.

The purpose of licensing, as noted earlier, is to protect the public health, safety, and welfare. For this reason, tests used for licensing must be able to help identify those who possess the knowledge, skills, and abilities to perform *critical tasks* [emphasis added] in a manner that will adequately safeguard the public health, safety, and welfare. (Shimberg, 1981, p. 1140).

The same is true for certification testing. Those tasks and underlying knowledge, skills, and abilities that are *critical* to job performance should be identified and included on a certification exam. For example, when thinking about food safety professionals, it is crucial to identify those tasks that are most critical to public health so that a greater emphasis may be placed on those tasks when developing a credentialing exam for food safety professionals.

Lastly, a need at entry scale is frequently used to rate tasks. This is especially important when the purpose of the job analysis is to develop a credentialing exam (Raymond, 2001). In a typical task inventory analysis or DACUM job analysis, a group of SMEs identify all of the tasks that are typically performed on a job and the underlying knowledge, skills and abilities required to perform those tasks. The SMEs that identify the requisite tasks have often been working in the field for a variety of years – the meeting participants would never be made up of all entry level practitioners. However, when one is developing a credentialing exam, it is important to know what tasks, and underlying knowledge, skills, and abilities, are actually required at entry into the profession, or required at that initial point of licensure, and what tasks are typically learned later in a career. Tasks and subsequent knowledge, skills, and abilities that are learned or mastered at a later point in time do not belong on a credentialing exam. For

example, a job analysis for a Building Operator might include a task called “Create an Annual Budget”. However, one wouldn’t expect that a newly credentialed Building Operator would be able to perform such a task, as this is a task that is typically mastered while on the job.

Table 1.  
*A Description of Rating Scales Used for Survey Validation Studies*

Task Rating Scale	Description of Rating Scale	Absolute or Relative
Criticality or consequence of error	The risk or adverse consequence of not performing the task correctly or not at all.	Relative rating scale
Difficulty to learn	The amount of time or effort that is required to learn how to perform the task.	Relative rating scale
Level of responsibility	Whether or not the person rating the task is personally responsible for performing the task, and if so, his or her level of responsibility.	Relative rating scale
Need at entry	The extent to which an entry-level individual should be able to perform the task.	Absolute rating scale
Task complexity or difficulty	The difficulty or complexity of the task.	Relative rating scale
Task frequency	The frequency with which the task is performed.	Absolute or relative rating scale
Task importance	The relative importance of knowing how to or being able to perform a task.	Relative rating scale
Time spent	The amount of time spent performing the task, usually described as the amount of time spent during a typical workday.	Absolute or relative rating scale

Considering the varying types of task rating scales, the job analyst must choose one or more rating scales to use for a survey validation study. And if more than one rating scale is chosen, the job analyst must decide how the rating scales will be combined. As previously mentioned, there is little research on the different types of task rating scales used for survey validation studies or how to combine those rating scales. And the research that is available is often times conflicting.

### *Relationships of Task Rating Scales*

Most of the published literature on the types of rating scales used for survey validation studies, as well as how rating scales should be combined, is anecdotal. There are only three empirical research studies examining the relationship between rating scales used for survey validation studies for job analyses. The three studies were published between 1989 and 1992, and are discussed in more detail below.

Sanchez and Levine (1989) published one of the first research studies on this topic. In it, Sanchez and Levine administered task inventory surveys to 60 incumbents spread across four different jobs: community services officer (CSO), engineering technician, librarian, and police officer. Each incumbent rated tasks related to his or her job on a total of six task rating scales. The number of tasks rated ranged from 19 tasks (librarians) to 109 tasks (engineering technicians). The rating scales used (in the order they were presented) were: time spent, task difficulty, task criticality, task responsibility, difficulty of learning the task, and overall task importance.

Sanchez and Levine evaluated the relationship between (a) overall task importance ratings, (b) relative time spent ratings, (c) a composite of task criticality times task difficulty plus relative time spent [criticality\*difficulty + relative time spent], and (d) task criticality ratings plus difficulty of learning ratings divided by two [(criticality + difficulty of learning)/2]. The relationships between scales were evaluated by computing the correlation between scales and by using multiple regression to determine how much of the variability in the overall task importance rating could be explained by the remaining five rating scales.

There were several significant findings from this study. First, they found that overall task importance was highly correlated with both task criticality and difficulty of learning. The correlations between overall importance and task criticality ranged from .78 to .90, while the correlations between overall importance and difficulty of learning ranged from .62 to .75. They also found that both task criticality and difficulty of learning were the best predictors of overall task importance. Task criticality was found to be a significant predictor of overall task importance for all four jobs ( $p < .05$ ) and difficulty of learning was found to be a significant predictor of overall task importance for two out of the four jobs ( $p < .05$ ).

Second, Sanchez and Levine found the task difficulty rating scale and the difficulty of learning rating scale to be highly correlated ( $r$ s ranged from .66 to .91). This finding suggests a degree of redundancy between task difficulty and difficulty of learning rating scales.

Third, the composite rating formed by task criticality and difficulty of learning (the second composite described above), produced the highest interrater reliability scores. Sanchez and Levine recommend that a composite of two or more simplified tasks would produce more reliable task ratings than a single holistic rating scale, and that the combination of task criticality and difficulty of learning should be used for the jobs mentioned in this study and perhaps other jobs.

This study had several limitations. First, the sample size was small. Although survey validation studies for job analyses conducted as part of developing a credentialing exam may have small sample sizes, the sample sizes in this study were so low that it is hard to have confidence in the findings. Sample sizes in this study ranged from five

(police officers) to 27 (engineering technicians). Second, each participant was asked to rate their job tasks using six different task rating scales. In reality, if job incumbents were asked to rate each job task using six different rating scales, the attrition rate would likely be too high to have confidence in the results. Third, the order with which the rating scales were presented was not varied. All participants saw all six scales in the same order. Because this study was analyzing job analysis rating scale data for four professions, it would have been ideal if the presentation order of the six rating scales was varied.

In a follow-up study, Sanchez and Fraser (1992) administered job analysis surveys to 101 incumbents from 25 different jobs in the service industry. The number of tasks rated ranged from 14 to 78, with a median of 34. Survey respondents rated each task on four individual rating scales: (a) relative time spent; (b) difficulty of learning the task; (c) criticality, or consequence of error; and (d) overall importance. Survey respondents were asked to rate each task on each of the four scales before moving onto the next task.

In addition to evaluating the relationship or correlation between each of the individual scales, Sanchez and Fraser also evaluated the relationship between the overall importance rating scale and a set of composites. The overall task importance ratings were compared to the following composites: (a) difficulty of learning times criticality plus relative time spent, (b) criticality plus difficulty divided by two, (c) relative time spent times task importance, and (d) task criticality times relative time spent.

When evaluating the relationship between individual rating scales, Sanchez and Fraser found that the ratings of task criticality and overall importance were highly

correlated ( $r$ s ranged from .60 to .99), and thus somewhat redundant. When evaluating the relationship between overall task importance and the four composites, Sanchez and Fraser found that “the choice of composite is not likely to alter the final rank ordering of tasks to a large extent. However, the inclusion of difficulty of learning and criticality in the composite may provide the best prediction of average task importance across SMEs” (p. 552). Overall, Sanchez and Fraser argued that if time was a concern for job analysts, an overall task importance rating might provide results that are comparable to those obtained by a composite rating.

Again, this study had several limitations. Like Sanchez and Levine (1989), Sanchez and Fraser (1992) had small a sample size. The overall sample size for the study was 101, spread across 25 different jobs. The number of respondents for each of the 25 jobs ranged from a low of one to a high of 13, which is far too few survey participants considered acceptable for the development of a credentialing exam. Second, presentation order was not varied. Survey respondents rated each task on all scales before moving onto the next task. Task rating scales were not varied in any way. Finally, this study was performed for the service industry. One might wonder if the same results would be found if the survey were repeated in a different industry.

In 1990, Friedman conducted a similar study in which the redundancy between three task ratings was analyzed. A validation survey for a research and development (R&D) manager task inventory analysis was administered to R&D managers from nine organizations. The survey consisted of 244 tasks. The 117 respondents rated each task on three scales: 1) relative time-spent, 2) importance, and 3) frequency. All three scales were seven point scales, in which a “1” indicated the least amount of time, the least

important, or the least frequently performed and “7” indicated the most amount of time, the most important, or the most frequently performed. Survey respondents rated all tasks on one rating scale before moving onto the next task.

Friedman evaluated both the correlations between task ratings on pairs of scales and the absolute-value differences on each task for each pair of scales. By doing the latter, Friedman was able to identify the percentage of tasks that were rated within one or two points of each other on each pair of scales.

Although, the correlations between task ratings on the three pairs of scales were fairly low (from correlations of .32 to .55), Friedman concluded that time and importance scales were redundant, and that one should choose one or the other when developing a validation survey to rate job tasks. This conclusion was based on the fact that absolute differences between task ratings of importance and relative time-spent were within one point roughly 70% of the time, and were within two points of each other almost 90% of the time (Friedman, 1990). Friedman did not find a strong relationship between time-spent and frequency ratings or importance and frequency ratings. Friedman postulated that the relationship between relative time-spent and importance may be due to the fact that people spend the most time on tasks they consider important, or people view tasks as important if they spent a great deal of time on them.

This study has a few limitations. First, the presentation order was not varied. Survey respondents rated all tasks on one scale before moving onto the next scale, and scale order wasn't varied across the respondents. Second, the study took place in one industry, evaluating research and development managers. One would again be curious to see if the results would be repeated if the study was conducted in a different industry.



Lastly, the finding that there is redundancy between a relative time-spent scale and other scales may not be helpful in that there have been multiple studies cautioning against using relative time-spent scales for task inventories (Wilson & Harvey, 1990; Pass & Robertson, 1980).

In both Sanchez and Levine (1989) and Sanchez and Fraser (1992), survey respondents rating each task on multiple rating scales before moving onto the next task. When survey respondents are presented with one task at a time, and asked to rate each task on multiple rating scales at once, the survey appears to be shorter which is why many organizations tend to use this presentation order. If scales are presented next to each other, the correlation between the two scales may be inflated. For example, if survey respondents are asked to rate the importance of a task followed by the frequency with which that task is performed before moving onto the next task, survey respondents are more likely to respond similarly to both rating scales. If survey respondents are asked to think about their job as a whole and consider how important each task is to successful performance of the job before moving onto how frequently each task is performed, the correlation between the two scales is assumed to be lower. For this reason, it is critical that survey validation studies included in this study represent both models (survey respondents rate all tasks on one scale before moving onto the next scale and survey respondents rate each task on all scales before moving onto the next task).

#### *Survey Design Related to Scale Placement*

At this point, it might be helpful to address some of the survey design issues faced by job analysts administering survey validation studies. As previously mentioned, presentation order is of huge concern in developing a validation survey. Dillman, Smyth,

and Christian (2009) state that asking respondents to rate two or more scales at once causes a cognitive challenge. Whenever possible, survey respondents should be presented with all of the tasks and asked to rate each task on one rating scale before moving onto subsequent rating scales. This basic survey design principle is often ignored when developing survey validation studies, as the perceived length of the survey is of greater concern to job analysts.

A study by Funke, Reips, and Thomas (2011), found that survey respondents had a statistically significantly higher drop-off rate when taking a web-based survey with slider rating scales than a survey with radio buttons. Additionally, Funke et al. found that the response time of slider scale items is significantly longer than the response time of items utilizing radio buttons. This is an important finding as when survey respondents are asked to rate all tasks on one scale before moving onto the next scale in an online survey, radio buttons are typically used, as illustrated in Figure 4. When survey respondents are asked to rate one task at a time across all scales in an online survey, drop-down menus are used to rate scales, as illustrated in Figure 5. While the scale presentation illustrated in Figure 5 not the same as a traditional slider scale used in web-based survey design, the two scales are similar.

This finding illustrates that there are significant differences in response rates and drop-off rates of surveys based on the structure of the survey. However, most organizations continue to present task rating scales in survey validation surveys as illustrated in Figure 5 because it makes the survey *appear* shorter in length, when in actuality it might take longer to complete.

**Please rate these tasks.**

	Very important	Important	Somewhat important	Not important
Task 1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Task 2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Task 3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Please rate these tasks.**

	Always performed	Performed often	Seldom performed	Never performed
Task 1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Task 2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Task 3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

*Figure 4.* Scales presented one at a time, in which survey respondents rate all tasks on one scale before moving onto the next scale.

**Please rate these tasks.**

	Frequency	Importance
Task 1	<input type="text"/>	<input type="text"/>
Task 2	<input type="text"/>	<input type="text"/>

Always performed  
Performed often  
Performed sometimes  
Never performed

*Figure 5.* Scales presented together, in which survey respondents rate all tasks on one scale before moving onto the next task.

Job analysts face another issue when trying to decide whether the scale ratings should be presented all on one page or split across multiple pages. When the presentation order is design so that a survey respondent rates all tasks on one scale before moving onto another, different scales are usually presented on different pages in the survey. For example, in an online job analysis validation survey, participants might be asked to rate the frequency with which they perform all tasks on one page, and then how important those tasks are on a subsequent page. When survey respondents are asked to rate one

tasks at a time across all rating scales, as illustrated in Figure 5, both scales are inevitably presented on the same page of the survey.

Tourangeau, Couper, and Conrad (2004) found that the correlation between two or more items was increased when presented on the same page or screen in a survey, as survey respondents perceived that two items presented next to each other were related. This finding is important for this study, as the correlation between two scales may be inflated based on presentation order. By evaluating the relationship between two scales when they are separated (presented on separate pages of a survey), one might be able to state with greater confidence that the relationship between those two scales is based on other factors not related to presentation order.

#### *Data Analysis and the Development of Examination Blueprints*

Unlike the disagreement in the choice of rating scales used for survey validation studies of task inventories, there is little disagreement as to how the resulting data should be analyzed and how the subsequent exam blueprint should be created. Raymond (1996) illustrated a common method for establishing examination blueprint weights. The general method described by Raymond is to 1) combine two or more rating scales into a single composite scale (the subsequent steps will be the same regardless of how the scales were combined), 2) determine an average task rating for each task based on the composite scale, 3) sum all of the averaged task ratings, 4) divide the average task rating for each task by the sum of all average task ratings, and then 5) multiply that number by 100 to arrive at a percentage for each task. It is important to note that when only one rating scale is used, the last three steps outlined above will still apply with the single rating scale.

Kane, Kingsbury, Colton, and Estes (1989) also describe a method for obtaining examination blueprint weights. The method that is used to establish exam blueprint weights by Kane et al. is virtually identical to that described by Raymond (1996). The only difference between the method described by Raymond and the method described by Kane et al. is in the first step. In Raymond's explanation on how to establish examination blueprint weights, he describes combining frequency and criticality scales by doubling criticality and adding it to frequency to obtain an overall importance rating (2\*criticality + frequency = overall importance). Kane et al. recommend a multiplicative model for combining criticality and frequency ratings for an overall importance rating (criticality\*frequency = overall importance).

Regardless of the composite used, the relationship between the composite and its constituents is standardized, as illustrated by Ghiselli, Campbell, and Zedeck (1981). The correlation between a composite and its constituents can be defined by Equation 1.

$$r_{z_1c_z} = \frac{1 + (k - 1)\bar{r}_{1i}}{\sqrt{k + k(k - 1)\bar{r}_{uv}}} \quad (1)$$

In the formula,  $\bar{r}_{1i}$  is equal to the average correlation of the single constituent with each of the other constituents, whereas  $\bar{r}_{uv}$  is the average of the coefficients of correlation among all the constituents in the composite. This finding has important implications for this study. As the correlations will be computed for each pair of rating scales for each survey validation study, Equation 1 can be used to determine the correlation between a composite and its constituents.

### **The Use of Correlations in Meta-Analysis Research**

One of the greatest limitations to the studies that have been described is the lack of variability in study characteristics. In each of the aforementioned studies, the

relationship between task rating scales was analyzed within some predefined context – for example, taking place in one industry or looking at tasks in a fixed presentation order. While each of these studies has value, they are limited in that they are not generalizable to varying contexts. According to Borenstein, “we live in a world where the utility of almost any intervention will be tested repeatedly, and that rather than looking at any theory in isolation, we need to look at the body of evidence” (2009, p. xxi). By using meta-analytic techniques, and evaluating the relationship between task rating scales across multiple contexts, one will be able to generalize the results of this study with more confidence.

Meta-analytic techniques will be incorporated into this study by using the correlations derived from sets of scales as effect sizes for a meta-analysis. The concept of using correlations as effect sizes has been discussed in many texts (Borenstein , 2009; Lipsey & Wilson, 2001; Hedges & Olkin, 1985). According to Lipsey and Wilson, “the correlation coefficient is already a standardized index and therefore is useable as a meta-analytic effect size statistic in its raw form even if the variables being correlated are differently operationalized” (2001, p. 63).

The method for using the correlation coefficient as an effect size is outlined below. First, the correlations are transformed using Fisher’s  $Z_r$ -transformation, as illustrated in the Equation 2, with the variance of the  $Z_r$ -transformed correlation illustrated in Equation 3.

$$ES_{Z_r} = .5 \log_e \left[ \frac{1+r}{1-r} \right], \text{ where } r \text{ is the correlation coefficient} \quad (2)$$

$$\omega_{Z_r} = \frac{1}{SE_{Z_r}^2} = n - 3 \quad (3)$$

After converting the correlations using Fisher's  $Z_r$ -transformation, the next step is to compute a weighted mean effect size, and a standard error around that mean effect size, as illustrated in the Equations 4 and 5 below.

$$\overline{ES} = \frac{\sum(\omega_i ES_i)}{\sum \omega_i}, \text{ where } ES_i \text{ are the values on the effect size statistics used}$$

and  $\omega_i$  is the inverse variance weight associated with effect size  $i$  (4)

$$SE_{\overline{ES}} = \sqrt{\frac{1}{\sum \omega_i}}, \text{ where } SE_{\overline{ES}} \text{ is the standard error of the effect size mean (5)}$$

The standard error will be used to create a confidence interval around the weighted mean effect size, as illustrated in Equations 6 and 7.

$$\overline{ES}_L = \overline{ES} - z_{(1-\alpha)}(SE_{\overline{ES}}) \quad (6)$$

$$\overline{ES}_U = \overline{ES} + z_{(1-\alpha)}(SE_{\overline{ES}}) \quad (7)$$

Lastly, the  $Z_r$ -transformed correlation will be transformed back to a standard correlation using the Equation 8.

$$r = \frac{e^{2ES_{Zr}} - 1}{e^{2ES_{Zr}} + 1} \quad (8)$$

## CHAPTER THREE: METHODS

### **Purpose**

The purpose of this study was to determine the relationship between individual and composite rating scales; examine how that relationship varies across industries, sample sizes, task presentation order, and number of tasks rated; and evaluate whether examination blueprint weightings would differ based on the choice of scales used. A secondary data analysis was performed using data from survey validation studies from 20 different job or task analyses. The 20 sample studies were from job analyses conducted for eight different professional industries. The sample sizes varied from less than 100 survey respondents to over 1,000 respondents. The relationship between individual and composite task ratings was compared when the scales were rated one scale at a time, as well as when scales were rated all at once, one task at a time (presentation order). The relationships between individual and composite task ratings were compared for small task lists (50 tasks or less), medium task lists (51-100 tasks rated), and large task lists (more than 100 tasks rated). Lastly, sample examination blueprint weights were generated based upon the each individual and composite scale to determine if the examination blueprint weighting would differ based on the choice of scale or composite of scales used to create the blueprint.

### **Research Questions**

There are four overarching research questions for this study:



1. What is the relationship between the different types of individual and composite rating scales?
  - a. What is the relationship between different types of individual rating scales?
  - b. What is the relationship between different types of composite rating scales?
  - c. What is the relationship between different types of individual rating scales with different types of composite scales?
2. To what extent do the relationships of individual and composite rating scales vary across industries?
3. To what extent do the relationships of individual and composite rating scales vary across survey design factors?
  - a. To what extent do the relationships of individual and composite rating scales vary across varying numbers of survey respondents?
  - b. To what extent do the relationships of individual and composite rating scales vary across scale presentation order?
  - c. To what extent do the relationships of individual and composite rating scales vary across the number of tasks rated?
4. To what extent are examination blueprint weightings different based on the choice of scale composites used in the survey validation study?

### **Overview of Research Design**

The secondary data analysis included survey validation data from job or task analyses conducted during a five year period (January 2007 to December 2011). Data

from 20 different surveys were included in the analysis. The range of sample sizes was from a small sample size of 37 to a large sample size of 3,185. Data from eight industries were included in the sample analysis (e.g., accommodation and food services, construction, and healthcare and social assistance). The four rating scales that were included in the study were frequency, importance, criticality or consequence of error, and need at entry. The composites that were used in the analysis are listed below:

- Composite 1 = 2\*Importance + Frequency
- Composite 2 = Criticality\*Frequency
- Composite 3 = 2\*Importance + 2\*Criticality + Frequency
- Composite 4 = 2\*Importance + Frequency + Need at Entry

As previously mentioned, there is no consistent literature on what composite should be used to derive an examination blueprint from survey validation data. The composites outlined above have been used by different psychometric organizations, and were identified by reviewing public job analysis reports.

### **Sample Studies**

A secondary data analysis was performed on a sample of 20 survey validation studies for job analyses in which the task inventory analysis or DACUM method was used. Each job analysis was performed for the purpose of developing (nine studies) or revalidating (11 studies) a licensure (three studies) or certification exam (17 studies), and all took place during a five year period. The sample studies were obtained using convenience sampling, as it is difficult to obtain job analysis data from many organizations because most organizations consider survey validation data confidential. A more detailed breakdown of the 20 sample studies is included in Appendix A.

The process used to obtain the 20 survey validation studies was to contact all of the psychometricians within a business network via email, and ask them to either 1) provide data from a survey validation study, or 2) recommend someone who may have survey validation data. In either case, the only requirement for survey validation data to be used in this study was that each study had to include two or more task rating scales. There are a handful of psychometricians who use only one scale when collecting survey validation data, and often that single scale is a hybrid of two scales, as illustrated in Figure 6. Survey validation studies that had included the single rating scale were not included in this study.

- |                          |
|--------------------------|
| (0) Not Performed        |
| (1) Of No Importance     |
| (2) Of Little Importance |
| (3) Moderately Important |
| (4) Very Important       |
| (5) Extremely Important  |

*Figure 6.* Illustration of single scale used in survey validation studies.

### **Representativeness of Survey Respondents in Sample Studies**

One additional consideration was whether or not survey respondents were representative of the population invited to respond to the survey in the 20 studies included in this analysis. All of the psychometricians who provided studies for this analysis confirmed that the survey respondents were representative of the target population. One way to verify representativeness is to ask the examination committee or stakeholder group to identify a set of demographic questions for survey respondents to answer. The demographic questions should cover every characteristic of the target

population that might influence respondents' task ratings. For example, if there is a concern that people with fewer years of experience may rate tasks differently than people with a lot of experience, then the number of years a respondent has been working in the industry may be of particular concern and should be included as a demographic question in the survey.

The demographic backgrounds of all survey respondents have been provided for two of the studies included in this analysis. The first study in which the demographic background of survey respondents was analyzed was for a Journeyman Plumber licensing exam. There were 100 journeyman plumbers invited to participate in the survey validation study of the job analysis. Of the 100 invited to participate in the survey, 65 responded to the survey. Survey respondents were asked the 11 demographic questions presented below:

1. In which state do you primarily work?
2. Is a state or local Journeyman's license required where you work?
3. If yes, do you have a state or local Journeyman's license?
4. What additional certifications have you obtained?
5. Which plumbing code do you follow?
6. Have you completed a Department of Labor approved apprenticeship program?
7. If not, what type of training have you had?
8. What is your highest level of education?
9. On what type of plumbing installation do you primarily work?
10. How many years have you been working as a Journeyman Plumber?

## 11. How old are you?

Survey respondents reported working in 13 States (note: two survey respondents reported working in “Multiple States”, however it is unknown in which states those respondents were primarily working). The majority of survey respondents (37 or 56.9%) reported that a state or local Journeyman’s license was required where they worked. Twenty-one respondents (21 or 32.3%) reported that a state or local Journeyman’s license was not required where they worked. The remaining respondents were either unsure whether or not a license was required (four respondents) or did not respond to the question (three respondents). Of the 37 who reported that a license was required, 35 of them reported being licensed.

Survey respondents reported having a variety of additional credentials, as illustrated in Figure 7. They also reported following a variety of plumbing codes, as indicated in Figure 8. The majority of survey respondents (54 or 83.1%) reported completing a Department of Labor approved apprenticeship program.

When asked to report their highest level of education, survey respondents had varying levels of education. The largest number of respondents (20 or 30.7%) reported “some college”, as illustrated in Figure 9.

Survey participants were asked report the type of plumbing installation in which they primarily worked. The majority of respondents (41 or 63%) indicated working on commercial installations. The remaining respondents were split evenly between industrial, institutional, residential, service, and other installations.

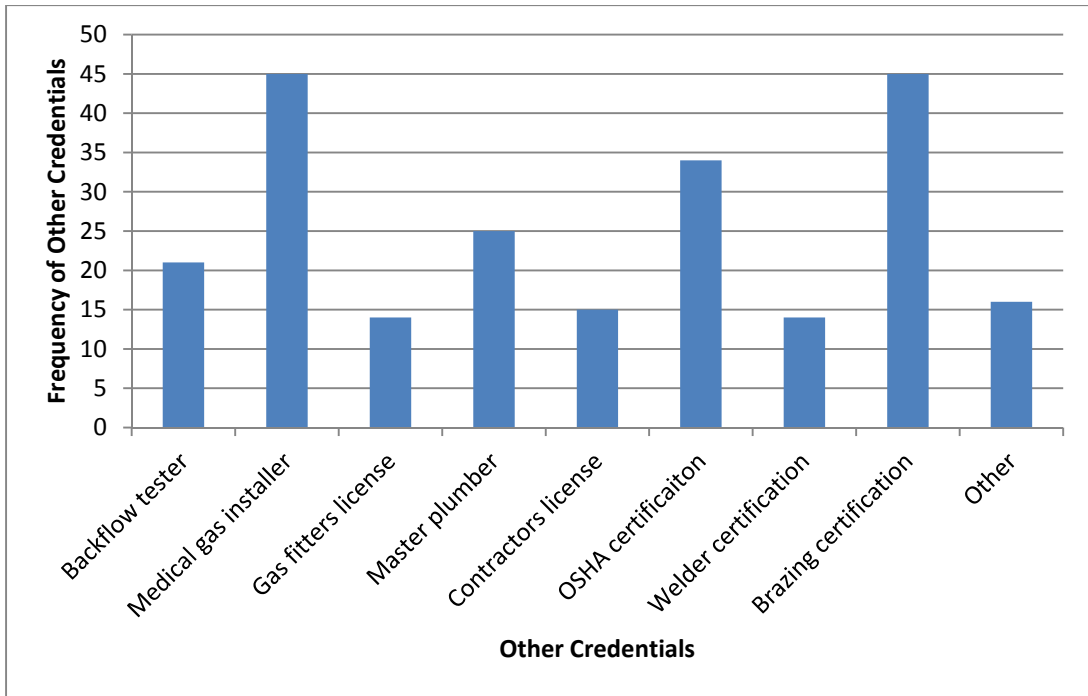


Figure 7. Other credentials obtained by survey respondents in one of the studies included in this analysis.

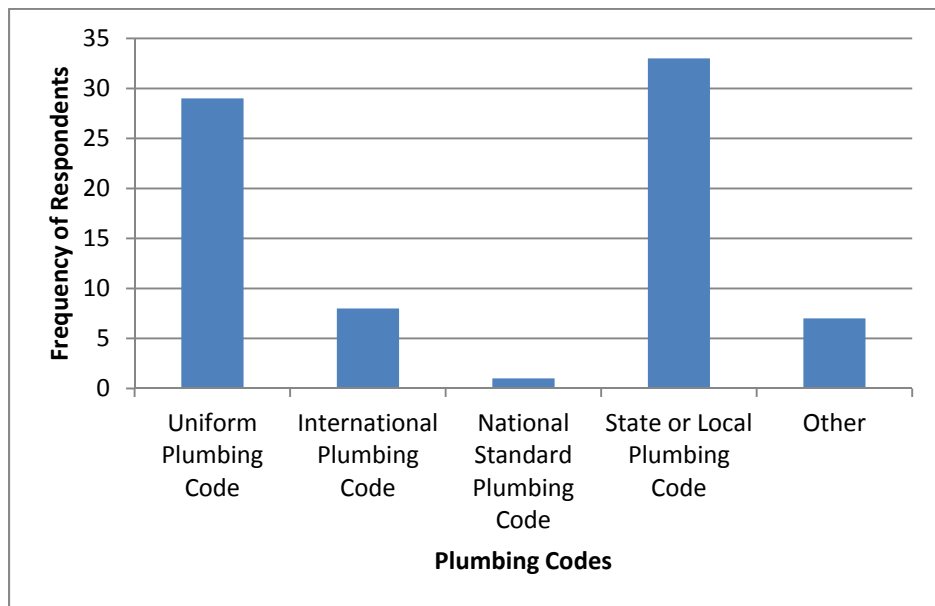


Figure 8. Plumbing codes followed by survey respondents in one of the studies included in this analysis.

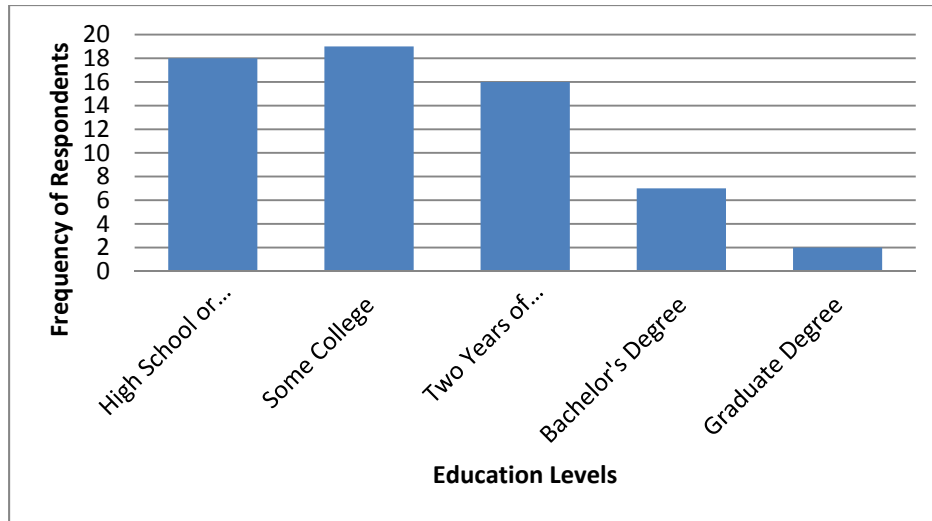


Figure 9. Highest level of education reported by survey respondents in one of the studies included in this analysis.

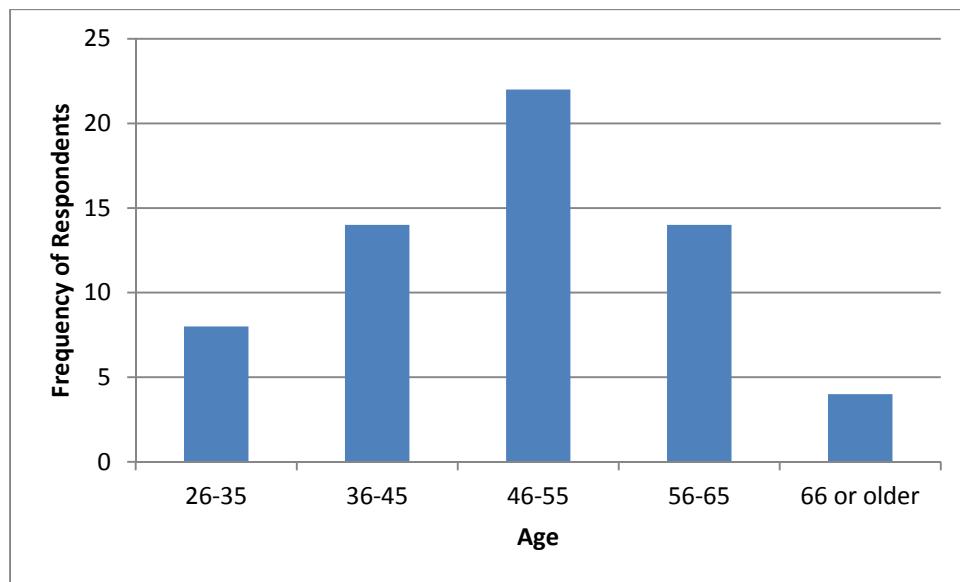
The last two questions in the demographics section referred to the number of years spent working as a Journeyman Plumber and the respondents' age. The majority of respondents reported working as a Journeyman Plumber for 21 or more years (39 or 60%). Respondents' reported age varied from 26-35 years, up to 66 or older, with the largest number of respondents between the ages of 46 and 55 (22 or 35%), as indicated in Figure 10.

It was determined by the organization that conducted this job analysis that the sample of journeyman plumbers who responded to this survey was representative of the population journeyman plumbers who seek to obtain this journeyman plumber license.

The second study in which the demographic background of survey respondents was analyzed was for a Phlebotomy certification exam. There were 1,914 participants invited to respond to the survey validation study. Of the 1,914 invited to participate in the survey, 400 responded to the survey. Each participant was asked 8 demographic questions. The questions are listed below:

1. What is your gender?
2. In which state do you work?
3. In what type of institution do you work?
4. How many physicians work in the Phlebology portion of the practice or group in which you work?
5. How many vascular technologists work in the Phlebology portion of the practice or group in which you work?
6. What is your highest level of education?
7. As a physician, what is your background?
8. How many years have you worked in Phlebology?

Of the 400 participants who responded to the survey, 267 (69.5%) were male, and 117 (30.5%) were female, as illustrated in Figure 11.



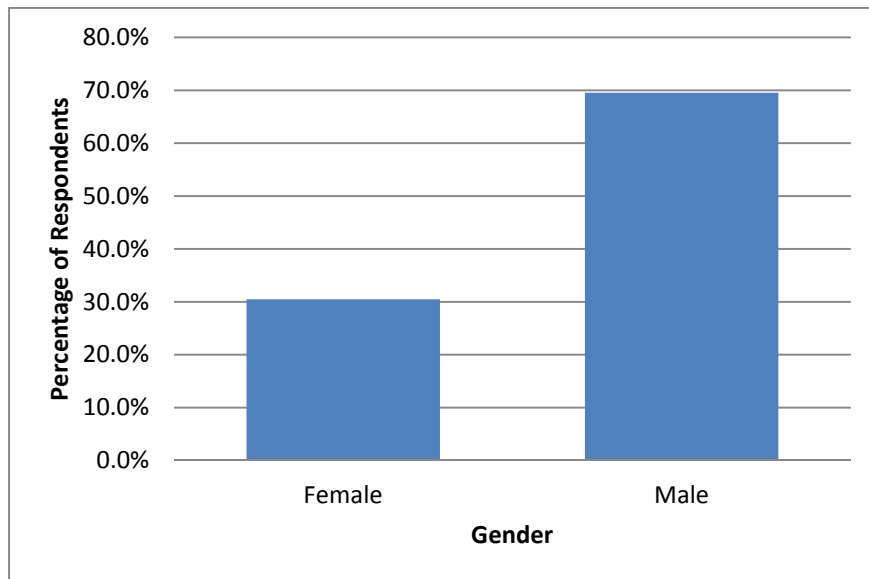
*Figure 10.* Reported age of survey respondents in one of the studies included in this analysis.

Next, survey participants were asked to report the state in which they worked.

Survey respondents reported working in 48 out of the 50 states. The largest number of



respondents reported working in Florida, followed closely by California, Texas, and Indiana, as illustrated in Table 2. No respondents reported working in Delaware or South Dakota. Thirty-eight survey participants did not respond to this item.



*Figure 11.* Reported gender of survey respondents in one of the studies included in this analysis.

Next, survey respondents were asked to report the type of institution in which they worked. The majority of respondents reported working in an individual private practice (194 or 50.1%), with the next largest group of respondents working in a group private practice (118 or 30.5%). Less than one percent of respondents reported working in a government hospital, mobile traveling ultrasound, or tertiary care center, as illustrated in Table 3. Respondents were also given the option of choosing “other” if the institution in which they worked was not represented. Fourteen survey participants did not respond to this item.

Table 2.  
*States in Which Respondents Reported Working*

States	Frequency	Percent
Alabama	3	0.8%
Alaska	1	0.3%
Arizona	12	3.3%
Arkansas	1	0.3%
California	29	8.0%
Colorado	9	2.5%
Connecticut	5	1.4%
Florida	30	8.3%
Georgia	13	3.6%
Hawaii	2	0.6%
Idaho	2	0.6%
Illinois	16	4.4%
Indiana	25	6.9%
Iowa	3	0.8%
Kansas	2	0.6%
Kentucky	1	0.3%
Louisiana	1	0.3%
Maine	2	0.6%
Maryland	6	1.7%
Massachusetts	6	1.7%
Michigan	18	5.0%
Minnesota	9	2.5%
Mississippi	2	0.6%
Missouri	9	2.5%
Montana	2	0.6%
Nebraska	2	0.6%
Nevada	1	0.3%
New Hampshire	1	0.3%
New Jersey	8	2.2%
New Mexico	3	0.8%
New York	13	3.6%
North Carolina	13	3.6%
North Dakota	1	0.3%
Ohio	14	3.9%
Oklahoma	7	1.9%
Oregon	3	0.8%
Pennsylvania	13	3.6%
Rhode Island	1	0.3%
South Carolina	3	0.8%
Tennessee	3	0.8%
Texas	26	7.2%
Utah	4	1.1%
Vermont	1	0.3%
Virginia	6	1.7%

Table 2.  
*States in Which Respondents Reported Working*

States	Frequency	Percent
Washington	10	2.8%
West Virginia	2	0.6%
Wisconsin	5	1.4%
Wyoming	3	0.8%
Multiple States	11	3.0%

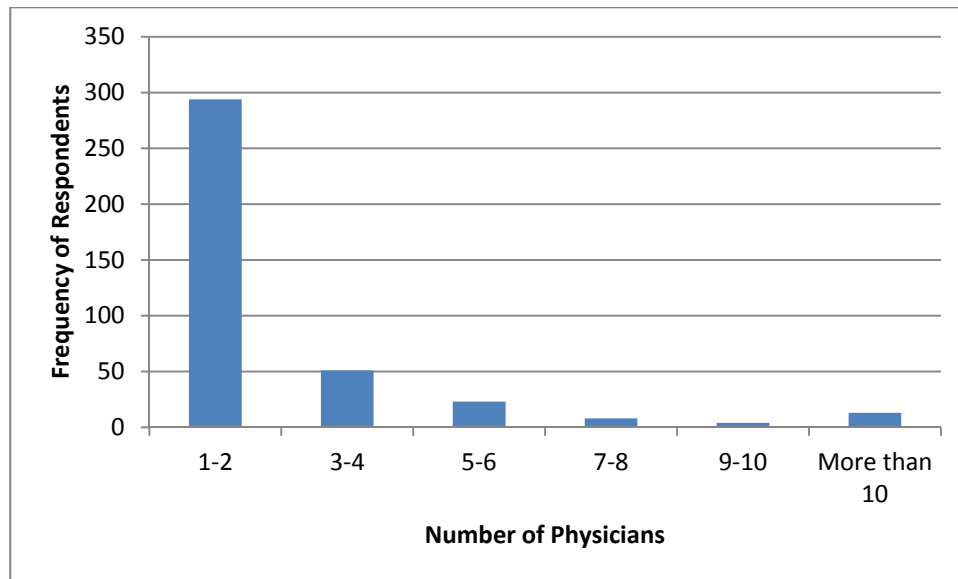
Table 3.  
*Institutions in Which Respondents Reported Working*

Institution	Frequency	Percent
Individual Private Practice	194	50.1%
Group Private Practice	118	30.5%
Community Hospital	41	10.6%
Independent Test Facility (IDTF)	6	1.6%
Mobile (Traveling) Ultrasound	3	0.8%
Government (Military, VA) Hospital	2	0.5%
Tertiary Care Center	2	0.5%

Survey participants were then asked to report the number of physicians and vascular technologists that were working in the Phlebology portion of the practice in which they worked. The majority of respondents reported having between one and two physicians (294 or 74.8%) and between one and two vascular technologists (290 or 78.6%) working in the Phlebology portion of the practice, as represented in Figures 12 and 13. Eight survey participants did not report the number of physicians working in the practice, while 32 did not report the number of vascular technologists working in the practice.

Survey participants were asked to report their highest level of education. The majority of respondents chose Doctor of Osteopathic Medicine degree or Medical Doctor as their highest level of education (274 or 69.9%), followed by a graduate degree (49 or 12.5%). The fewest number of respondents reported “high school or equivalent” as their

highest level of education, as illustrated in Table 4. Nine survey respondents did not respond to the item.



*Figure 12.* The number of physicians reported as working in the Phlebology portion of the practice in one of the studies included in this analysis.

Next, survey respondents were asked to report their backgrounds, as most do not have a background in Phlebology. A large number of survey respondents reported that they were not physicians (94 or 28.8%), as illustrated in Table 5. Of those physicians who reported their background, there was a nice spread between the background choices, with the largest number of respondents reporting general surgery as their background. Sixty-nine respondents reported “other” backgrounds.

The last background question was on the number of years spent in Phlebology. The largest number of respondents reported working in Phlebology for three to five years (121 or 30.6%). The fewest number of respondents reported working in Phlebology for more than 20 years (49 or 12.4%), as illustrated in Figure 14. Six survey participants did not respond to this item.

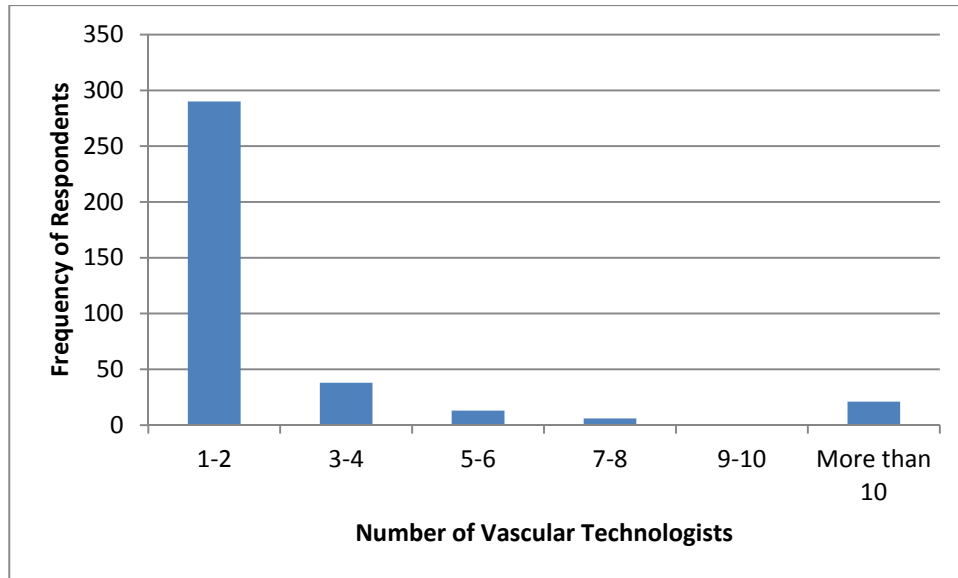


Figure 13. The number of vascular technologists reported as working in the Phlebology portion of their practice in one of the studies included in this analysis.

Table 4.  
*Survey Respondents Highest Reported Education*

Education	Frequency	Percent
High School or Equivalent	2	0.5%
Some College	4	1.0%
Two Years of College/Technical School/Community College	37	9.4%
Bachelor's Degree	26	6.6%
Graduate Degree	49	12.5%
Doctor of Osteopathic Medicine/Medical Doctor	274	69.9%

Table 5.  
*Reported Backgrounds of Physicians*

Background	Frequency	Percent
Vascular Surgery	79	24.2%
General Surgery	60	18.4%
Family Practice	39	11.9%
Interventional Radiology	30	9.2%
Internal Medicine	15	4.6%
Obstetrics and Gynecology	7	2.1%
Dermatology	3	0.9%
I am not a physician	94	28.8%

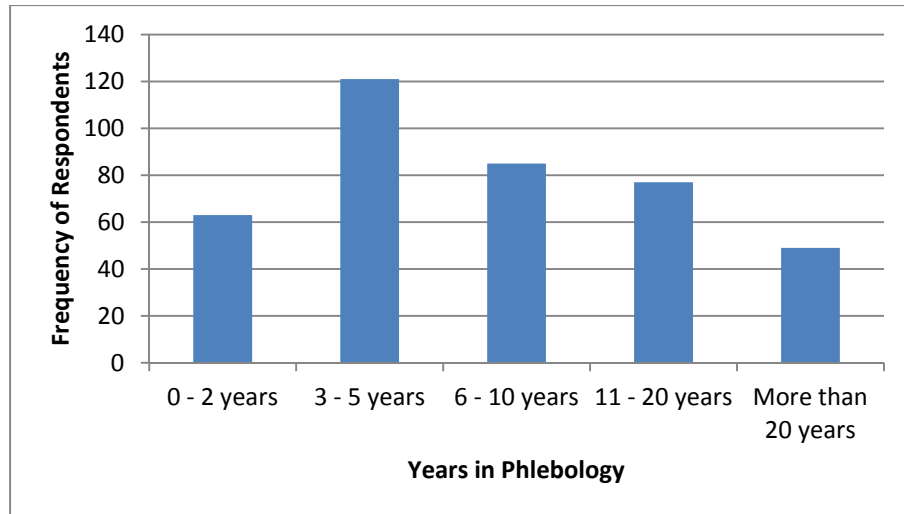


Figure 14. The number of years respondents reported working in Phlebotomy in one of the studies included in this analysis.

Again, it was determined by the organization that performed this job analysis and survey validation study that the sample of survey respondents was representative of the population of individuals working in Phlebotomy.

### Coding of Sample Studies

The 20 sample studies were coded based on five factors: task rating scale used on the survey, industry for which the job analysis was performed, number of survey respondents (sample size), presentation order, and number of tasks rated.

In each of the 20 studies, either two (14 studies) or three (six studies) rating scales were used to evaluate task statements. All 20 studies included a *task frequency* rating scale and 15 studies included a *task importance* rating scale. A *task criticality* or *consequence of error* rating scale was used in nine of the studies, and a *need at entry* rating scale was used in only three studies. The number and types of scales used in each study is presented in Table 6.

In order to separate the 20 studies into the different industries for which they were performed, a list of industries needed to be identified. A list of 21 industries was obtained from the *Occupational Information Network* (O\*NET), an online resource of jobs created by the United States government (Brannick et al., 2007). Of the 21 industries described in O\*NET, seven of the industries were represented in the 20 sample studies used in this analysis. The seven industries in which sample job analysis studies are separated include accommodation and food services; construction; educational services; healthcare and social assistance; information; professional, scientific, and technical services; and utilities; as illustrated in Table 7. The industries that are represented by the 20 sample studies seem to represent the areas in which large numbers of licensure and certification tests are utilized.

The number of survey respondents in the 20 job analyses included in this secondary data analysis ranged from a low of 37 to a high of 3,185, and were coded into one of four categories: less than 100 respondents, between 100-500 respondents, between 501-1,000 respondents, and more than 1,000 respondents. Three surveys had fewer than 100 respondents, eight surveys had between 100-500 respondents, three surveys had between 501-1,000 respondents, and six surveys had more than 1,000 respondents.

The 20 job analyses included in this study were coded based on whether survey respondents were asked to evaluate one task at a time based on all scales at once, or if they were asked to evaluate all of the tasks based on one scale and then all of the tasks again based on the next scale (referred to as presentation order). The majority of studies included in this analysis (14 or 70%) were structured the former way, whereby survey

respondents rated one task a time and considered all rating scales for each task at once, as illustrated in Figure 15.

Table 6.  
*A Breakdown of The Number and Types of Tasks Used in Each Survey Validation Study*

Study	Task Frequency	Task Importance	Task Criticality/ Consequence of Error	Need at Entry
1	X	X		X
2	X	X		
3	X	X		
4	X	X	X	
5	X	X		
6	X	X	X	
7	X	X		X
8	X	X	X	
9	X	X	X	
10	X		X	
11	X	X		
12	X	X		
13	X	X		
14	X	X		
15	X	X		
16	X	X		X
17	X		X	
18	X		X	
19	X		X	
20	X		X	

Each of the 20 job analyses were coded based on the number of tasks rated in the survey validation survey. The number of tasks rated in each of the job analyses ranged from a low of 18 to a high of 330. The 20 survey validation studies were coded into three categories: 0-50 tasks, 51-100 tasks, and 101 or more tasks. The greatest number of survey validation studies (10 or 50%) fell in the 0-50 category. Four studies fell into the 51-100 category, and the remaining six studies fell into the 101 or more category.



Table 7.

*A Breakdown of the 20 Sample Studies Based on Industry, Sample Size, Presentation Order, and Number of Tasks Rated*

	Number of Survey Respondents				Presentation Order		Number of Tasks		
	less than 100	101- 500	501- 1000	1001 or more	By Scale	By Task	0-50	51-100	101 or more
Accommodation and Food Services			9		9		9		
Construction	11, 12	8			8	11, 12	12	8, 11	
Educational Services				1	1		1		
Health Care and Social Assistance	5	2, 4, 6	3	16	4, 6	2, 3, 5, 16	3, 4, 5, 6	2	16
Information		15			15		15		
Professional, Scientific, and Technical Services				7, 10		7, 10		10	7
Utilities		13, 14, 20	18	17, 19		13, 14, 17, 18, 19, 20	13, 14		17, 18, 19, 20
<b>Number of Tasks</b>									
0-50	5, 12	4, 6, 13, 14, 15	3, 9	1	1, 4, 6, 9, 15	3, 5, 12, 13, 14			
51-100	11	2, 8		10	8	2, 10, 11			
101 or more		20	18	7, 16, 17, 19		7, 16, 17, 18, 19, 20			
<b>Presentation Order</b>									
By Scale		4, 6, 8, 15	9	1					
By Task	5, 11, 12	2, 13, 14, 20	3, 18	7, 10, 16, 17, 19					

	Scale 1	Scale 2	Scale 3
Task 1	Rating 1	Rating 2	Rating 3
Task 2	Rating 4	Rating 5	Rating 6
Task 3	Rating 7	Rating 8	Rating 9

*Figure 15.* Rating one task at a time, based on multiple scales.

To ensure that the coding made sense, a second researcher was given 10 out of the 20 studies and was asked to code all 10 studies based on the four coding criteria listed above – task rating scales used on the survey, industry for which the job analysis was performed, number of survey respondents (sample size), presentation order, and number of tasks rated. The coding of the 10 studies performed by the second researcher matched the original coding performed by the primary researcher.

Prior to beginning the data analysis, all potential moderators, along with a few additional factors (percentage of eliminated responses, whether the study was conducted for a new credential or revalidation of an existing credential, and whether the study was conducted for a licensure exam or certification exam) were correlated. There were four statistically significant correlations: 1) the number of tasks with the percent of survey respondents eliminated  $r = .791$ ; 2) the number of tasks with whether the study was performed for a new credential or revalidating an existing credential  $r = .587$ ; 3) whether the study was performed for a new credential or revalidation an existing credential with presentation order  $r = .504$ ; and 4) the number of scales used on the survey with presentation order  $r = -.663$ . A more detailed description of the results are presented in Appendix B.

## Data Analysis

All data analyses were performed through one of two software packages - SAS 9.3 or Microsoft Excel 2010. The first step in the data analysis process was to look at some of the statistical properties of the 20 individual job analysis and survey validation studies. The reliability of each scale on each survey was computed using Cronbach alpha. As each study had either two or three individual scales and between one and three composite scales, there were between three and six Cronbach alphas computed for each study. The Cronbach alphas across all scales across all 20 studies were between .85 and .99. This finding suggests that the scales used in each of the studies were reliable. This finding also suggests that all of the tasks “hang together”. As such, all scales from all 20 studies were included in this analysis.

### *Missigness Analysis*

Next, the amount of missingness was analyzed across survey respondents. Any respondent who completed less than 75% of the survey was removed from the final analysis. For example, in survey validation for journeyman plumbers, there were 120 ratings (60 tasks x 2 scales = 120 total ratings), so anyone who provided fewer than 90 ratings were eliminated from the final analysis. For this study, that meant eliminating 20 out of 65 survey respondents (anyone below the red line) from the final analysis, as illustrated in Figure 16. Note, 19 survey respondents did not provide any task ratings. The reason these individuals were included in the dataset is because they most likely provided demographic information and as such were originally included in the data analysis.

For the survey validation for individuals working in Phlebology, there were 102 ratings (34 tasks x 3 scales = 102 total ratings), so anyone who provided fewer than 77 ratings were eliminated from the final analysis. For this study, that meant eliminating 149 survey respondents out of 400 survey respondents (anyone below the red line) from the final analysis, as illustrated in Figure 17. Note, 68 survey respondents in the dataset did not provide any task ratings.

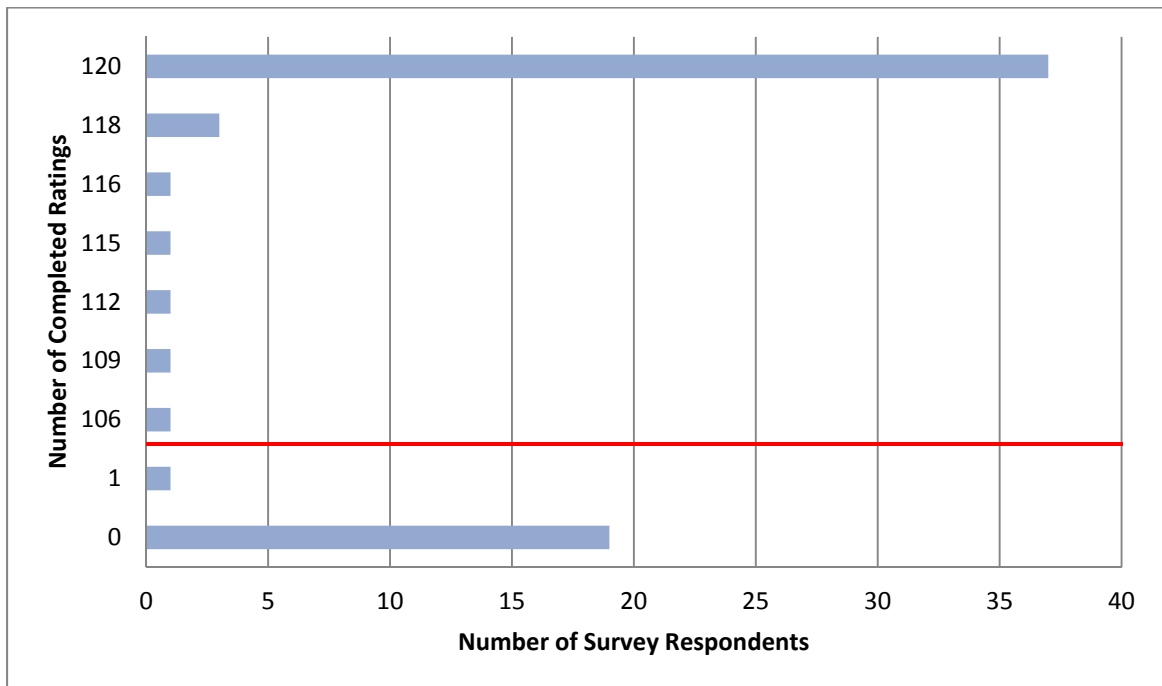
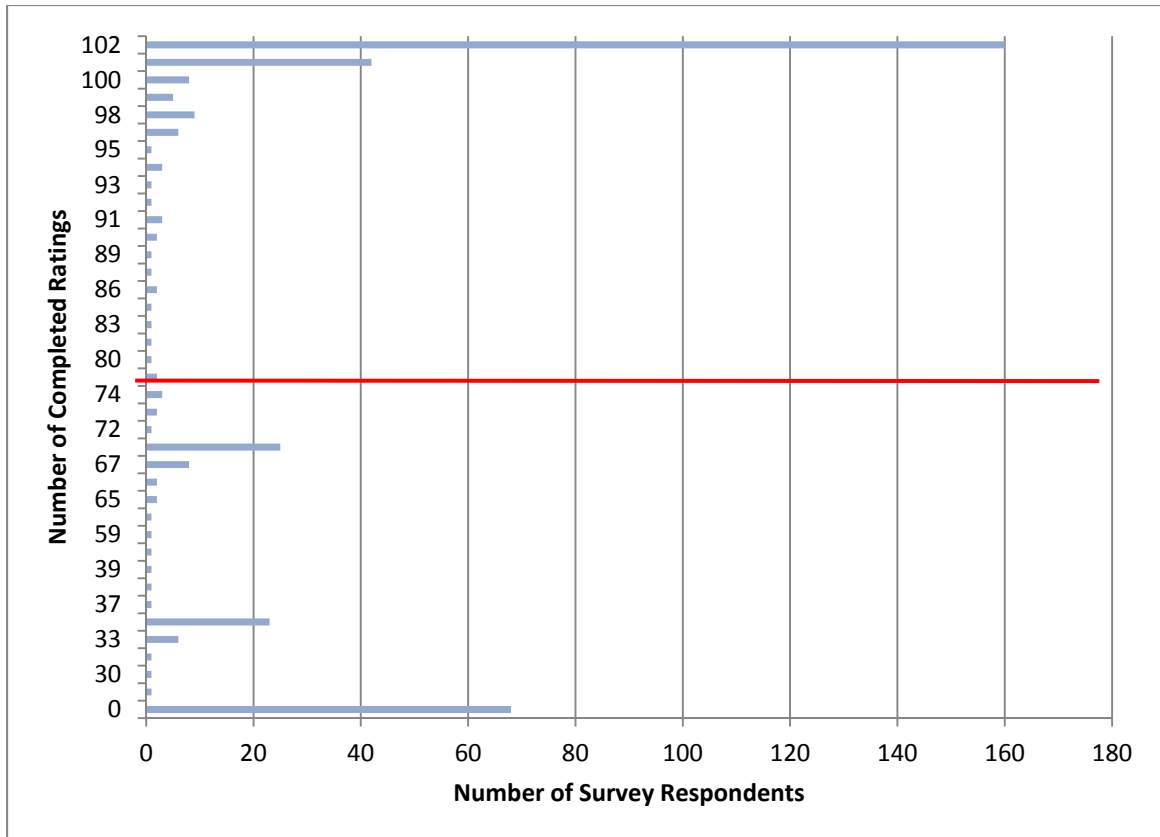


Figure 16. Number of survey respondents who completed the journeyman plumber validation survey.

The amount of missingness and task ratings were evaluated by computing a correlation between the two variables. The average correlation between missingness and task ratings was .02 and the median correlation was .01, as illustrated in Figure 18. This finding suggests that there is essentially no relationship between how people respond to task ratings and when they stop responding.

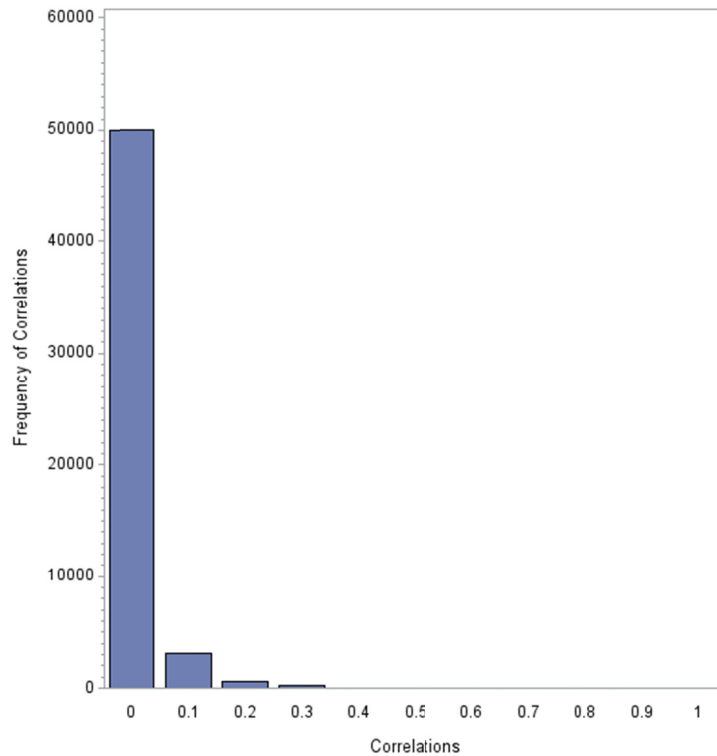


*Figure 17.* Number of survey respondents who completed the Phlebology validation survey.

Additionally, all missing tasks were correlated with one another. The average correlation between missingness was .66 and the median correlation was .75, as illustrated in Figure 19. This finding suggests that survey respondents who didn't complete the survey stopped responding to task ratings early in the survey and then left the remaining part of the survey blank (as opposed to jumping around and intentionally leaving some task ratings blank and responding to others).

Task ratings for each rating scale were aggregated across all survey respondents (minus the survey respondents who did not complete at least 75% of the survey). "This aggregation process is typically done in job analysis, because it is assumed that individual

biases are overcome by aggregating the data across subjects” (Sanchez & Levine, 1989, p. 38).



*Figure 18.* Distribution correlations between task ratings and missingness. N=478,079.

### *Outlier Analysis*

Using SAS 9.3, Pearson product moment correlation coefficients (PPMCCs) were computed for every pair of individual and composite task rating scales to answer the first three overarching research questions. The correlations were reviewed to determine if there were outliers. Four outliers were discovered. The four outlier correlations are presented in Table 8.

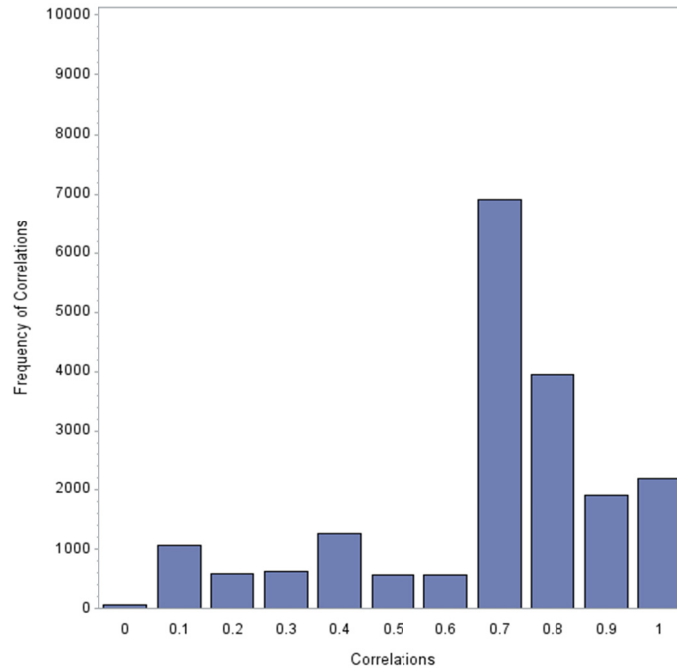


Figure 19. Distribution correlations between the amount of missingness across all task rating scales. N=222,064.

Table 8.  
*Description of Outlier Correlations*

Study	Variable	With Variable	N Tasks	Corr	Industry	Sample Size	Presentation Order
16	Imp	Need	123	.27	Health	1,798	Task
16	Need	Comp1	123	.31	Health	1,798	Task
16	Freq	Need	123	.33	Health	1,798	Task
10	Freq	Crit	87	.37	Prof, Sci	3,043	Task

Three of the outlier correlations included the Need at Entry rating scale, however, there were 12 total correlations that included the Need at Entry rating scale ranging from .52 to .96. Three of the outlier correlations came from the same study, study 16, however, study 16 contributed a total of 10 correlations, the other seven of which fell between .91 and .98. All of the outlier correlations came from studies with higher sample sizes, but again, there were 39 other studies with large sample sizes that produced much higher correlations.

The one unique finding related to the outlier correlations is that every correlation that included the Need at Entry Scale *and* came from a study in which the presentation order was task-based was an outlier. This finding may suggest that when a survey respondent is presented with the Need at Entry rating scale directly next to one or more rating scales, the Need at Entry rating scale is rated very differently.

Due to the fact that there were only four outlier correlations, all of the subsequent analyses were conducted both with and without the four outliers to see if the results of the study would be different if the outliers were removed. Ultimately, including the outliers did not change the outcome of any of the findings of this study; a decision was made to include all outliers in the final analysis.

To answer research question 1A, the average task rating was computed for each task and each rating scale, by each study. Then, a mean correlation was computed for each pair of individual rating scales in each study, as illustrated in Figure 20.

Studies	Tasks	Scale 1	Scale 2
Correlation for Study 1	Task 1	Mean Rating(Person1-Personj)	Mean Rating(Person1-Personj)
	Task 2	Mean Rating(Person1-Personj)	Mean Rating(Person1-Personj)
	Task j	Mean Rating(Person1-Personj)	Mean Rating(Person1-Personj)
Correlation for Study 2	Task 1	Mean Rating(Person1-Personj)	Mean Rating(Person1-Personj)
	Task 2	Mean Rating(Person1-Personj)	Mean Rating(Person1-Personj)
	Task j	Mean Rating(Person1-Personj)	Mean Rating(Person1-Personj)
Correlation for Study j	Task 1	Mean Rating(Person1-Personj)	Mean Rating(Person1-Personj)
	Task 2	Mean Rating(Person1-Personj)	Mean Rating(Person1-Personj)
	Task j	Mean Rating(Person1-Personj)	Mean Rating(Person1-Personj)

*Figure 20.* Illustration of data analysis method to answer research question 1A.

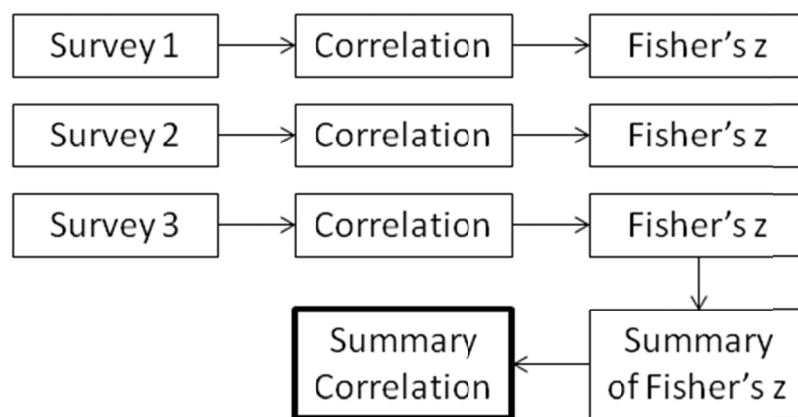
The same process was used to answer research questions 1B and 1C, but rather than looking at the correlation between two individual rating scales, the correlations were computed between all sets of composites and between individual rating scales with



composites scales. As Sanchez and Fraser indicated "... the more highly correlated the composite scales, the higher the reliability of the composite" (1992, p. 552).

Upon obtaining correlations for each set of rating scales, meta-analytic techniques were used to compare the mean correlation between the rating scales. According to Lipsey and Wilson, "the correlation coefficient is already a standardized index and therefore is usable as a meta-analytic effect size statistic in its raw form even if the variables being correlated are differently operationalized" (2001, p. 63). To use a correlation as an effect size, the correlations were transformed using Fisher's  $Z_r$ -transformation, as illustrated in Figure 21.

The correlations obtained from each pairing of task rating scales were transformed using Fisher's  $Z$ -transformation. All of the transformed effect sizes were compiled to create a summary effect size (mean weighted transformed correlation), and a confidence interval was built around that summary effect size. If the confidence interval contained zero, the mean correlation was not statistically significantly different from zero. If the confidence interval did not contain zero, the mean differs significantly from zero – meaning there was a statistically significant relationship between two different rating scales. Additionally, prediction intervals were calculated around each of the weighted mean correlations. The prediction intervals illustrate how true effects are disbursed around summary effects. The wider the prediction interval, the greater the distribution is of true effects.



*Figure 21.* Illustration of how the summary correlation are be derived for each of the pairs of rating scales and/or composites using Fisher’s  $z_r$ -transformation. Adapted from “Correlations are analyzed in Fisher’s  $z$  units” by M. Borenstein, L.V. Hedges, J.P.T. Higgins and H.R. Rothstein, 2009, *Introduction to Meta-Analysis*, p. 42. Copyright 2009 by John Wiley & Sons, Ltd.

For the first three research questions, a random effects model was chosen. The random effects model was selected because according to Lipsey and Wilson (2001) random effects models should be used over fixed effects models when there is a smaller number of effect sizes used in the comparison and there may be variability in the effect sizes that is due to something other than the rating scale chosen. To perform this analysis Proc Mixed was used to account for the violation of independence because multiple effect sizes coming from a single study.

For research questions 2 and 3, an analysis of variance (ANOVA) analog was used to group and compare mean effect sizes based on the four moderator variables introduced in the research questions: industry (research question 2), sample size (research question 3A), presentation order (research question 3B), and number of tasks (research question 3C). The ANOVA analog was chosen because the four variables listed above will all be treated as categorical, and because “the ANOVA analog is best suited to

testing a limited set of a priori hypotheses regarding moderator variables” (Lipsey & Wilson, 2001, p. 120).

For research question 4, draft examination blueprints were created for each of the 20 job analyses based on composites of the rating scales included in the job analysis. Examination blueprints were created using the method previously described, as outlined by Raymond (1996) and Kane, Kingsbury, Colton, and Estes (1989). It is important to emphasize that examination blueprints were created using both individual scales and composites of scales. For example, if a survey validation study included three rating scales, an examination blueprint was created from all three individual scales, as well as one or more composites of those scales to see if differences exist.

To identify the differences between examination blueprints, the relative weightings for each task and overarching duty (or content area) were compared to see if differences existed based on the scale, or combination of scales, chosen to produce the examination blueprint weightings. For example, the highest weighted duty area derived by one scale was compared to the highest weighted duty area derived by a second scale, and the composite of scales, to determine if they are the same or different. Additionally, the absolute differences between the percent of the examination blueprint dedicated to each overarching duty or content area was compared across all examination blueprints.

Both the relative differences between duty areas and absolute differences between duty areas were evaluated for all blueprints derived from individual and composite ratings. The relative and absolute examination blueprint weights were compared to the relative and absolute examination blueprint weights on the actual examination blueprints.

Lastly, examination blueprint weights were created with all duty areas equally weighted to see how the equally weighted blueprints would compare to examination blueprints derived from the individual and composite scales, as well as actual examination blueprints.

## CHAPTER FOUR: RESULTS

This chapter details the results of the study in relation to each of the four research questions. The chapter is divided into three sections and organized in the order of the research questions. First, the results from the first research question will be presented. This includes the relationship between a) individual rating scales and other individual rating scales, b) composite rating scales and composite rating scales, and c) individual rating scales and composite rating scales. The relationships are described in terms of the weighted and unweighted average correlations between each set of individual and composite rating scales.

Second, the results from research questions two and three will be presented. Again, the correlations between each set of individual and composite rating scales will be presented; however, the correlations will be grouped by the level of each of the four potentially moderating variables: industry, number of survey respondents, presentation order, and number of tasks rated. Additionally, the results of the ANOVA analog will be presented.

Third, results from the fourth research question will be presented. This includes the correlations between the relative rankings of content areas from examination blueprints derived from individual or composites of rating scales when exam blueprints were derived from individual or composite rating scales.

### **Research Questions**

The four overarching research questions analyzed in this study are:

1. What is the relationship between the different types of individual and composite rating scales?
  - a. What is the relationship between different types of individual rating scales?
  - b. What is the relationship between different types of composite rating scales?
  - c. What is the relationship between different types of individual rating scales with different types of composite scales?
2. To what extent do the relationships of individual and composite rating scales vary across industries?
3. To what extent do the relationships of individual and composite rating scales vary across survey design factors?
  - a. To what extent do the relationships of individual and composite rating scales vary across varying numbers of survey respondents?
  - b. To what extent do the relationships of individual and composite rating scales vary across scale presentation order?
  - c. To what extent do the relationships of individual and composite rating scales vary across the number of tasks rated?
4. To what extent are examination blueprint weightings different based on the choice of scale composites used in the survey validation study?

### **Research Question One Results**

In total, there were 129 correlations computed across the 20 studies included in the analysis. Each study provided three, 10, or 15 correlations to the analysis. A

breakdown of the 20 studies, the number and type of scales used in each study, the number of tasks included in the correlations, and the number of correlations that each study contributed to the analysis is presented in Table 9.

Table 9.  
*Distribution of Correlations Between 20 Studies*

Studies	N Scales	Freq	Imp	Crit	Need	Comp1	Comp2	Comp3	Comp4	N Tasks	N Corrs
1	3	X	X		X	X			X	47	10
2	2	X	X			X				51	3
3	2	X	X			X				33	3
4	3	X	X	X		X	X	X		34	15
5	2	X	X			X				37	3
6	3	X	X	X		X	X	X		30	15
7	3	X	X		X	X			X	190	10
8	3	X	X	X		X	X	X		59	15
9	3	X	X	X		X	X	X		32	15
10	2	X		X			X			87	3
11	2	X	X			X				60	3
12	2	X	X			X				50	3
13	2	X	X			X				36	3
14	2	X	X			X				18	3
15	2	X	X			X				19	3
16	3	X	X		X	X			X	123	10
17	2	X		X			X			305	3
18	2	X		X			X			222	3
19	2	X		X			X			331	3
20	2	X		X			X			180	3
Totals		20	15	9	3	15	9	4	3		129

The range of the obtained correlations was .27 to 1.00. The unweighted mean correlation was .87, and the weighted mean correlation for all 129 correlations was .92. A histogram of all obtained correlations is presented in Figure 22.

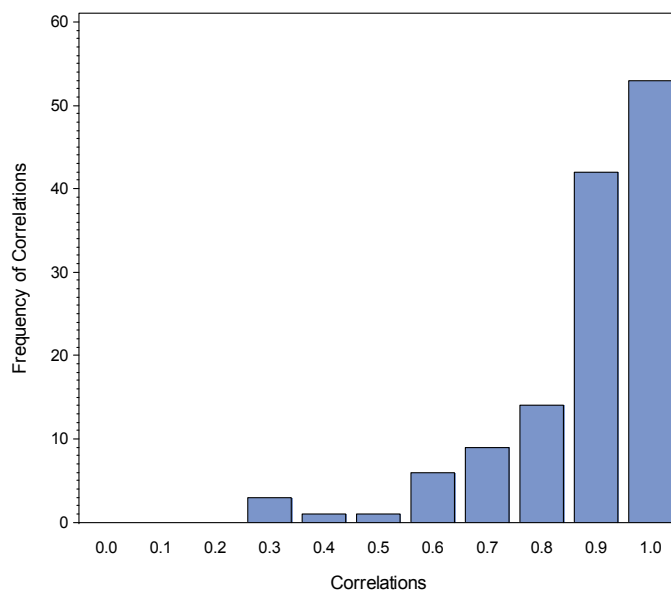
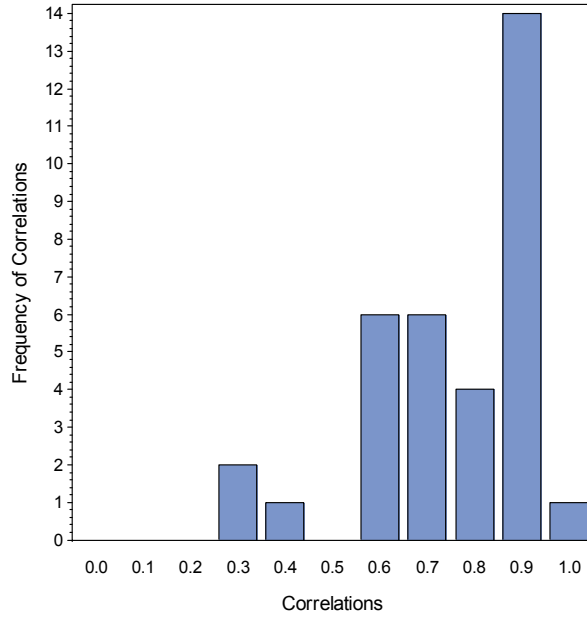


Figure 22. Distribution of all obtained correlations. N=129.

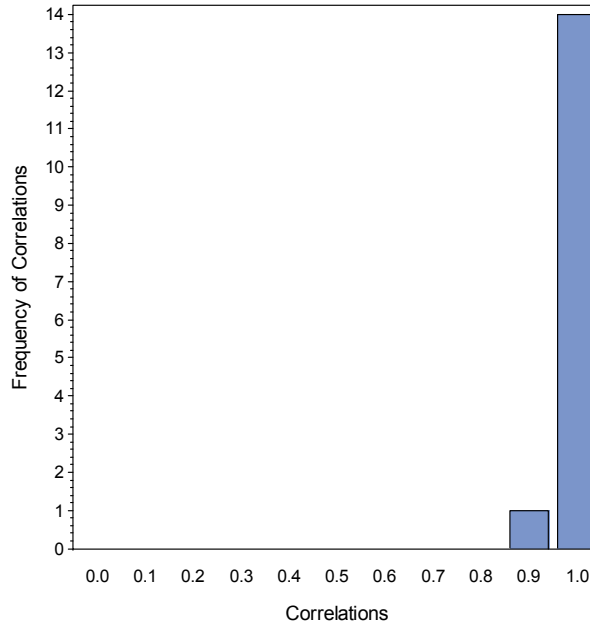
Of the 129 obtained correlations, 34 were between two individual rating scales, 15 were between two composite rating scales, and 80 were between an individual scale and composite scale, as illustrated in Figures 23, 24, and 25 respectively. The unweighted mean correlation between two individual scales was .75 and the mean weighted correlation between pairings of individual scales was .79. The mean unweighted correlation between two composite scales was .98, and the mean weighted correlation was .99. Finally, the mean between an individual and composite scale was .91, and the mean weighted correlation between individual and composite scales was .94.

The distribution of correlations between two individual scales had the most variability, ranging from .27 to .95. The correlation between two composites had the least variability, ranging from .95 to 1.00. The range of correlations between all individual and composite scales was .30 to .99.





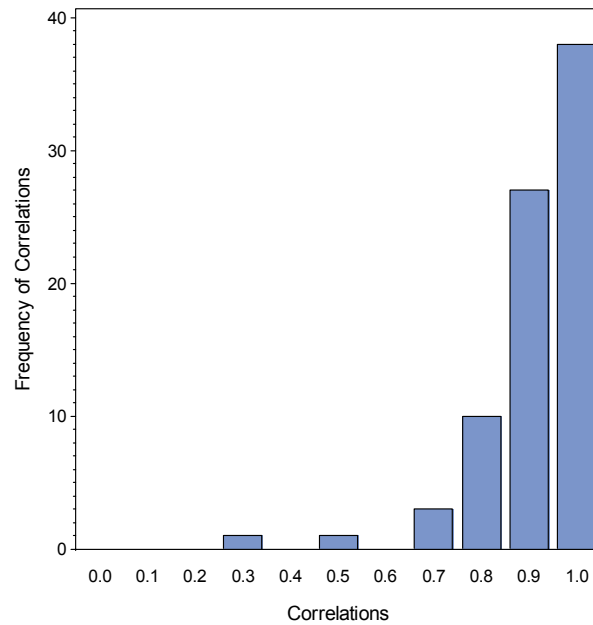
*Figure 23.* Distribution correlations between two individual scales. N=34.



*Figure 24.* Distribution correlations between two composite scales. N=15.

Among the 129 correlations, there were 22 pairings of scales, as illustrated in Figure 26. Of the 22 correlations, five were between two individual scales, four were

between two composite scales, and 13 were between and individual and composite scale, as illustrated in Figure 27.



*Figure 25.* Distribution correlations between a composite and individual scale. N=80.

The five pairings between two individual scales were

- Frequency with Criticality,
- Frequency with Importance,
- Frequency with Need at Entry,
- Importance with Criticality, and
- Importance with Need at Entry.

There were no correlations between the Criticality and Need at Entry scales as the Criticality and Need at Entry scales were not used on the same survey in any of the 20 sample studies included in this study. The distribution of correlations between the five individual scales is presented in Figure 28. Of the five combinations of individual scales,

the correlations between Importance and Criticality ratings were the highest with the least amount of variability.

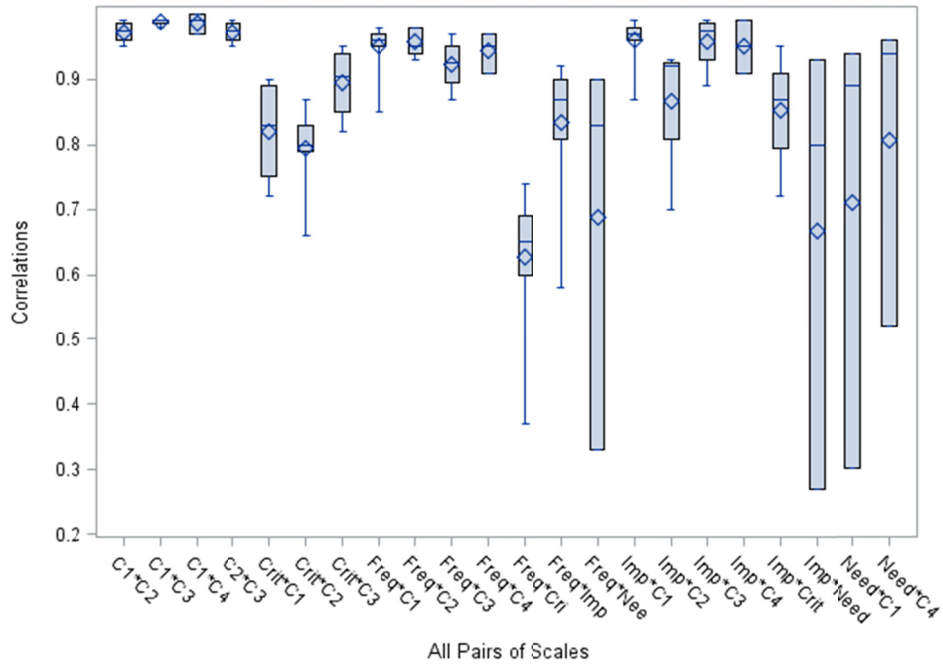


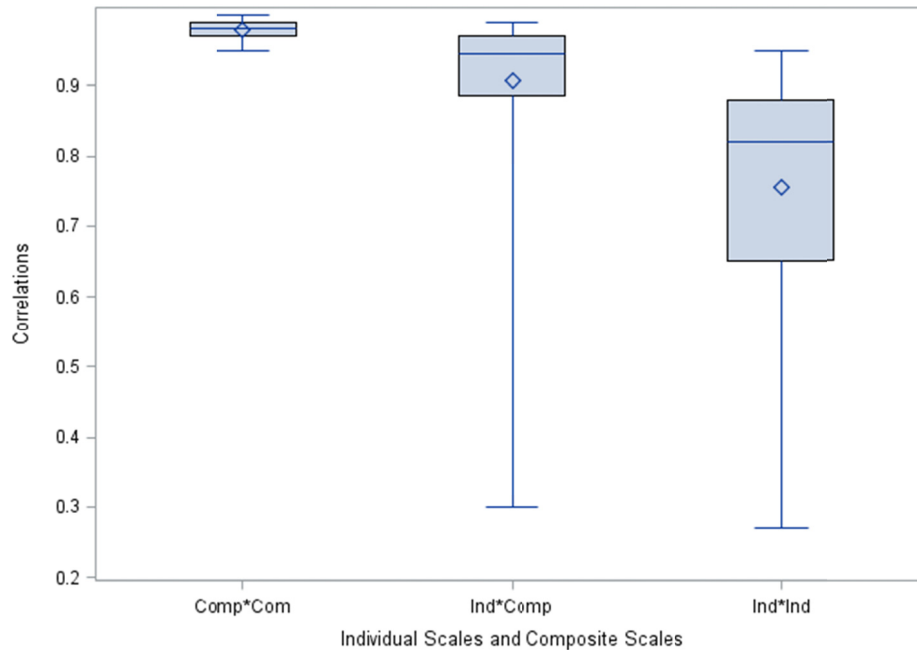
Figure 26. Distribution of all obtained correlations by all combinations of scales. N=129.

There were 15 correlations between two pairings of composite scales. The four combinations of composites rating scales include

- Composite 1 with Composite 2,
- Composite 1 with Composite 3,
- Composite 1 with Composite 4, and
- Composite 2 with Composite 3.

There were no correlations between the Composite 2 and Composite 4 and Composite 3 and Composite 4. Composite 2 includes the Criticality and Frequency scales and Composite 3 includes the Importance, Criticality, and Frequency scales, whereas Composite 4 includes the Importance, Frequency, and Need at Entry scales. As

previously mentioned, there were no studies included in this analysis that had both the Criticality and Need at Entry rating scales. Therefore there was no opportunity to have Composite 4 correlated with Composite 2 or 3.



*Figure 27.* Distribution of all obtained correlations for composites with composites, individual scales with composite scales, and individual scales with individual scales. N=129.

The range of correlations between all four pairs of composites was small because most of the correlations between composites were very high, as indicated in Figure 29.

There were 13 pairings of individual and composite rating scales, as illustrated in Figure 30. The Criticality rating scale was correlated to Composites 1, 2, and 3; the Frequency and Importance rating scales were correlated with all four Composites; and the Need at Entry rating scale was correlated with Composites 1 and 2. Note, responses to the Need at Entry scale were correlated with Composite 1 (which is defined derived from a combination of Frequency and Importance scales, and does not include ratings from the

Need at Entry rating scale) because three of the surveys included in this study included Frequency, Importance, and Need at Entry rating scales.

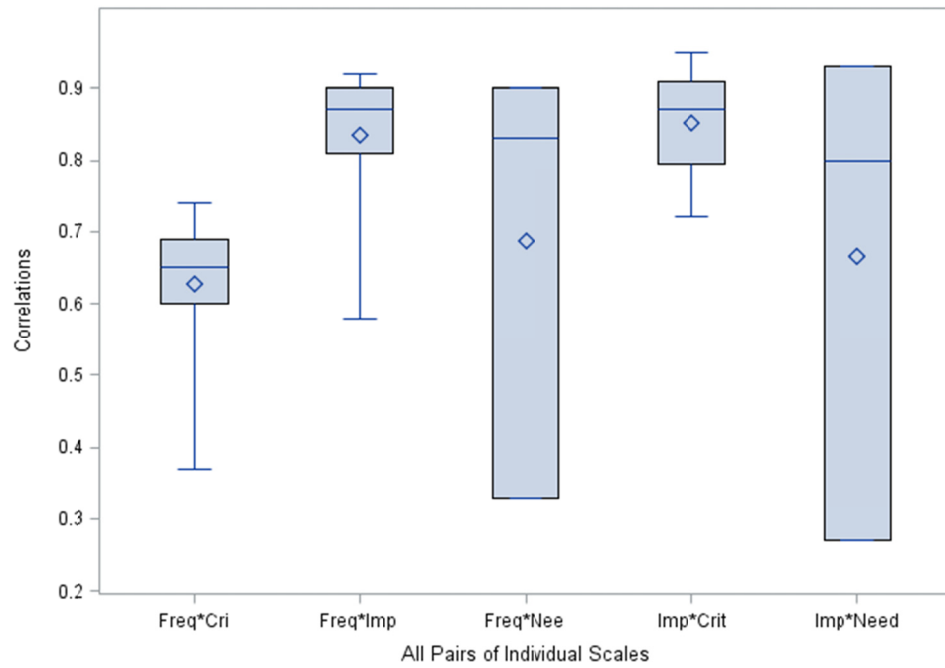


Figure 28. Distribution of correlations between individual scales. N=34.

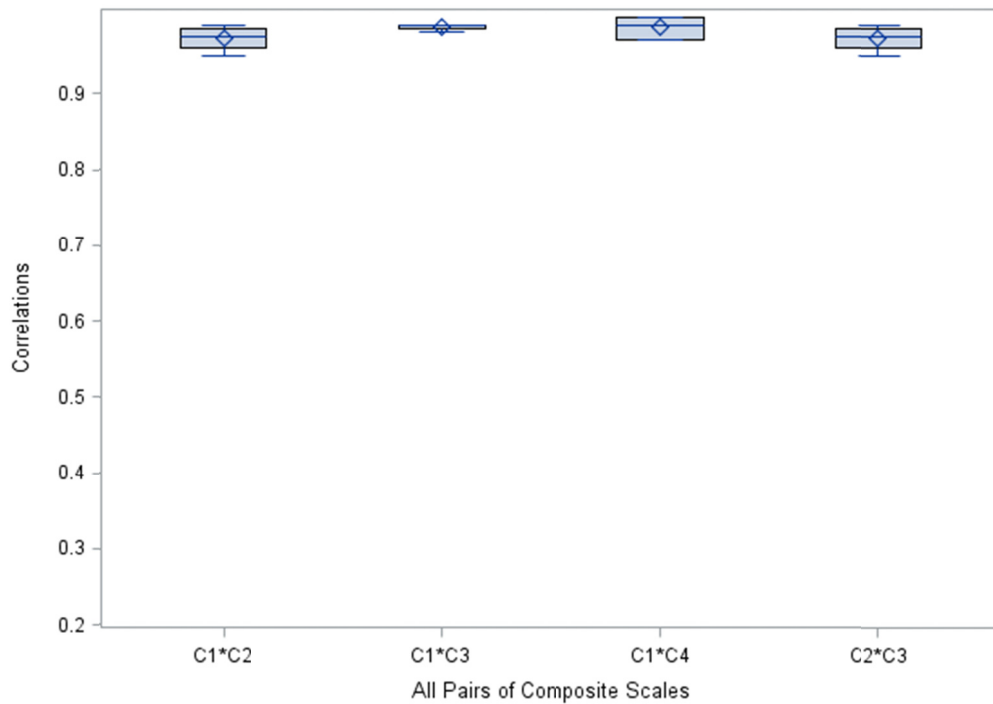


Figure 29. Distribution of all correlations between composite scales. N=15.

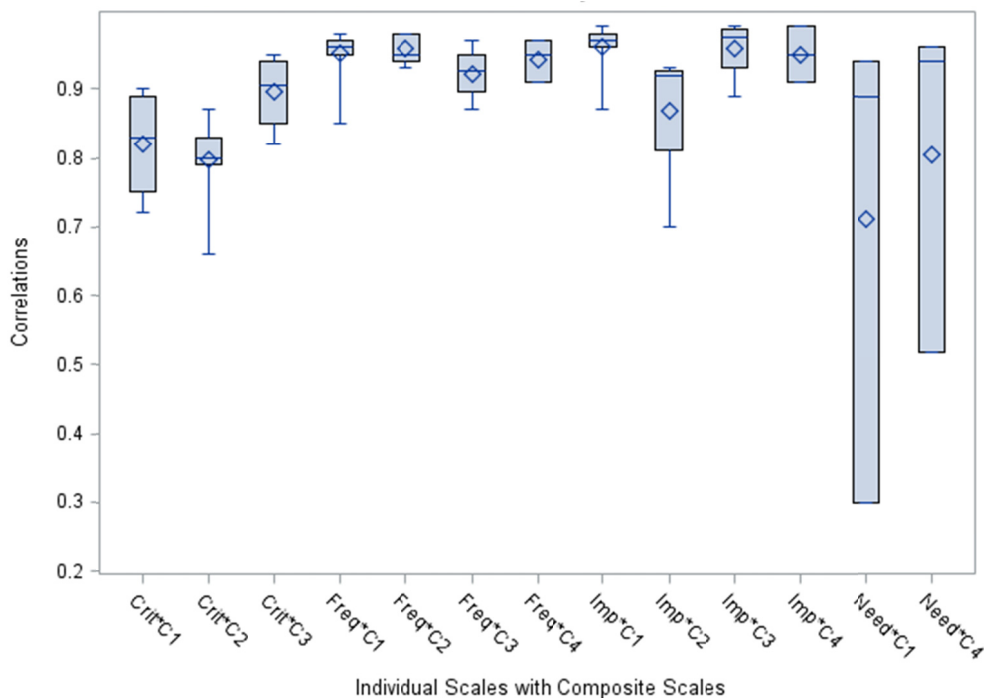


Figure 30. Distribution of all correlations between individual and composite scales. N=80.

Again, there was a lot of variability in the range of correlations between individual and composite scales. Both the correlations between Composite 1 and Composite 4 with the Need at Entry scale were on average lower than the other pairings of scales, and had the greatest amount of variability.

### Research Questions Two and Three Results

#### *Research Questions Two and Three for All Correlations*

To answer research questions two and three, five ANOVA analogs were computed using SAS 9.3. Each of the ANOVA analogs included all 129 obtained correlations as dependent variables. The five ANOVA analogs were analyzed both with all 129 obtained correlations and with the four outlier correlations removed. The results were the very similar regardless of whether or not outlier correlations were included. The *F* statistics of the ANOVA analogs on average differed by 0.1 depending on whether or

not the outlier correlations were included. The significance of the  $p$  values of all five ANOVA analogs did not change based on whether the outlier correlations were included. As such, a decision was made to include all 129 correlations in the five ANOVA analogs.

The first ANOVA analog included the “industry” as the moderating variable, the second included “sample size” as the moderating variable, the third included “presentation order” as the moderating variable, and the fourth included “number of tasks” as the moderating variable. The fifth ANOVA analog included all four moderating variables.

The relationship between all combinations of scales by industry is presented in Figure 31. The mean weighted correlations for all combinations of scales, their confidence intervals, and prediction intervals are also presented by industry in Table 10. The prediction intervals are substantially higher than the confidence intervals suggesting that the distribution of actual correlations is great. There was no statistically significant relationship between the industry in which the survey validation study was performed and the observed correlations between all rating scales,  $F(6,122) = 0.39, p = 0.8830$ , as illustrated in Table 11.

Table 10.  
*Mean Weighted Correlation, CIs and PIs for All Combinations of Scales by Industry*

Industry	Weighted Mean Correlation	95% CI	95% PI
Construction	0.92	[0.83, 0.96]	[0.66 , 0.98]
Education	0.91	[0.68, 0.98]	[0.50 , 0.99]
Food	0.94	[0.79, 0.98]	[0.65 , 0.99]
Healthcare	0.94	[0.89, 0.97]	[0.74 , 0.99]
Information	0.89	[0.63, 0.97]	[0.42 , 0.98]
Professional	0.93	[0.84, 0.97]	[0.68 , 0.98]
Utilities	0.89	[0.82, 0.94]	[0.59 , 0.97]

Note: CI =Confidence Intervals, PI=Prediction Intervals

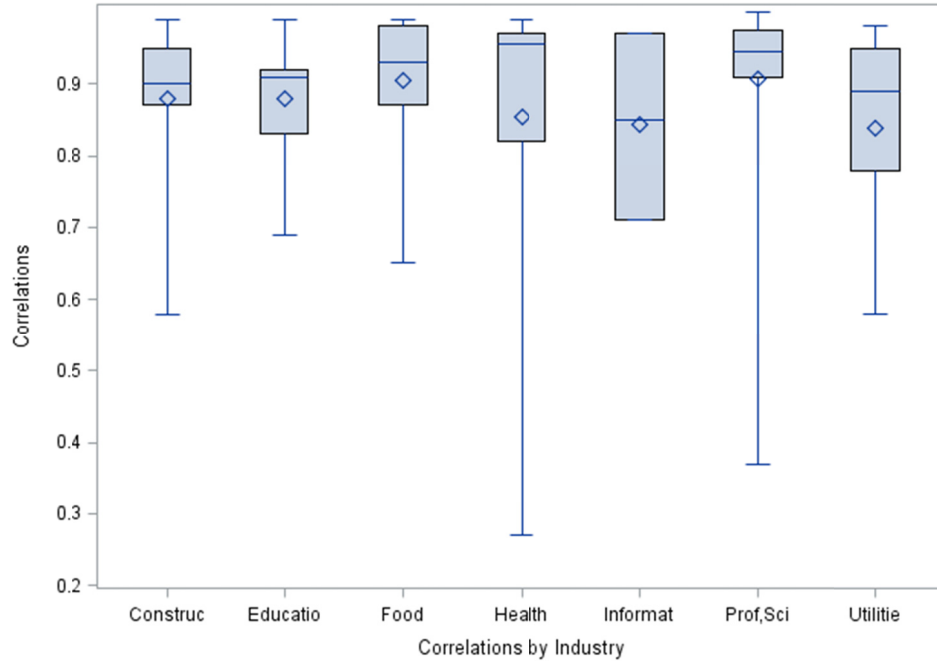


Figure 31. Distribution of all obtained correlations by industry. N=129.

The relationship between all combinations of scales by sample size is presented in Figure 32. The mean weighted correlations for all combinations of scales and their confidence intervals are also presented by sample size in Table 12. There was no statistically significant relationship between sample size and the correlations between all rating scales,  $F(3,125) = 1.14, p = 0.3349$ , as illustrated in Table 13.

Table 11.

*Fixed and Random Effects for Industries on Correlations*

Effect	Estimate (SE)
<b>Fixed Effects</b>	
Intercept	1.43 (0.14)
Construction	0.17 (0.25)
Education	0.09 (0.38)
Food	0.32 (0.38)
Healthcare	0.29 (0.21)
Information	0.00 (0.38)
Professional	0.20 (0.25)
Utilities	Reference
<b>Random Effects</b>	
Study	0.08 (0.03)



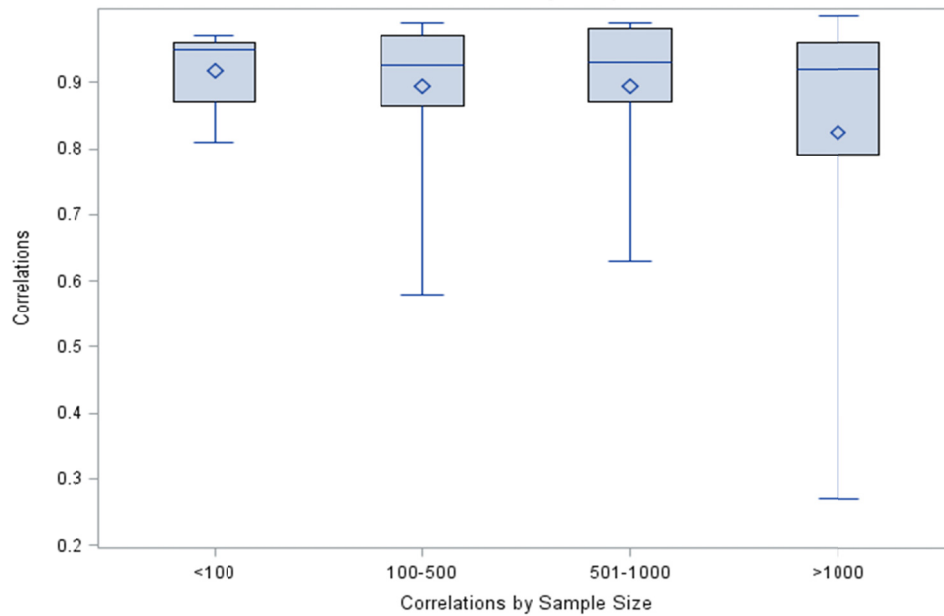


Figure 32. Distribution of all obtained correlations by sample size. N=129

Table 12.  
Mean Weighted Correlation, CIs and PIs for All Combinations of Scales by Sample Size

Sample Size	Weighted Mean Correlation	95% CI	95% PI
<100	0.93	[0.87, 0.97]	[0.75, 0.98]
100-500	0.93	[0.89, 0.95]	[0.77, 0.98]
501-1000	0.92	[0.85, 0.96]	[0.72, 0.98]
>1000	0.88	[0.81, 0.93]	[0.62, 0.97]

Note: CI=Confidence Intervals, PI=Prediction Intervals

Table 13.  
Fixed and Random Effects for Sample Size on Correlations

Effect	Estimate (SE)
Fixed Effects	
Intercept	1.38 (0.13)
<100	0.31 (0.22)
100-500	0.28 (0.17)
501-1000	0.23(0.22)
>1000	Reference
Random Effects	
Study	0.10 (0.03)

The relationship between all combinations of scales by presentation order is presented in Figure 33. The mean weighted correlations for all combinations of scales and their confidence intervals are also presented by presentation order in Table 14. There was no statistically significant relationship between presentation and the correlations between all rating scales,  $F(1,127) = 1.69, p = 0.1964$ , as illustrated in Table 15.

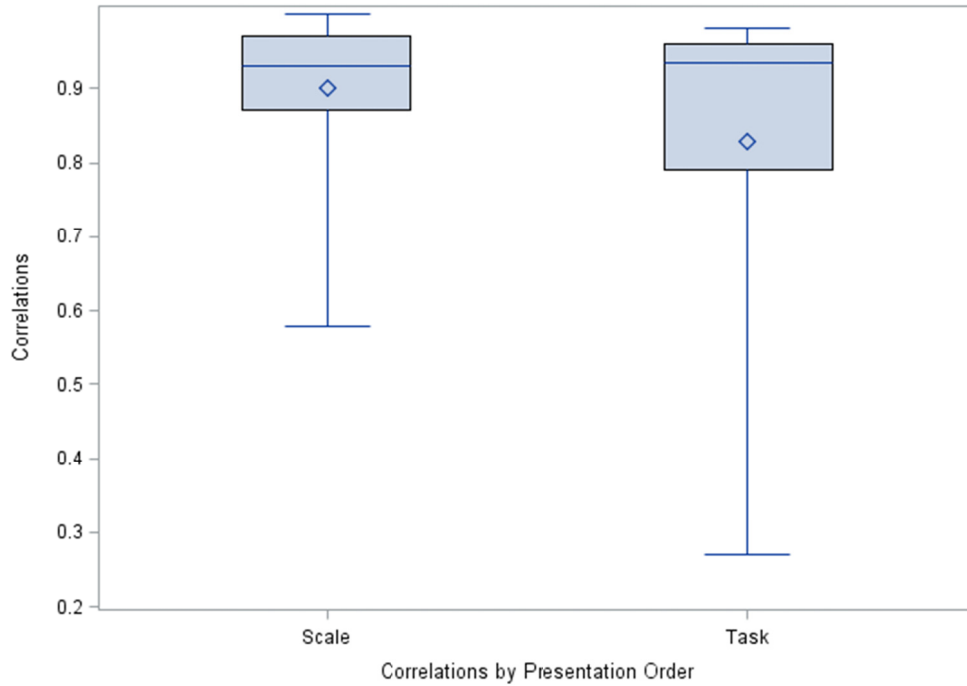


Figure 33. Distribution of all obtained correlations by presentation order. N=129.

Table 14.  
Mean Weighted Correlation, CIs and PIs for All Combinations of Scales by Presentation Order

Presentation Order	Weighted Mean Correlation	95% CI	95% PI
By Scale	0.94	[0.90, 0.96]	[0.78, 0.98]
By Task	0.91	[0.87, 0.93]	[0.7, 0.97]

Note: CI =Confidence Intervals, PI=Prediction Intervals

Table 15.  
*Fixed And Random Effects for Presentation Order on Correlations*

Effect	Estimate (SE)
Fixed Effects	
Intercept	1.51 (0.03)
Scale	0.19 (0.14)
Task	Reference
Random Effects	
Study	0.10 (0.03)

The relationship between all combinations of scales by number of tasks is presented in Figure 34. There was a statistically significant relationship between the number of tasks on the correlations obtained between all rating scales,  $F(2,126) = 3.64$ ,  $p = 0.0291$ , as illustrated in Table 16. The mean weighted correlations for all combinations of scales and their confidence intervals are also presented by number of tasks in Table 17.

A follow-up Tukey test indicated that there was a statistically significant difference in correlations between rating scales when a small number of tasks were rated (0-50 tasks) compared to studies in which a large number of tasks were rated (more than 100 tasks),  $t(126) = 2.59$ ,  $p = 0.0107$ .

Table 16.  
*Fixed and Random Effects for Number of Tasks on Correlations*

Effect	Estimate (SE)
Fixed Effects	
Intercept	1.37 (0.11)
0-50	0.37 (0.14)
51-100	0.12 (0.18)
>101	Reference
Random Effects	
Study	0.08 (0.03)

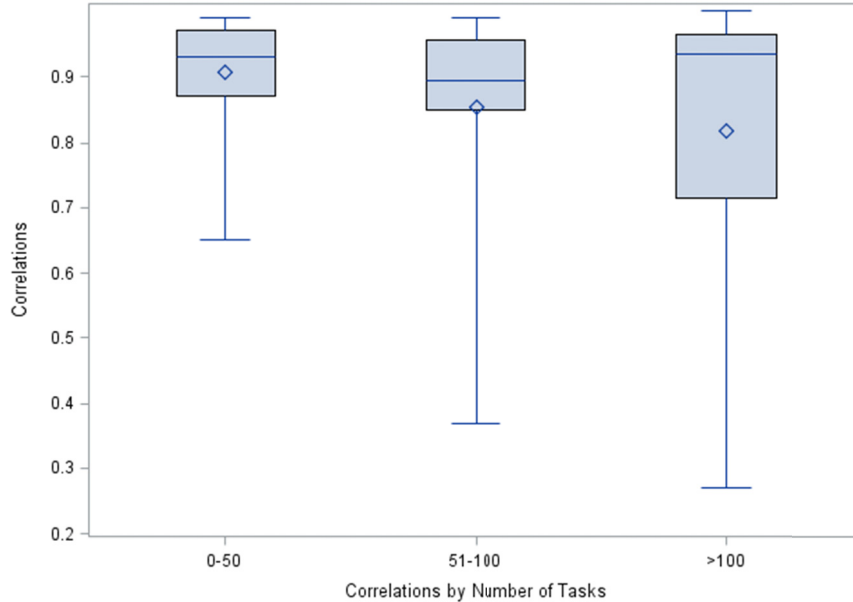


Figure 34. Distribution of all obtained correlations by number of tasks. N=129.

Table 17.

*Mean Weighted Correlation, CIs and PIs for All Combinations of Scales by Number of Tasks*

Number of Tasks	Weighted Mean Correlation	95% CI	95% PI
0-50 Tasks	0.96	[0.90, 0.99]	[0.82 , 0.98]
51-100 Tasks	1.00	[0.98, 1.00]	[0.70 , 0.97]
More than 100 Tasks	1.00	[0.99, 1.00]	[0.65 , 0.96]

Note: CI =Confidence Intervals, PI=Prediction Intervals

There was no statistically significant effect of all four variables together on the correlations between all rating scales, as illustrated in Table 18.

*Research Questions Two and Three for Correlations between Individual Scales*

The same five ANOVAs were computed with only pairings of individual scales. The first ANOVA analog included the “industry” as the moderating variable, the second included “sample size” as the moderating variable, the third included “presentation order” as the moderating variable, and the fourth included “number of tasks” as the moderating variable. The fifth ANOVA analog included all four moderating variables.

Table 18.  
*Fixed And Random Effects for All Potential Moderator  
 Variables in All Correlations on All Pairings of Scales*

Effect	Estimate (SE)
Fixed Effects	
Intercept	1.25 (0.21)
Construction	-0.07 (0.58)
Education	-0.25 (0.76)
Food	-0.14 (0.73)
Healthcare	0.12 (0.32)
Information	-0.49 (0.55)
Professional	0.06 (0.48)
Utilities	Reference
<100	0.27 (0.46)
100-500	0.14 (0.36)
501-1000	0.11 (0.35)
>1000	Reference
Scale	0.26 (0.39)
Task	Reference
0-50	0.26 (0.32)
51-100	0.02 (0.42)
>100	Reference
Random Effects	
Study	0.13 (0.07)

The relationship between pairings of individual rating scales by industry is presented in Figure 35. The mean weighted correlations for all pairings of individual scales and their confidence intervals are also presented by industry in Table 19. There was no statistically significant relationship between the industry in which the survey validation study was being performed and the correlations between all individual rating scales,  $F(6,27) = 1.05, p = 0.4140$ , as illustrated in Table 20.

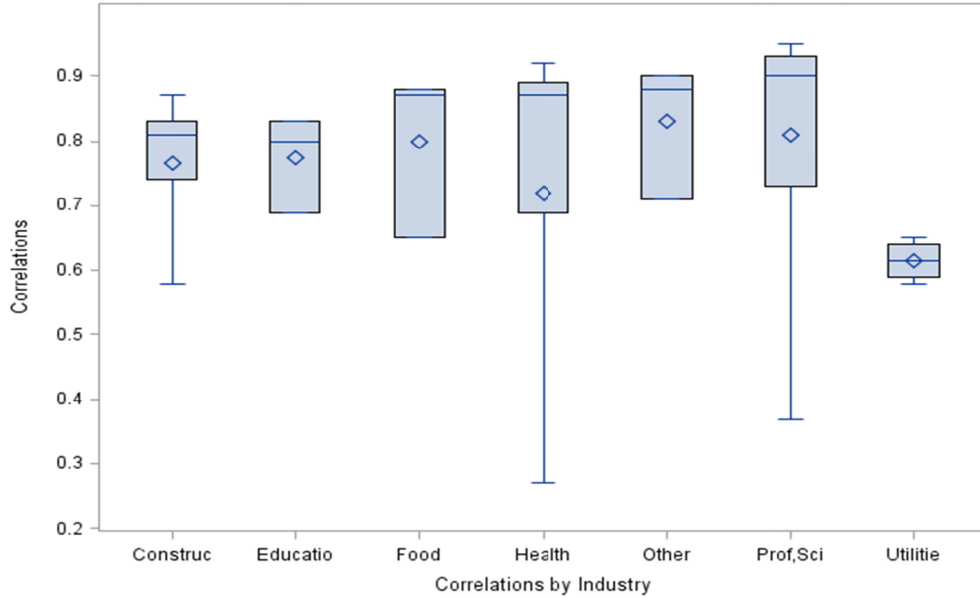


Figure 35. Distribution of correlations between individual rating scales by industry. N=34.

Table 19.  
*Mean Weighted Correlation, CIs and PIs for All Pairings of Individual Scales by Industry*

Industry	Weighted Mean Correlation	95% CI	95% PI
Construction	0.80	[0.63, 0.90]	[0.35 , 0.95]
Education	0.78	[0.39, 0.93]	[0.14 , 0.96]
Food	0.82	[0.48, 0.95]	[0.25 , 0.97]
Healthcare	0.83	[0.71, 0.90]	[0.44 , 0.95]
Information	0.85	[0.70, 0.92]	[0.47 , 0.96]
Professional	0.80	[0.63, 0.90]	[0.35 , 0.95]
Utilities	0.62	[0.38, 0.78]	[0.00 , 0.89]

Note: CI =Confidence Intervals, PI=Prediction Intervals

The relationship between all pairings of individual rating scales by sample size is presented in Figure 36. The mean weighted correlations for all pairings of individual scales and their confidence intervals are also presented by sample size in Table 21. There was no statistically significant relationship between sample size and the correlations between all rating scales,  $F(3,30) = 1.24, p = 0.3127$ , as illustrated in Table 22.

Table 20.

*Fixed and Random Effects for Industries on Correlations*

Effect	Estimate (SE)
<b>Fixed Effects</b>	
Intercept	0.72 (0.16)
Construction	0.38 (0.25)
Education	0.33 (0.36)
Food	0.44 (0.36)
Healthcare	0.46 (0.22)
Information	0.53 (0.25)
Professional	0.39 (0.25)
Utilities	Reference
<b>Random Effects</b>	
Study	0.10 (0.04)

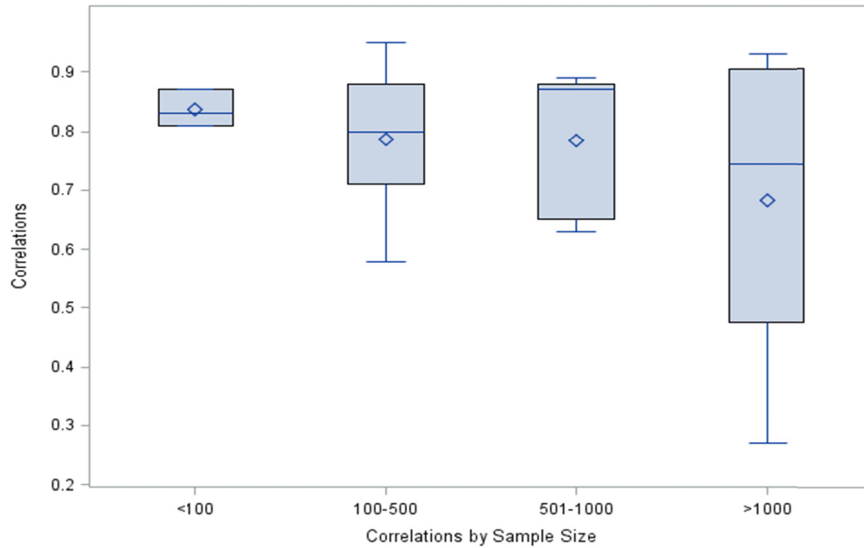


Figure 36. Distribution of correlations between individual rating scales by sample size. N=34.

Table 21.

*Mean Weighted Correlation, CIs, and PIs for All Pairings of Individual Rating Scales By Sample Size*

Sample Size	Weighted Mean Correlation	95% CI	95% PI
<100	0.84	[0.69, 0.92]	[0.45, 0.96]
100-500	0.82	[0.73, 0.88]	[0.44, 0.95]
501-1000	0.80	[0.63, 0.90]	[0.36, 0.95]
>1000	0.70	[0.54, 0.81]	[0.17, 0.91]

Note: CI=Confidence Intervals, PI=Prediction Intervals

Table 22.  
*Fixed And Random Effects for Sample Size on Correlations*

Effect	Estimate (SE)
Fixed Effects	
Intercept	0.86 (0.13)
<100	0.36 (0.23)
100-500	0.29 (0.17)
501-1000	0.24 (0.23)
>1000	Reference
Random Effects	
Study	0.10 (0.04)

The relationship between all combinations of individual rating scales by presentation order is presented in Figure 37. The mean weighted correlations for all combinations individual scales and their confidence intervals are also presented by presentation order in Table 23. There was no statistically significant relationship between presentation order on the correlations between all rating scales,  $F(1,32) = 0.90$ ,  $p = 0.3497$ , as illustrated in Table 24.

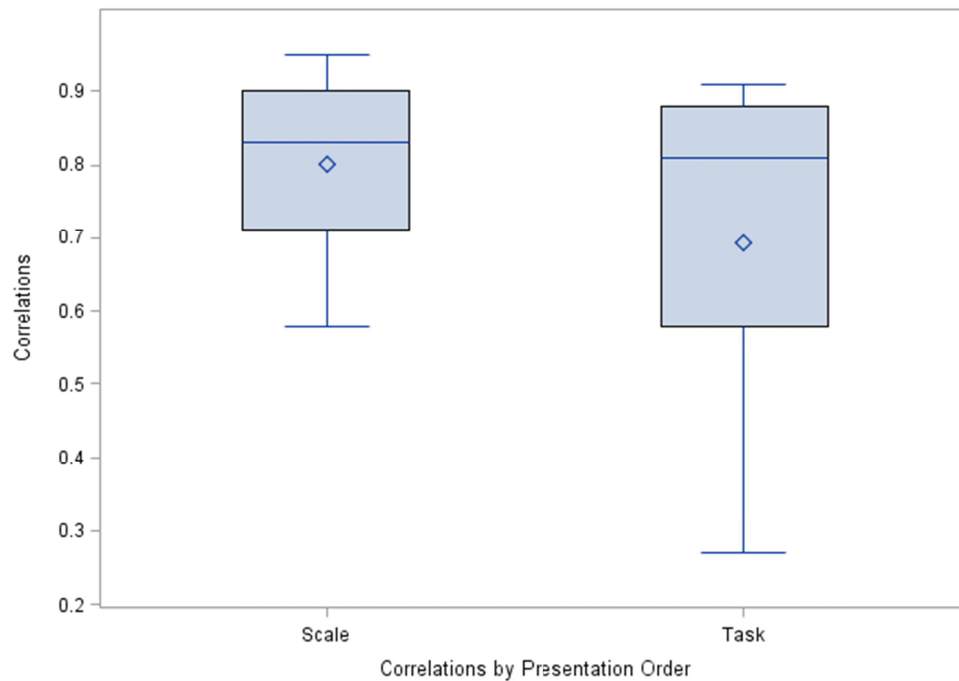


Figure 37. Distribution of correlations between individual rating scales by presentation order. N=34.



Table 23.

*Mean Weighted Correlation, CIs, and PIs for All Pairings of Individual Rating Scales by Presentation Order*

Presentation Order	Weighted Mean Correlation	95% CI	95% PI
By Scale	0.82	[0.72, 0.89]	[0.44, 0.95]
By Task	0.77	[0.68, 0.83]	[0.33, 0.93]

Note: CI =Confidence Intervals, PI=Prediction Intervals

Table 24.

*Fixed and Random Effects for Presentation Order on Correlations*

Effect	Estimate (SE)
Fixed Effects	
Intercept	1.01 (0.09)
Scale	0.15 (0.15)
Task	Reference
Random Effects	
Study	0.10 (0.04)

The relationship between all combinations of individual rating scales by number of tasks is presented in Figure 38. The mean weighted correlations for all combinations of individual rating scales and their confidence intervals are also presented by number of tasks in Table 25. There was no statistically significant relationship between the number of tasks on the correlations obtained between all rating scales,  $F(2,31) = 3.19, p = 0.0550$ , as illustrated in Table 26.

Table 25.

*Mean Weighted Correlation, CIs, and PIs for All Pairings of Individual Rating Scales by Number of Tasks.*

Number of Tasks	Weighted Mean Correlation	95% CI	95% PI
0-50 Tasks	0.84	[0.78, 0.89]	[0.55, 0.95]
51-100 Tasks	0.75	[0.59, 0.85]	[0.31, 0.92]
More than 100 Tasks	0.70	[0.55, 0.80]	[0.23, 0.90]

Note: CI =Confidence Intervals, PI=Prediction Intervals

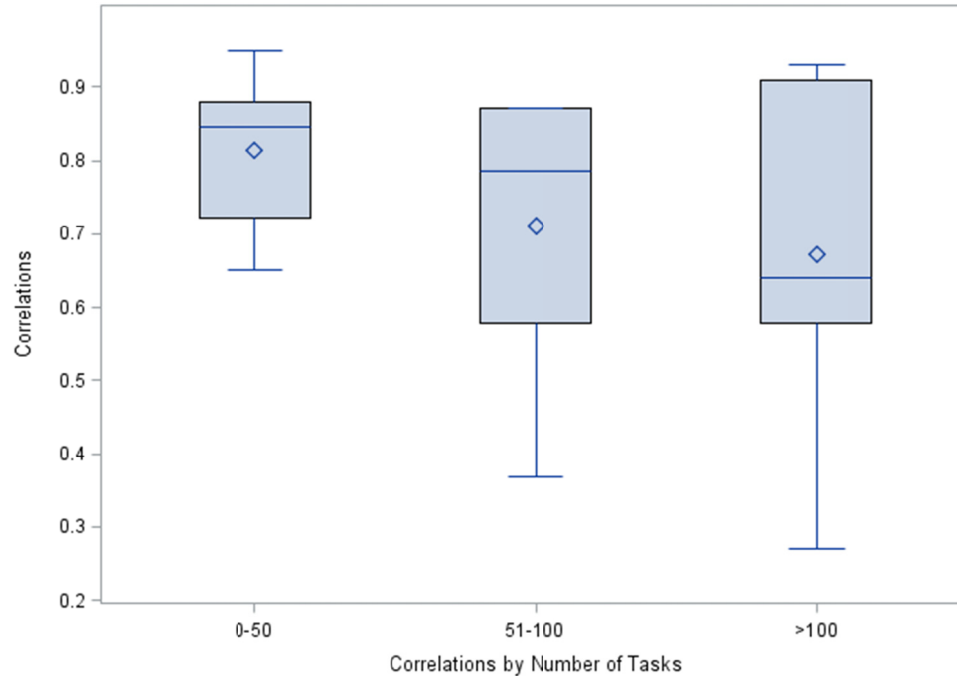


Figure 38. Distribution of correlations between individual rating scales by number of tasks. N=34.

Table 26.

*Fixed and Random Effects for Number of Tasks on Correlations*

Effect	Estimate (SE)
Fixed Effects	
Intercept	0.86 (0.12)
0-50	0.37 (0.15)
51-100	0.11 (0.19)
>101	Reference
Random Effects	
Study	0.08 (0.03)

There was no statistically significant effect of all four variables together on the correlations between all rating scales, as illustrated in Table 27.

*Research Questions Two and Three for Correlations between Composite Scales*

The same five ANOVAs were computed with only pairings of composite scales. The relationship between pairings of composite rating scales by industry is presented in Figure 39. The mean weighted correlations for all pairings of composite scales and their

confidence intervals are also presented by industry in Table 28. There was no statistically significant relationship between industry for which the survey validation study was performed and the correlations between all rating scales,  $F(4,10) = 0.60$ ,  $p = 0.6699$ , as illustrated in Table 29.

Table 27.  
*Fixed and Random Effects for All Potential Moderator Variables on All Correlations of Pairings of Individual Scales*

Effect	Estimate (SE)
Fixed Effects	
Intercept	0.52 (0.24)
Construction	0.52 (0.56)
Education	0.82 (0.77)
Food	0.53 (0.69)
Healthcare	0.58 (0.38)
Information	0.60 (0.47)
Professional	0.71 (0.47)
Utilities	Reference
<100	0.49 (0.49)
100-500	0.40 (0.35)
501-1000	0.40 (0.38)
>1000	Reference
Scale	-0.03 (0.31)
Task	Reference
0-50	-0.26 (0.43)
51-100	-0.44 (0.43)
>100	Reference
Random Effects	
Study	0.14 (0.07)

The relationship between all combinations of composite rating scales by sample size is presented in Figure 40. The mean weighted correlations for all combinations of composite scales and their confidence intervals are also presented by sample size in Table 30. There was no statistically significant relationship between sample size and the

correlations between all rating scales,  $F(2,12) = 0.66$ ,  $p = 0.5364$ , as illustrated in Table 31.

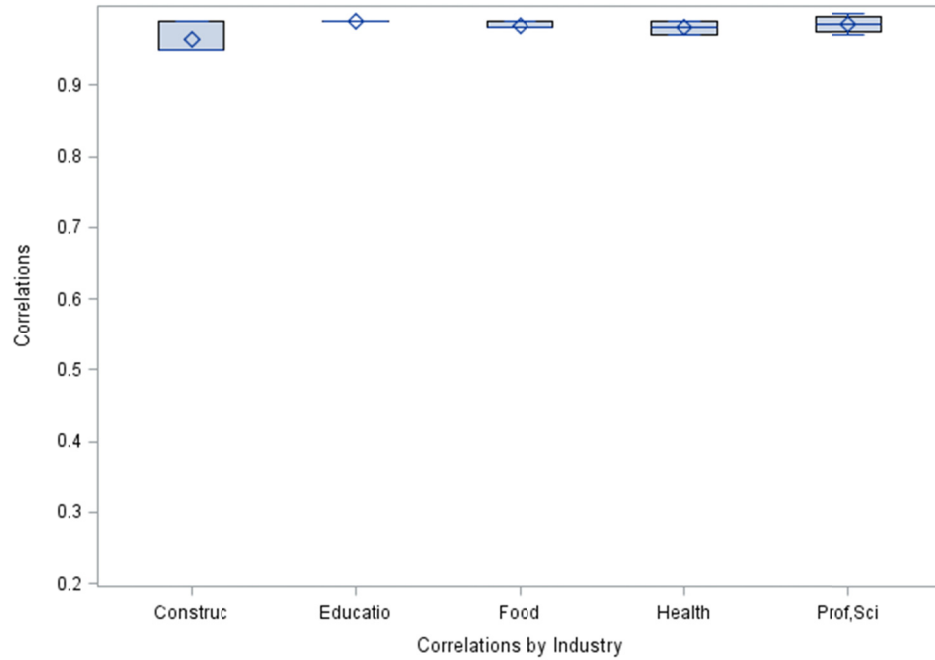


Figure 39. Distribution of correlations between pairings of composite rating scales by industry. N=15.

Table 28.  
Mean Weighted Correlation, CIs, and PIs for All Combinations of Composite Rating Scales by Industry.

Industry	Weighted Mean Correlation	95% CI	95% PI
Construction	0.97	[0.80, 0.96]	[0.61, 1.00]
Education	0.99	[0.93, 1.00]	[0.86, 1.00]
Food	0.98	[0.89, 1.00]	[0.76, 1.00]
Healthcare	0.98	[0.92, 0.99]	[0.79, 1.00]
Professional	0.99	[0.97, 1.00]	[0.93, 1.00]

Note: CI =Confidence Intervals, PI=Prediction Intervals

Table 29.  
*Fixed and Random Effects for Industries on Correlations*

Effect	Estimate (SE)
Fixed Effects	
Intercept	2.86 (0.35)
Construction	-0.78 (0.60)
Education	-0.21 (0.60)
Food	-0.50 (0.60)
Healthcare	-0.60 (0.49)
Professional	Reference
Random Effects	
Study	0.24 (0.24)

The relationship between all combinations of composite rating scales by presentation order is presented in Figure 41. The mean weighted correlations for all combinations of composite scales and their confidence intervals are also presented by presentation order in Table 32. There was no statistically significant relationship between presentation order and the correlations between all rating scales,  $F(1,13) = 0.81$ ,  $p = 0.3851$ , as illustrated in Table 33.

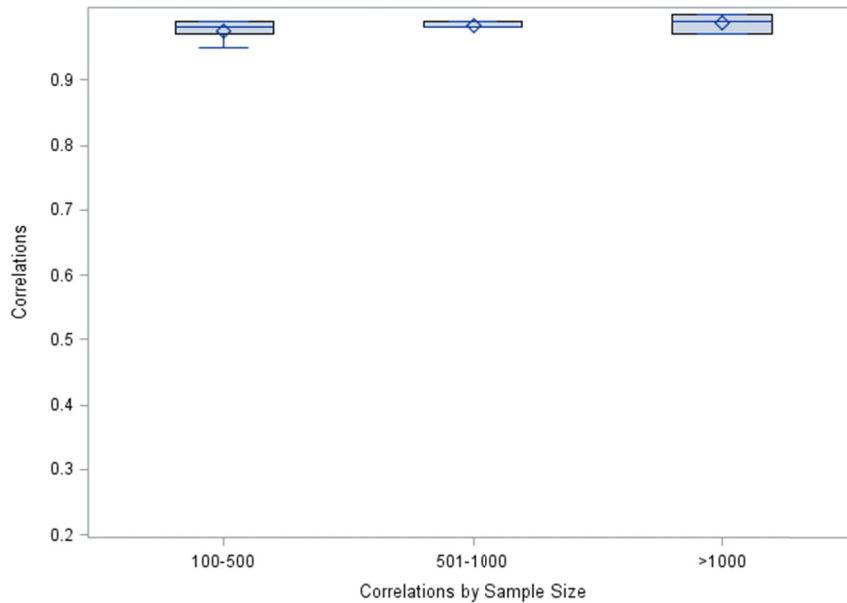


Figure 40. Distribution of correlations between pairings of composite rating scales by sample size. N=15.

Table 30.  
*Mean Weighted Correlation, CIs, and PIs for All Combinations of Composite Rating Scales by Sample Size.*

Sample Size	Weighted Mean Correlation	95% CI	95% PI
100-500	0.98	[0.95, 0.99]	[0.86, 1.00]
501-1000	0.98	[0.90, 1.00]	[0.80, 1.00]
>1000	0.99	[0.98, 1.00]	[0.93, 1.00]

Note: CI =Confidence Intervals, PI=Prediction Intervals

Table 31.  
*Fixed and Random Effects for Sample Size on Correlations*

Effect	Estimate (SE)
Fixed Effects	
Intercept	2.70 (0.26)
100-500	-0.41 (0.37)
501-1000	-0.34 (0.52)
>1000	Reference
Random Effects	
Study	0.20 (0.14)

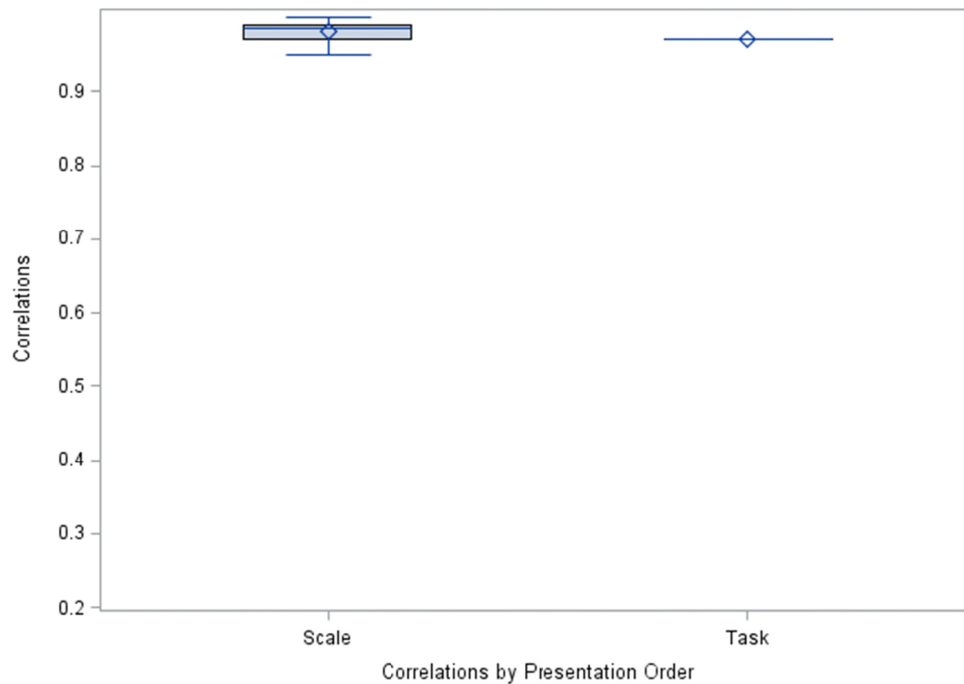


Figure 41. Distribution of correlations between pairings of composite rating scales by presentation order. N=15.

Table 32.  
*Mean Weighted Correlation, CIs, and PIs for All Combinations of Composite Rating Scales by Presentation Order*

Presentation Order	Weighted Mean Correlation	95% CI	95% PI
By Scale	0.99	[0.98, 0.99]	[0.93 , 1.00]
By Task	0.97	[0.86, 0.99]	[0.73 , 1.00]

Note: CI =Confidence Intervals, PI=Prediction Intervals

Table 33.  
*Fixed and Random Effects for Presentation Order on Correlations*

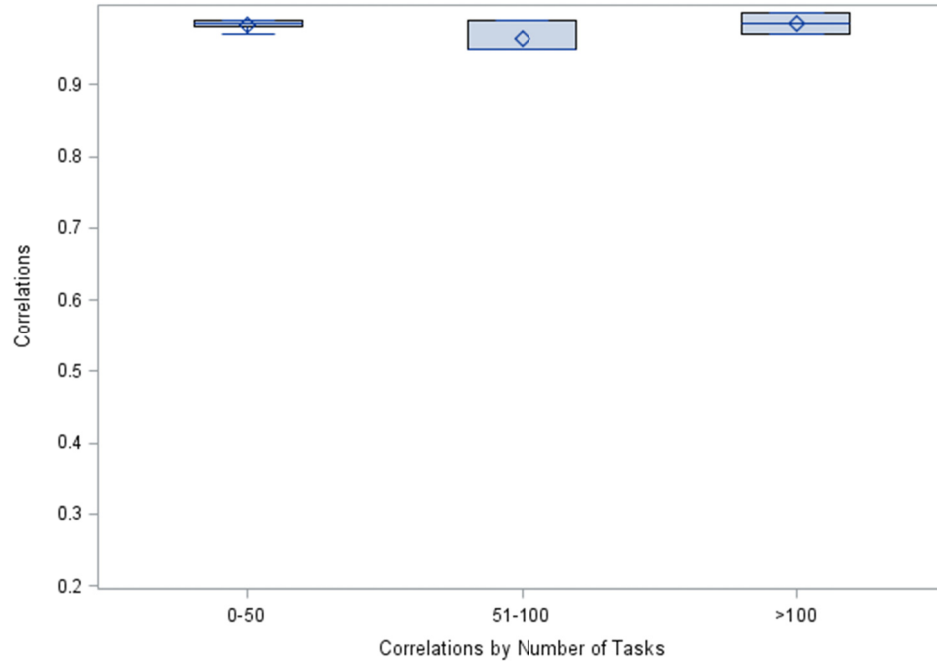
Effect	Estimate (SE)
Fixed Effects	
Intercept	2.13 (0.43)
Scale	0.42 (0.46)
Task	Reference
Random Effects	
Study	0.18 (0.12)

The relationship between all combinations of composite rating scales by number of tasks is presented in Figure 42. The mean weighted correlations for all combinations of composite scales and their confidence intervals are also presented by number of tasks in Table 34. There was no statistically significant relationship between the number of tasks on the correlations obtained between all rating scales,  $F(2,12) = 0.71, p = 0.5089$ .

Table 34.  
*Mean Weighted Correlation, CIs, and PIs for All Combinations of Composite Rating Scales by Number of Tasks*

Number of Tasks	Weighted Mean Correlation	95% CI	95% PI
0-50 Tasks	0.99	[0.97, 0.99]	[0.90 , 1.00]
51-100 Tasks	0.97	[0.84, 0.99]	[0.69 , 1.00]
More than 100 Tasks	0.99	[0.97, 1.00]	[0.93 , 1.00]

Note: CI =Confidence Intervals, PI=Prediction Intervals



*Figure 42.* Distribution of correlations between pairings of composite rating scales by number of tasks. N=15.

When all four potential moderator variables were put into one model, the large number of empty cells prevented the estimation of all relevant relationships.

*Research Questions Two and Three for Correlations between Individual and Composite Scales*

The same five ANOVAs were computed with pairings of individual and composite rating scales. The relationship between pairings of individual and composite rating scales by industry is presented in Figure 43. The mean weighted correlations for all pairings of individual and composite scales and their confidence intervals are also presented by industry in Table 35. There was no statistically significant relationship between the industry in which the survey was conducted and the correlations between all rating scales,  $F(6,73) = 1.63, p = 0.1501$ , as illustrated in Table 36.



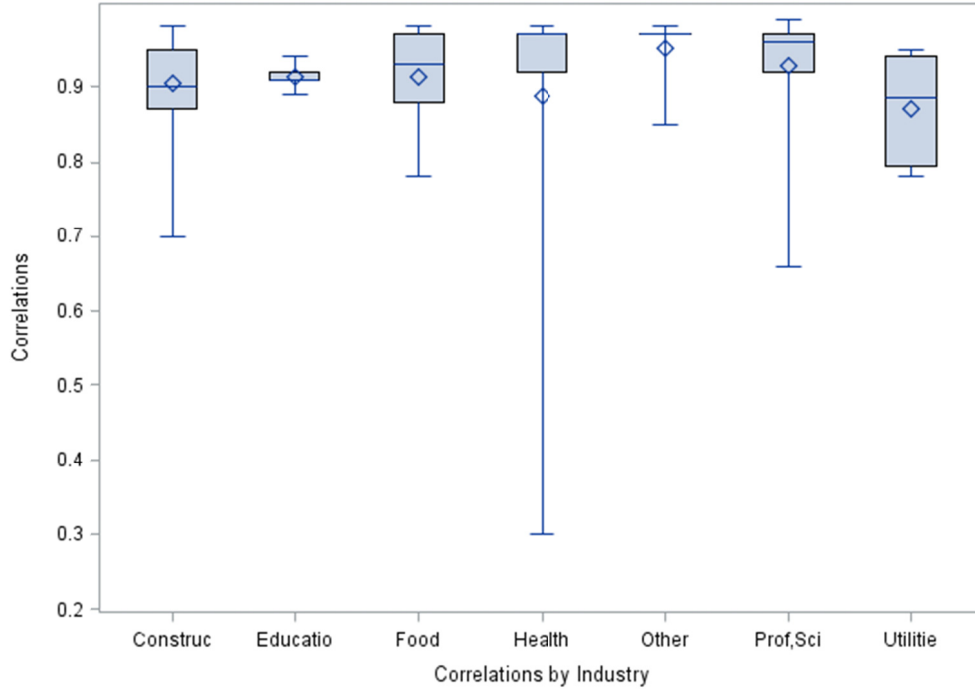


Figure 43. Distribution of correlations between individual and composite rating scales by industry. N=80.

Table 35.

Mean Weighted Correlation, CIs, and PIs for Pairings of Individual and Composite Rating Scales by Industry

Industry	Weighted Mean Correlation	95% CI	95% PI
Construction	0.94	[0.83, 0.96]	[0.77, 0.99]
Education	0.92	[0.89, 0.97]	[0.62, 0.98]
Food	0.94	[0.78, 0.97]	[0.71, 0.99]
Healthcare	0.96	[0.84, 0.98]	[0.83, 0.99]
Information	0.96	[0.93, 0.97]	[0.75, 0.98]
Professional	0.94	[0.93, 0.98]	[0.76, 0.99]
Utilities	0.89	[0.83, 0.94]	[0.74, 0.98]

Note: CI =Confidence Intervals, PI=Prediction Intervals

The relationship between pairings of individual and composite rating scales by sample size is presented in Figure 44. The mean weighted correlations for pairings of individual and composite scales and their confidence intervals are also presented by sample size in Table 37. There was no statistically significant relationship between

sample size and the correlations between all rating scales,  $F(3,76) = 1.50, p = 0.2220$ , as illustrated in Table 38.

Table 36.  
*Fixed and Random Effects for Industries on Correlations*

Effect	Estimate (SE)
Fixed Effects	
Intercept	1.67 (0.13)
Construction	0.08 (0.24)
Education	-0.10 (0.36)
Food	0.07 (0.36)
Healthcare	0.22 (0.20)
Information	0.04 (0.36)
Professional	0.06 (0.24)
Utilities	Reference
Random Effects	
Study	0.11 (0.04)

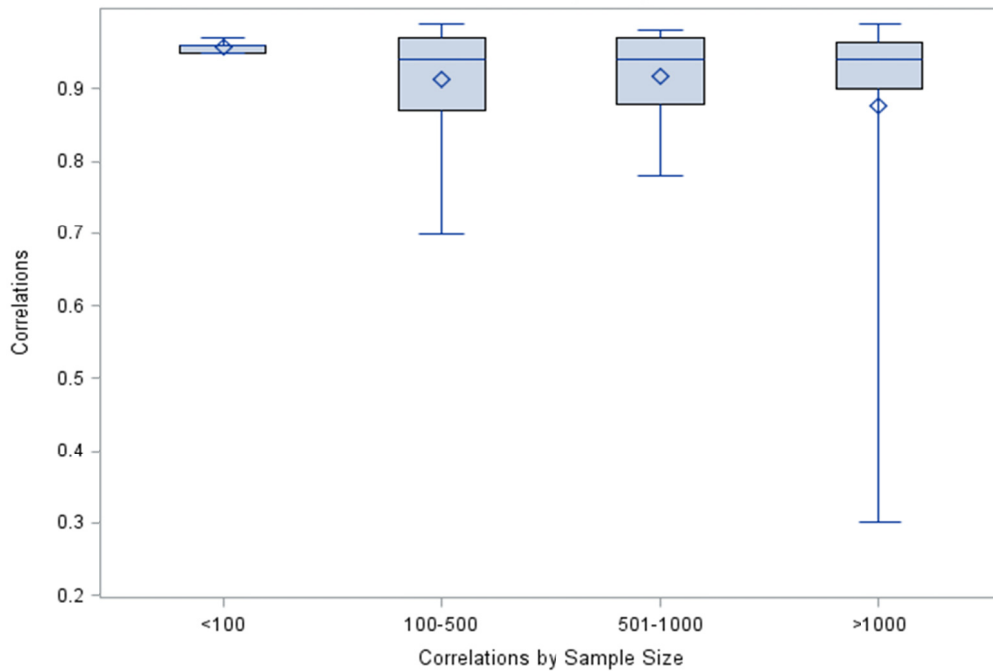


Figure 44. Distribution of correlations between individual and composite rating scales by sample size. N=80.

Table 37.  
*Mean Weighted Correlation, CIs, and PIs for Pairings of Individual and Composite Rating Scales by Sample Size*

Sample Size	Weighted Mean Correlation	95% CI	95% PI
<100	0.96	[0.92, 0.98]	[0.86 , 0.99]
100-500	0.95	[0.92, 0.96]	[0.84 , 0.98]
501-1000	0.94	[0.90, 0.97]	[0.81 , 0.98]
>1000	0.91	[0.87, 0.94]	[0.74 , 0.97]

Note: CI =Confidence Intervals, PI=Prediction Intervals

Table 38.  
*Fixed and Random Effects for Sample Size on Correlations*

Effect	Estimate (SE)
Fixed Effects	
Intercept	1.55 (0.12)
<100	0.38 (0.20)
100-500	0.25 (0.15)
501-1000	0.22 (0.20)
>1000	Reference
Random Effects	
Study	0.08 (0.03)

The relationship between the pairings of individual and composite rating scales by presentation order is presented in Figure 45. The mean weighted correlations for all pairings of individual and composite scales and their confidence intervals are also presented by presentation order in Table 39. There was no statistically significant relationship between presentation order and the correlations between all rating scales,  $F(1,78) = 0.00, p = 0.9753$ , as illustrated in Table 40.

Table 39.  
*Mean Weighted Correlation, CIs, and PIs for Pairings of Individual and Composite Rating Scales by Presentation Order*

Presentation Order	Weighted Mean Correlation	95% CI	95% PI
By Scale	0.94	[0.91, 0.96]	[0.8 , 0.98]
By Task	0.94	[0.92, 0.96]	[0.81 , 0.98]

Note: CI =Confidence Intervals, PI=Prediction Intervals

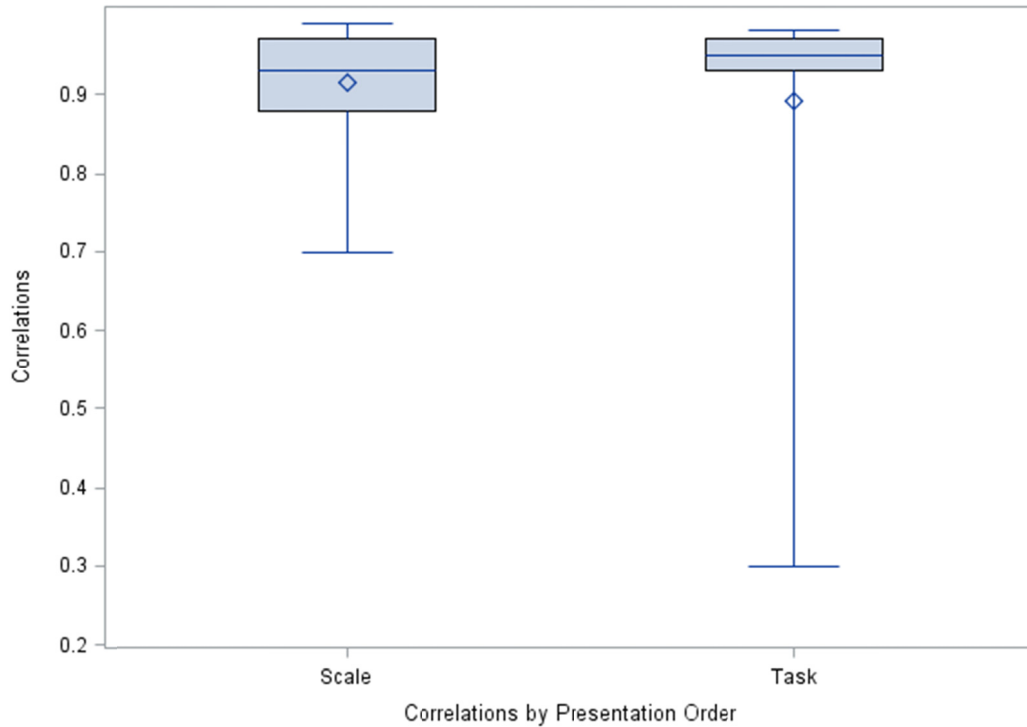


Figure 45. Distribution of correlations between individual and composite rating scales by presentation order. N=80.

Table 40.

*Fixed and Random Effects for Presentation Order on Correlations*

Effect	Estimate (SE)
Fixed Effects	
Intercept	1.75 (0.08)
Scale	-0.00 (0.14)
Task	Reference
Random Effects	
Study	0.09 (0.03)

The relationship between all pairings of individual and composite rating scales by number of tasks is presented in Figure 46. The mean weighted correlations for all combinations of composite scales and their confidence intervals are also presented by number of tasks in Table 41. There was no statistically significant relationship between the number of tasks on the correlations obtained between all rating scales,  $F(2,77) = 2.94$ ,  $p = 0.0588$ , as illustrated in Table 42.

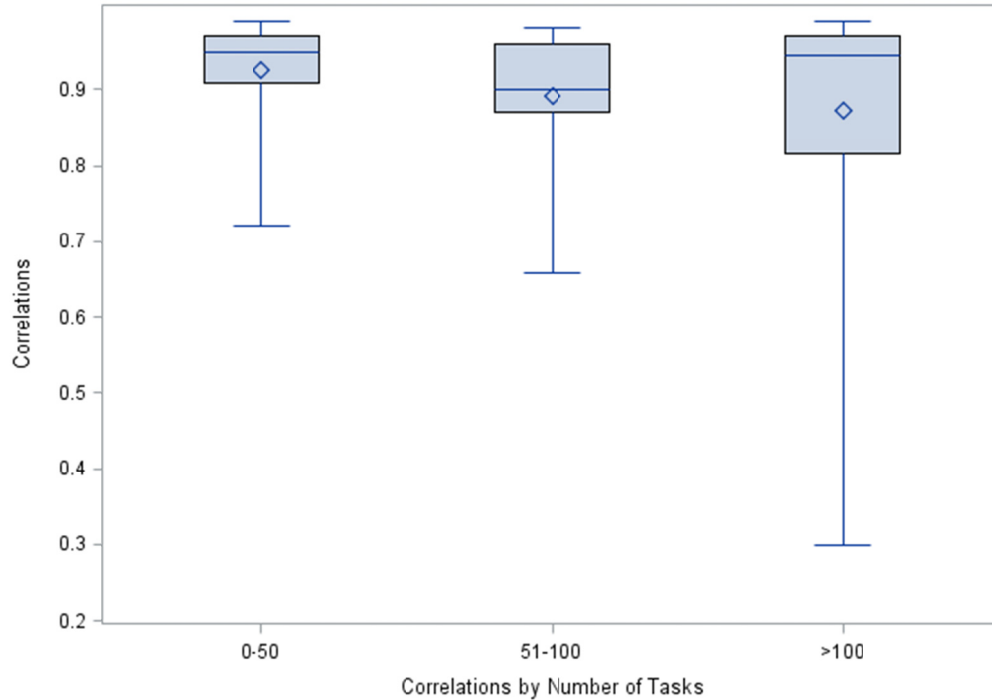


Figure 46. Distribution of correlations between individual and composite rating scales by number of tasks. N=80.

Table 41.  
Mean Weighted Correlation, CIs, and PIs for Pairings of Individual and Composite Rating Scales by Number of Tasks.

Number of Tasks	Weighted Mean Correlation	95% CI	95% PI
0-50 Tasks	0.95	[0.94, 0.97]	[0.80, 0.99]
51-100 Tasks	0.93	[0.89, 0.96]	[0.65, 0.99]
More than 100 Tasks	0.92	[0.87, 0.94]	[0.62, 0.98]

Note: CI =Confidence Intervals, PI=Prediction Intervals

Table 42.  
Fixed and Random Effects for Number of Tasks on Correlations

Effect	Estimate (SE)
Fixed Effects	
Intercept	1.56 (0.11)
0-50	0.32 (0.14)
51-100	0.10 (0.17)
>101	Reference
Random Effects	
Study	0.07 (0.02)

There was no statistically significant effect of all four variables together on the correlations between all rating scales, as illustrated in Table 43.

Table 43.  
*Fixed and Random Effects for All Potential Moderator Variables on All Correlations of Pairings of Individual and Composite Rating Scales.*

Effect	Estimate (SE)
Fixed Effects	
Intercept	1.50 (0.22)
Construction	-0.14 (0.60)
Education	-0.25 (0.77)
Food	-0.20 (0.74)
Healthcare	0.08 (0.33)
Information	-0.23 (0.56)
Professional	0.05 (0.49)
Utilities	Reference
<100	0.29 (0.497)
100-500	0.12 (0.36)
501-1000	0.11 (0.35)
>1000	Reference
Scale	0.05 (0.40)
Task	Reference
0-50	0.28 (0.33)
51-100	0.06 (0.43)
>100	Reference
Random Effects	
Study	0.13 (0.07)

### Research Question Four Results

Examination blueprints were created for all of the 20 survey validation studies, derived from all individual and composite rating scales used in each study. The blueprint weights were analyzed at the overarching duty or content area (ranging from four duty areas to 26 duty areas), rather than the individual tasks. This is due to the fact that many

organizations only publish examination blueprint weights at the duty level, as illustrated in Figure 47 below.

Duties and Tasks	Exam Weight
Duty A	40%
Task A.01	
Task A.02	
Task A.03	
Duty B	45%
Task B.01	
Task B.02	
Task B.03	
Task B.04	
Duty C	15%
Task C.01	
Task C.02	
Totals	100%

*Figure 47.* Sample examination blueprint.

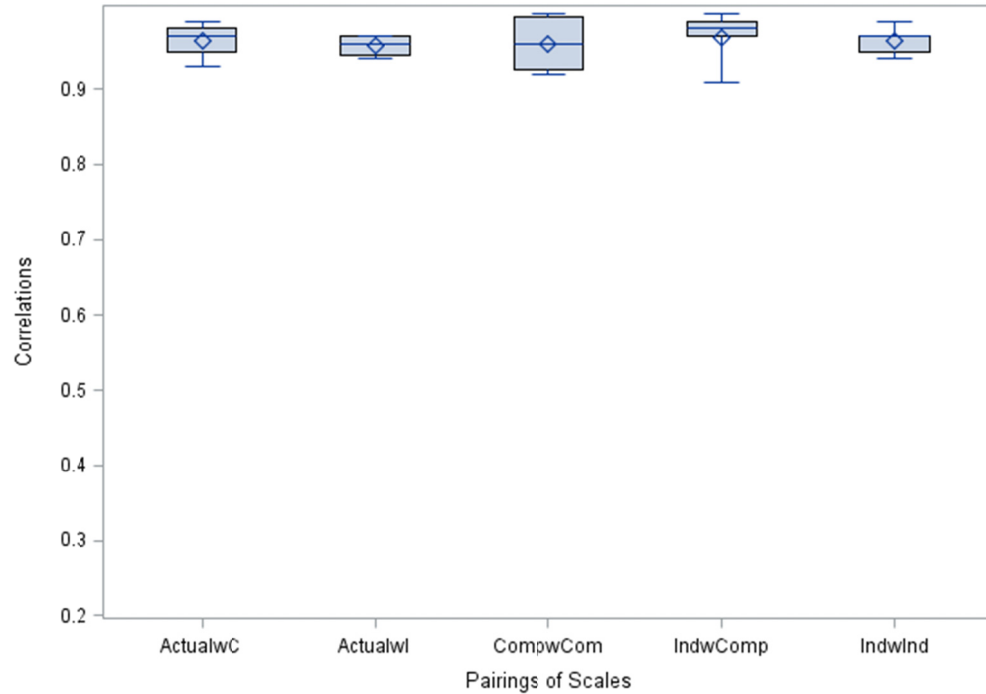
The calculated examination blueprints were compared to each other, as well as to the actual examination blueprint weights used for each of the licensure or certification exams for which the job analysis was performed. Each overarching duty area was rank ordered from greatest weight on an exam to the least weight on an exam (a “1” meant that the greatest portion of the exam was devoted to that section, a “2” meant the next greatest portion of the exam was devoted to that section, and so on). The relative ranking for each duty area on all derived examination blueprints was correlated with the rank order on the actual examination blueprint used for a licensure or certification exam. The relative ranking for each duty area on all derived examination blueprints were also compared with each other. The range of obtained correlations from relative rankings between pairings of

derived examination blueprints and derived examination blueprints with actual examination blueprints was .91 to 1.00, as illustrated in Table 44 and Figure 48.

Table 44.  
*Correlations Between Relative Rankings of Duty Areas on Derived Examination Blueprints and Actual Examination Blueprints*

Variable	With Variable	N	Correlation
Importance	Composite4	31	1.00
Composite1	Composite4	31	1.00
Importance	Composite1	114	0.99
Importance	Composite3	30	0.99
Criticality	Composite3	30	0.99
Actual Exam	Composite4	31	0.99
Importance	Criticality	34	0.99
Composite1	Composite3	30	0.99
Frequency	Composite4	31	0.98
Criticality	Composite1	34	0.98
Frequency	Composite1	114	0.98
Need	Composite4	31	0.98
Frequency	Composite2	128	0.97
Actual Exam	Composite1	114	0.97
Actual Exam	Importance	114	0.97
Frequency	Importance	114	0.97
Importance	Need	31	0.97
Need	Composite1	31	0.97
Actual Exam	Frequency	212	0.97
Actual Exam	Composite2	128	0.97
Actual Exam	Need	31	0.95
Frequency	Need	31	0.95
Frequency	Criticality	132	0.94
Actual Exam	Criticality	132	0.94
Criticality	Composite2	128	0.94
Actual Exam	Composite3	30	0.93
Composite1	Composite2	30	0.93
Frequency	Composite3	30	0.92
Composite2	Composite3	30	0.92
Importance	Composite2	30	0.91





*Figure 48.* Distribution of correlations between relative ranks of examination blueprints derived from individual and composite scale, as well as the actual examination blueprints used on the licensure or certification exam. N=30.

In addition to computing the relative differences between the rank order of the examination blueprint weights derived from all individual and composite scales with the actual examination blueprints used on the licensure or certification exam, the absolute differences between the weights derived from individual and composite scales with weights from the actual examination blueprints were computed. To do this, the absolute difference between the percent of the exam devoted to each overarching duty area when the exam blueprint was derived from individual or composite rating scales and the percent of the exam devoted to each overarching duty area from the actual examination blueprint was computed. An example of this is provided in Table 45. In this example, there are six duty areas. Three duty weights are provided for examination blueprints derived from individual scales (Frequency, Importance, and Criticality), three duty

weights are provided for examination blueprints derived from composite scales (Composite 1, 2, and 3), and the duty weights for the actual exam is provided for comparison.

Table 45.

*Comparison Between Duty Weights on Actual and Derived Examination Blueprints for a Certification Exam*

Scales	Duty 1	Duty 2	Duty 3	Duty 4	Duty 5	Duty 6	Totals
Actual	17.29%	24.31%	26.13%	11.28%	11.75%	9.24%	100.00%
Frequency	17.15%	21.92%	28.58%	11.70%	10.88%	9.77%	100.00%
Importance	16.89%	25.00%	26.16%	11.31%	11.43%	9.22%	100.00%
Criticality	18.20%	24.17%	25.28%	10.95%	12.38%	9.02%	100.00%
Comp1	16.98%	23.87%	27.05%	11.45%	11.23%	9.42%	100.00%
Comp2	17.58%	19.98%	30.19%	11.15%	11.23%	9.88%	100.00%
Comp3	17.34%	23.96%	26.54%	11.31%	11.56%	9.31%	100.00%

The distribution of absolute differences between the weights on each duty area from examination blueprints derived from individual and composite scales and the weights on the duty areas from the actual examination blueprints is presented in Table 46. The weight represents the percentage of an exam devoted to a specific duty area. For example, if one of the duty areas was represented by 20% of the examination blueprint, that means that 20% of the items on the test should be written to that duty area. If one were to imagine a 100-item exam, 20% of those items on one duty area would mean 20 items written to that duty area.

One of the 20 studies had large absolute differences between all derived examination blueprints and the actual examination blueprint, was considered an outlier, and was not included in Table 45. The single study that was considered an outlier had four overarching duty areas. The absolute differences in the duty areas ranged from a low of 7.64% to a high of 14.80%. If one were to imagine the same 100-item exam, the absolute differences between each duty area on the derived examination blueprints and

the actual duty areas on the licensure or certification exam, a 7.64% to 14.80% absolute difference reflects a large change in the number of items devoted to each content area (7 or 8 items to 14 or 15 item differences between two examination blueprints).

Table 46.  
*Distribution of Absolute Differences Between the Weights of Derived Exam Blueprints and Actual Exam Blueprints*

Scale	Mean (SD)	Range	N
Frequency	1.19 (1.08)	[0.35,5.10]	19
Importance	1.10 (1.49)	[0.15,5.33]	14
Criticality	0.94 (1.01)	[0.08,3.47]	9
Need at Entry	1.06 (0.74)	[0.63,1.91]	3
Composite1	0.89 (1.47)	[0.00,5.22]	14
Composite2	0.92 (0.73)	[0.27,2.30]	9
Composite3	0.83 (1.47)	[0.02,3.03]	4
Composite4	0.43 (0.49)	[0.12,0.99]	3

The average absolute difference between the weights on exam blueprints derived from Composite 4 and the weights on the actual examination blueprints was the smallest. The average absolute difference between the weights from those two examination blueprints was 0.43%. Again, considering a 100-item exam, that represents less than one item difference between the number of items devoted to each content area on the examination blueprint derived from Composite 4 when compared to the actual examination blueprint. The greatest absolute difference between weights on derived exam blueprints and weights on actual exam blueprints was observed when exam blueprints were derived from the Frequency rating scale. The average absolute differences between the percent of the exam devoted to each content area when the exam was weighted using only the Frequency rating scales was on average 1.19% different than the percent of the exam devoted to each content area on the actual examination blueprint. Even though this the greatest absolute difference observed between the weights on all of

the content areas on the actual examination blueprints and the weights on all of the content areas on the derived examination blueprints, this is still a relatively small number.

When considering both relative and absolute differences between examination blueprints, there were three studies that had large relative rank order differences and absolute percentage differences between all derived examination blueprints and the actual examination blueprints. In those cases, it is possible that an examination committee made many modifications to the examination blueprint after the survey validation study.

In addition to comparing the absolute differences in duty weights on derived examination blueprints with actual examination blueprints, the absolute differences of duty weights on derived examination blueprints and actual examination blueprints were compared with examination blueprints in which all of the duty areas were equally weighted. For example, if one were to image an examination blueprint with five duty areas, all five duty areas would be worth 20% on the overall exam. If one were to imagine an examination blueprint with 25 duty areas, each duty area would be worth 4% on the overall exam.

The range of absolute differences between duty areas on actual examination blueprints and duty areas in that are all equally weighted was 9.76 to 1.47, as illustrated in Table 47. The mean absolute difference between duty areas on actual examination blueprints and duty areas on equally weighted blueprints was 4.76. Imagining our 100-item exam, this means that if we were to equally weight all of the duty areas on an exam compared to the weights of a real 100-items exam, there would be on average a 4.76 item difference between the two blueprints.

Table 47.

*Average Absolute Differences Between Duty Weights on Actual and Derived Examination Blueprints and Exam Blueprints in which All Duties are Equally Weighted*

Study	Actual	Freq	Imp	Crit	Need	Comp1	Comp2	Comp3	Comp4	N Duties
15	9.04	7.80	6.96			7.24				4
13	4.00	12.72	12.13			12.38				4
14	4.40	5.10	4.96			5.02				5
5	5.99	5.94	6.08			6.03				6
4	5.91	5.88	6.01	5.89		5.97	5.92	5.94		6
1	9.76	11.30	9.22		9.80	9.75			9.77	7
3	9.48	10.12	9.13			9.47				7
8	2.53	2.86	2.55	2.57		2.52	3.05	2.54		7
2	7.13	7.12	7.24			7.20				8
9	4.06	4.90	4.41	4.45		4.53	5.00	4.42		8
6	3.94	3.54	3.90	4.07		3.79	3.45	3.90		9
12	3.00	2.72	3.24			3.00				9
11	4.99	5.29	4.89			4.99				10
16	3.37	3.26	3.25		3.64	3.25			3.32	11
10	1.47	1.51		1.59			1.53			12
7	4.41	4.48	4.39		4.42	4.42			4.42	13
20	3.49	3.26		3.04			3.45			17
18	2.94	2.84		2.83			2.96			19
17	2.88	2.80		2.95			2.84			24
19	2.51	2.46		2.68			2.58			26
Mean	4.76	5.30	5.89	3.34	5.95	5.97	3.42	4.2	5.84	11
SD	2.41	3.11	2.66	1.27	3.35	2.80	1.31	1.41	3.45	
Lower	1.47	1.51	2.55	1.59	3.64	2.52	1.53	2.54	3.32	4
Upper	9.76	12.72	12.13	5.89	9.80	12.38	5.92	5.94	9.77	26
N	20	20	15	9	3	15	9	4	3	20

The average absolute differences between duty areas on actual or derived blueprints and duty areas on examination blueprints in which all duty areas are equal appear to decrease as the number of duty areas increases, as illustrated in Figure 49.

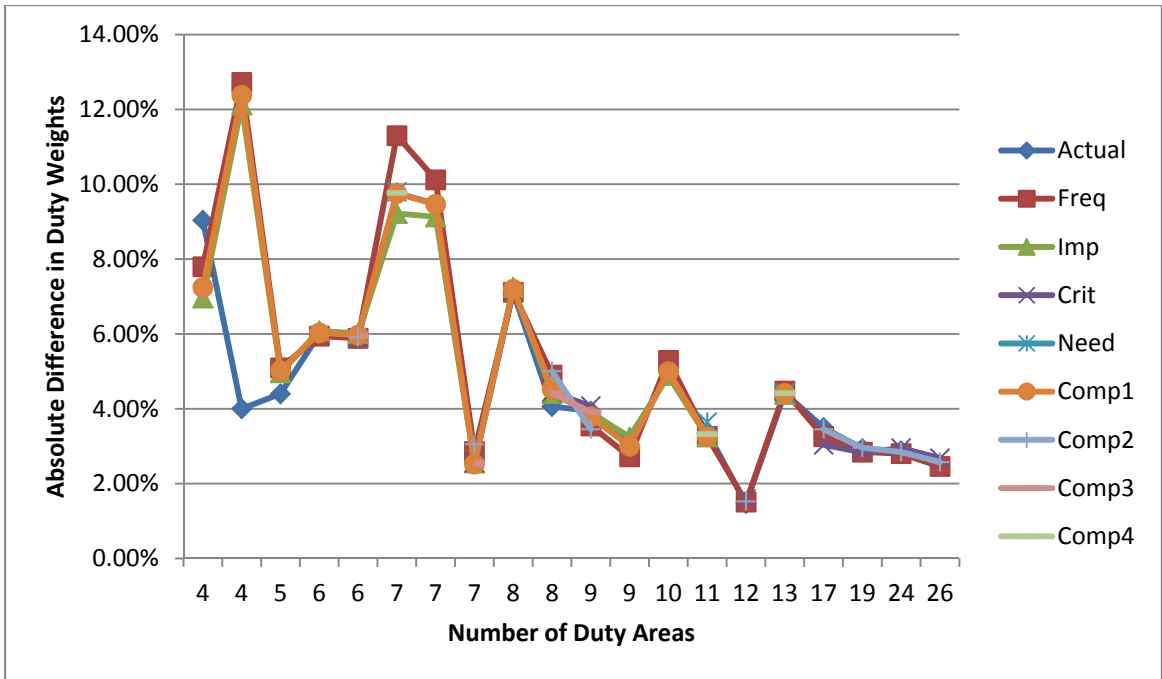


Figure 49. Absolute differences between duty areas on actual and derived examination blueprints and duty areas on examination blueprints in which all duty areas are equal.

## CHAPTER FIVE: DISCUSSION

The purpose of this study was to determine the relationship between individual and composite rating scales, examine how that relationship varies across industries, sample sizes, task presentation order, and number of tasks rated; and evaluate whether examination blueprint weights would differ based on the rating scales or composites of scales used to establish blueprints weights. A secondary data analysis was performed using data from survey validation studies from 20 different job or task analyses in which the industry for which the study was performed, the number of respondents, the order in which respondents rate tasks, and the number of tasks rated varied.

SAS 9.3 was used to calculate correlations between pairings of individual rating scales, pairings of composite rating scales, and pairings of individual rating scales with composite rating scales. To determine if the correlations between all pairings of rating scales varied based upon the four proposed moderator variables (industry, sample size, presentation order, and number of tasks) five ANOVAs were computed for all pairings of rating scales, pairings of only individual rating scales, pairings of only composite rating scales, and pairings of individual rating scales with composite rating scales. In total, 20 models were analyzed to determine if 1) there was a relationship between scales and 2) if there was a relationship between scales, did that relationship vary based on any of the four proposed moderating factors.

Additionally, examination blueprints derived from each individual and composite rating scale were compared to actual examination blueprints used on the licensure or

certification exams for which the 20 sample studies were performed. Comparisons were made between the weights devoted to each overarching duty area on each of the 20 exams with the derived examination blueprints weights to determine the extent to which derived examination blueprints varied from actual examination blueprints. In short, did it matter what scales were used to derive the examination blueprint or would examination blueprints look roughly the same regardless of how the weights were derived?

### **Summary of Individual and Composite Rating Scale Findings**

#### *Importance and Criticality Rating Scales*

There was a strong relationship between pairs of individual rating scales, pairs of composite rating scales, and individual and composite rating scales. When only considering the relationships between pairings of individual rating scales, the strongest relationship, defined by the largest correlation, was between Importance and Criticality rating scales with an unweighted average correlation of .85. This finding is not unique to this study. Both Sanchez and Levine (1989) and Sanchez and Fraser (1992) reported finding a strong relationship between Importance and Criticality rating scales. Sanchez and Levine report correlations between .78 and .90 for Importance and Criticality ratings, while Sanchez and Fraser reported correlations between .60 and .99. In this study, the range of correlations between Importance and Criticality ratings was between .72 and .95, which is in line with findings from the previous two studies.

It seems reasonable that the relationship between these two scales would be strong, as evaluating the *importance* of performing a task is not unlike evaluating how *critical* successful performance of that task is to a job or how great the *consequence of error* is if the task is performed incorrectly or not at all. For example, the task of



“Verifying patient identification” for the medical professional is considered highly important. It is also *critical* that someone working in the medical profession verifies patient identification and failure to perform this task, or performing it incorrectly could result in a huge consequence of error. In short, the perception of importance and criticality may be highly related, and thus asking a person to rate both Importance and Criticality may be redundant.

#### *Frequency and Importance Rating Scales*

The second strongest relationship between pairs of individual rating scales was between Frequency and Importance rating scales. The range of correlations between the pairing of Frequency and Importance rating scales was .58 to .92, with an unweighted average correlation of .83. This is similar to the finding that Friedman (1990) reported, in which observed correlations between Frequency and Importance ratings ranged from .37 to .93, with an average unweighted correlation of .71. At the time, Friedman did not describe the correlation as “high”.

In this study, however, the observed unweighted average correlation was much higher than what Friedman had previously observed. Additionally, there were more sample studies included in this analysis (15 validation studies compared to 11 validation studies in Friedman’s research). The sample sizes in the 15 studies included in this analysis ranged from a low of 37 to a high of 3,185, whereas the range of sample studies in Friedman’s 11 studies ranged from a low of 3 to a high of 18. Presentation order was not varied in Friedman’s study, whereas seven of the 15 studies in this analysis were presented in “scale” order while the other eight were presented in “task” order. Finally, there was a fixed number of tasks rated in the study (244 tasks – the same task analysis

survey was used for all 11 studies), while the number of tasks rated in the 15 sample studies included in this analysis ranged from low of 18 to a high of 190.

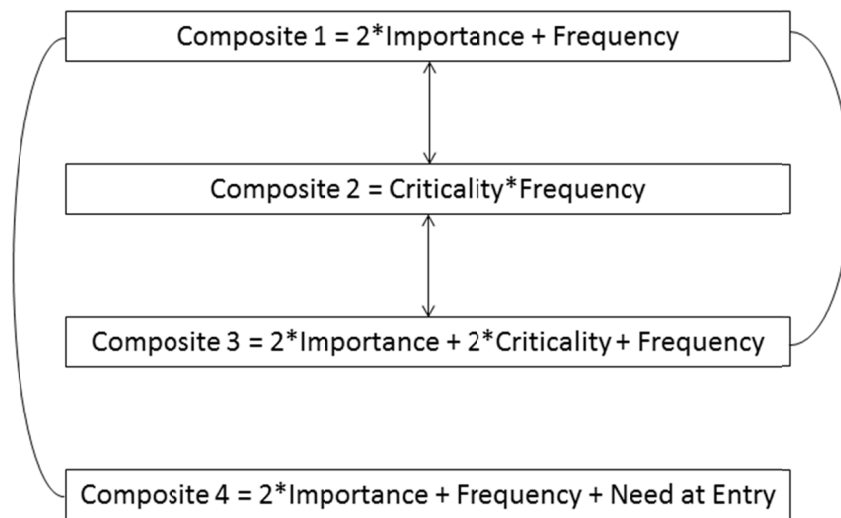
Due to the large amount of variation in the 15 studies included in the analysis between Frequency and Importance rating scales, and the small variation in the 11 studies included in Friedman's (1990) analysis of Frequency and Importance rating scales, it seems reasonable to conclude that there is a high degree of redundancy between the two rating scales and that including both rating scales on a survey validation study might not be the best use of survey respondents' time.

#### *Remaining Individual Rating Scales*

The weakest overall relationship, defined by the smallest unweighted average correlation ( $r=.63$ ), was between Frequency and Criticality ratings, indicating that the two scales are not highly related. Additionally, the relationship between both the Frequency and Importance rating scales with the Need at Entry scale was weak ( $r=.69$  and  $.67$  respectively), indicating that the Need at Entry rating scale might be assessing something different than task *frequency* or *importance*. The implications of which are that Frequency rating scales are evaluating something different than both Criticality and Need at Entry rating scales, and that Importance rating scales are evaluating something different than Need at Entry rating scales. Based on these findings, one might conclude that Need at Entry rating scales are truly assessing something different than the other scales evaluated in this study, and should be included on survey validation studies. Additionally, if offered a choice between including Frequency and Importance rating scales or Frequency and Criticality rating scales, it might be worth using Frequency and Criticality rating scales.

### *Composite Rating Scales*

Composite 1 was derived from Frequency and Importance rating scales; Composite 2 was derived from Criticality and Frequency rating scales; Composite 3 was derived from Importance, Criticality, and Frequency rating scales; and Composite 4 was derived from Frequency, Importance, and Need at Entry rating scales. The pairings obtained from this analysis were between Composite 1 and 2, Composite 1 and 3, Composite 1 and 4, and Composite 2 and 3, as illustrated in Figure 50. There were no pairings between Composite 2 and 4 or Composite 3 and 4, due to the fact that the Need at Entry scale and Criticality scale were never used on the same study.



*Figure 50.* Pairings of Composite scales used in data analysis.

As expected, there was a very strong relationship between the four observed pairings of composite rating scales. The average unweighted correlations ranged from a low of .97 (between Composite 2 and 3) to a high of .99 (between Composite 1 and 3). When considering the relationship between Composite 1 and 3, and Composite 1 and 4, both Composite 3 and 4 includes the same two scales from Composite 1 with the addition

of an extra rating scale. For example, Composite 1 is derived from two times Importance plus Frequency; while Composite 4 is two times Importance plus Frequency plus *Need at Entry*. The addition of the Need at Entry rating scale in the Composite didn't affect the relationship.

The same is true for the relationship between Composite 2 and 3. Composite 2 is derived from Frequency and Criticality rating scales; while Composite 3 includes Frequency, Criticality and *Importance* rating scales. These findings suggest the addition of a third rating scale, when added to the combination of Frequency and Importance or Frequency and Criticality rating scales, does not impact the relationship. The addition of either a Criticality rating scale or Need at Entry rating scales does not impact the magnitude of aggregate task ratings. For example, if tasks are performed frequently (receiving a high Frequency rating), very important (receiving a high Importance rating), they are also likely to be needed at entry into the profession (receiving a high Need at Entry rating).

Lastly, the relationship between Composite 1 and 2 is not surprising, as Composite 1 is derived from Frequency and Importance rating scales, while Composite 2 is derived from Frequency and Criticality ratings, and as previously mentioned, there is a very strong relationship between Importance and Criticality rating scales. This finding suggests that if Frequency, Importance, and Criticality rating scales are all used on the same survey, using a Composite that incorporates all three individual rating scales would produce largely similar results as a Composite that incorporates only Frequency and Importance or Frequency and Criticality.

### *Individual with Composite Rating Scales*

There were a total of 13 pairings between Individual and Composite rating scales. The range of unweighted, average correlations for pairs of Individual and Composite rating scales was .71 to .96. Ten of the pairings were between Individual rating scales and Composite rating scales in which the Individual rating scales were part of the Composite. Of those, the top three pairings (all with unweighted, average correlations of .96) were between Importance and Composite 1, Importance and Composite 3, and Frequency and Composite 2. This finding is reasonable since in all three cases the Individual scale was part of the Composite.

Three of the pairings were between Individual rating scales and Composites in which the individual scale was not included in the Composite. These three pairings were between the Need at Entry rating scale and Composite 1 (derived from Frequency and Importance scales); the Criticality rating scale and Composite 1; and the Importance rating scale and Composite 2 (derived from Frequency and Criticality). Those three pairings had lower unweighted average correlations than the other pairings between individual and composite scales, with the lowest unweighted average correlation between the Need at Entry rating scale and Composite 1 ( $r=.71$ ). As previously mentioned, the Need at Entry rating scale seems to be assessing something different than both the Frequency and Importance rating scales, so it is not surprising that the individual Need at Entry and Composite 1 rating scales had a relatively low correlation.

### **Summary of Potentially Moderating Variables**

In this study, the relationship between all pairings of rating scales was not statistically significantly affected by the four potential moderating variables – industry,

sample size, presentation order, or number of tasks. This finding implies that the redundancy (or lack thereof) between two rating scales would be observed regardless of the industry for which the job analysis was performed, the number of participants responding to the validation survey, the order in which scales or tasks are presented, or the number of tasks rated.

Although the four moderator variables did not *significantly* affect the relationships of rating scales, it is highly likely that some of the moderator variables do impact the relationship between rating scales. The sample sizes in this study were on the smaller side, which can affect power. For example, the industry for which the job analysis was performed might have had an impact on the relationship between scales, but due to small sample sizes, the effect of industry on the correlation between two rating scales may have been minimized. If the study were to be repeated with a larger sample size, statistical power may be boosted, and the effect may be more prominent.

Additionally, the correlations between rating scales was already very high to begin with, so assuming that a moderating variable would have a positive impact on the relationship between two rating scales, adding the moderating variable wouldn't significantly increase the correlation. Again, if this study were to be repeated with many more job analysis studies, we may find a greater range of correlations between scales, and we may be able to detect how those relationships are affected by any number of moderating variables.

### **Summary of Examination Blueprint Development Findings**

When considering the development of examination blueprints, the majority of psychometricians create examination blueprints based on the model presented earlier in

this text originally described by Raymond (1996), in which some combination of two or more scales is used to create examination blueprint weights. As such, there was little expectation that the examination blueprints derived from individual rating scales would resemble the examination blueprints derived from composite rating scales. However, this analysis was important because previous studies (Sanchez & Levine, 1989; Sanchez & Fraser, 1992) had postulated that one scale (in both cases, the overall Importance rating scale) would produce comparable results to a Composite scale.

To this end, both the relative rank order of the content areas on examination blueprints derived from the four individual scales and four composites, and the absolute difference between content areas on those derived examination blueprints, were compared to the actual examination blueprints. When looking at the relative rank order of content areas (would the greatest weighted content area on one examination blueprint be the greatest weighted content area on another examination blueprint), the derived examination blueprints that were most comparable to actual examination blueprints were examination blueprints derived from individual Frequency and Importance rating scales and Composites 1, 2 and 4. The correlations between the relative rank order of content areas in the examination blueprints derived from these four rating scales and the relative rank order of the content areas on the actual examination blueprints was between .97 and .99.

Examination blueprints derived from the Criticality rating scale and Composite 3 rating scale were the most dissimilar to the actual examination blueprints, with correlations of .93 and .94 respectively. This finding is somewhat counterintuitive, as the relationship between Importance and Criticality is high, one would expect that if an

examination blueprint based on the Importance rating scale alone was similar to actual examination blueprints, then an examination blueprint based on the Criticality rating scale alone would also be similar to actual examination blueprints.

However, correlations of .93 and .94 between the rank order of each duty area on all examination blueprints derived from the two lowest correlated scales and each duty area on all actual examination blueprints used on a licensure or certification is still quite high. This finding suggests there was a strong relationship between the relative rank orders of duty areas from *all* derived examination blueprints and the relative rank order of actual examination blueprints.

When considering the absolute differences between the percent of the exam devoted to each content area from derived examination blueprints versus the percent of the exam devoted to each content area from actual examination blueprints, examination blueprints derived from Composite 4 ratings were most similar to actual examination blueprints. On average, the absolute difference of the percent of the exam devoted to each content area on examination blueprints derived from Composite 4 compared to actual examination blueprints was 0.43%. This finding could be attributed to the fact that the actual examination blueprints for some of the studies in this analysis used Composite 4 (or something very similar to it) to derive those examination blueprints.

In fact, the percent of each content area on examination blueprints resulting from all four Composite ratings was similar to the actual examination blueprint in most cases. Again, this finding may be due to the fact that many psychometricians use some type of Composite scale to create examination blueprints, so any choice of Composite rating



scale is more likely to resemble the actual examination blueprint than any examination blueprint based on one Individual rating scale.

Lastly, the incorporation of the new examination blueprint in which all duty areas were equally weighted provided an additional level of analysis. The finding that as the number of duty (or content) areas increased, the absolute difference between actual and derived examination blueprints and equally weighted examination blueprints decreased seems intuitive. Nevertheless, it was a bit shocking to see that on average, examination blueprints derived from equally weighted duty areas differed from actual examination blueprints by less than 5%.

### **Implications for Practice**

What should we take away from this study? First, and most importantly, the choice of Composites used to create an examination blueprint does not seem to have an impact on the distribution of items on the final examination blueprint (with the exception of Composite 3). When developing a licensure or certification examination, the number of items devoted to each content area are most likely going to be the same (or very similar) regardless of the Composite rating scales used to derive the examination blueprint. This is due to the fact that very small changes in the percent of the exam devoted to each content area (around 1%), when multiplied by the number of items on an exam is only going to equate to a small difference between the number of items on a content area when the examination blueprint is derived from one Composite rating scale or another. For example, considering a 100-item exam, a 1% difference between content areas equates to a one item difference in each content area. To this end, the choice of Composites does not make a substantial difference in the weighting of examination

blueprints so psychometricians and job analysts should choose which Composite they feel most comfortable with and use the chosen Composite to create examination blueprints.

Second, as both Task Importance and Task Criticality are highly related, psychometricians should choose to use an Importance rating scale or a Criticality rating scale, but not both. Whether one decides to use an Importance rating scale or a Criticality rating scale might depend on the industry for which the job analysis is being performed. For example, if one were performing a job analysis for a dentist, choosing a Criticality or Consequence of Error rating scale may make more sense than choosing an Importance rating scale, as it may be easier for dentists to describe their job in terms of *critical* tasks rather than expressing the *importance* of tasks. Considering the task “Sterilize dental equipment”, asking a dentist to rate the severity of the consequences of not performing this task, or performing it incorrectly, may be easier than simply asking the dentist to rate its overall importance.

However, if one were performing a job analysis for a teacher, a Criticality or Consequence of Error rating scale may not be as good a fit as an Importance rating scale, as it might be much easier for teachers to think in terms of “How *important* is this activity for student success?” or “How *important* is this activity for achieving tenure?”. Again, a choice should be made by the job analyst or psychometrician on which of these two rating scales is a better fit.

An additional consideration in choosing between Criticality or Importance rating scales is the other scale(s) that are included along with the Criticality or Importance rating scale. Remember, examination blueprints derived from the Criticality rating scale

alone were dissimilar to the actual examination blueprints in this study. As such, one could argue for the use of an Importance rating scale over a Criticality rating scale, as examination blueprints seem to be more similar to task *importance* ratings.

Third, since the Need at Entry rating scale had relatively low correlations with the other individual rating scales, it seems reasonable to assume that the Need at Entry rating scale is assessing something different than the other rating scales. As such, organizations should consider including the Need at Entry rating scale when conducting survey validation studies for job analyses. This is especially true for organizations developing licensure exams, as licensure relates to minimal competence and any tasks that are obtained on the job, after years of working in a profession, may not be suitable for a licensure examination anyway.

### **Limitations and Implications for Future Research**

While this study contributed to the literature by confirming some of the findings from previous studies and weakening some findings from other studies, it by no means answered all of the questions related to the choice of scales that should be used on survey validation studies for job or task analyses for licensure or certification examinations. One of the limitations of this study was that it was a secondary data analysis, and as such, the variables in this study could not be manipulated. In the future, it would be beneficial to develop survey validation studies in which some of the variables of interest could be manipulated. For example, it would be valuable to create two versions of the same survey validation study in which one version presented the task ratings one task at a time and the other version presented the task ratings one scale at a time, and to randomly assign survey participants to one of the two versions. In this setting, one would be able to

better determine if presentation order had an impact on the relationships between rating scales as the presentation order variable wouldn't be fixed in one industry or with one respondent population.

Another limitation to this study is the relatively small number of studies that are included in the analysis. With only 20 studies, statistical power is not as high as we would like it to be, and thus the effects of moderator variables on the relationship between scales may not be as prevalent as one would like. As such, this study should be repeated with a larger number of sample studies. And the sample studies included in a future analysis could include a variety of additional moderator variables. For example, whether the job analysis was performed for a licensure or certification program may be an interesting moderating variable in a future study. Whether the job analysis and validation survey was performed for a startup credential or for an existing credential may be of interest. These additional moderator variables could be included in a follow-up analysis that included many more sample studies.

A third limitation to this study is the somewhat “unknown” quality of the surveys. Although there are generally accepted best practices for creating, disseminating, and analyzing survey validation studies for job analyses, it is unknown whether all of the organizations who conducted the studies included in this analysis followed those best practices. For example, one of the best practices associated with conducting survey validation studies, and survey research in general, is to pilot test the survey before administering it to a larger audience. This activity is performed to at least partially ensure that the interpretation of the rating scales is uniform across survey respondents. If this activity was not performed, and the rating scales were not interpreted as intended,

than one may be unsure of the survey results. While each of the organizations who contributed studies to this analysis stated that they followed best practices related to survey research, this could not be verified. However, in reviewing the standard errors of the mean (SEM) of each task rating on all 20 of the studies included in this analysis, one could argue that if these studies were repeated according to best practices, the results would be largely the same, as the SEMs for each task on each study were all relatively low.

A fourth limitation to this study and implication for future research is in the choice of rating scales used in this study. While the four individual scales analyzed in this study are the most common, there are some job analysts and psychometricians that use other individual rating scales (i.e., time-spent or difficulty of learning). It would be beneficial in the future if this study could be repeated with more studies that utilized a larger variety of task rating scales. Along those same lines, Composite 2 is the only one of the Composite rating scales that utilizes a multiplicative model versus an additive model. If a new Composite was created using an additive model with only the Frequency and Criticality ratings, it is possible that the new Composite would also resemble the other Composites.

## REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Andersson, B. & Nilsson, S. (1964). Studies in the reliability and validity of the critical incident technique. *Journal of Applied Psychology*, 48(6), 398-403.
- Blumrosen, A.W. (1972). Strangers in paradise: Griggs v. Duke power co. and the concept of employment discrimination. *Michigan Law Review*, 71(1), 59-110.
- Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (2009). *Introduction to meta-analysis*. Chichester, U.K.: John Wiley & Sons, Ltd.
- Brannick, M.T., Cadle, A., & Levine, E.L. (2012). Job analysis for KSAOs, predictor measures, and performance outcomes. In N. Schmitt (Ed.) *The Oxford handbook of assessment and selection*. Oxford University Press.
- Brannick, M. T., Levine, E. L., & Morgeson, F. P. (2007). *Job and work analysis: Methods, research and applications for human resource management*. Thousand Oaks, CA: Sage.
- Butterfield, L.D., Borgen, W. A., Amundson, N.E., & Maglio, A.T. (2005). Fifty ears of the critical incident technique: 1954-2004 and beyond. *Qualitative Research*, 5(4), 475-497.
- Christal, R. E., & Weissmuller, J. J. (1988). Job-task inventory analysis. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government: Volume II* (pp. 1036-1050). New York: John Wiley & Sons.
- Colton, A., Kane, M.T., Kingsbury, C., & Estes, C.A. (1991). A strategy for examining the validity of job analysis data. *Journal of Educational Measurement*, 28(4), 283-294.
- Corbally, J.E. (1956). The critical incident technique and educational research. *Educational Research Bulletin*, 35(3), 57-62.
- Council on Licensure, Enforcement and Regulation. (1992). *Principles of fairness: An examination guide for credentialing boards*. Retrieved from: <http://www.clearhq.org/Default.aspx?pageId=481186>

- Dillman, D. A., Smyth, J.D., Christian, L.M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method*. Hoboken, NJ: John Wiley and Sons, Inc.
- Fine, S.A. (1988). Functional job analysis. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government: Volume II* (pp. 1019-1035). New York: John Wiley & Sons.
- Fine, S. A. & Cronshaw, S. F. (1999). *Functional job analysis: A foundation for human resources management*. Mahwah, NJ: Erlbaum.
- Fine, S.A. & Getkate, M. (1995). *Benchmark Tasks for Job Analysis: A Guide for Functional Job Analysis Scales*. Mahwah, NJ: Erlbaum.
- Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin*, 51(4), 327-358.
- Fleishman, E.A., & Mumford, M.D. (1988). Ability requirement scales. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government: Volume II* (pp. 917-935). New York: John Wiley & Sons.
- Friedman, L. (1990). Degree of redundancy between time, importance, and frequency task ratings. *Journal of Applied Psychology*, 75, 748-752.
- Funke, F., Reips, U., & Thomas, R.K. (2011). Sliders for the smart: Type of rating scale on the web interacts with educational level. *Social Science Computer Review*, 29(2), 221-231.
- Gael, S. (1983). *Job analysis: A guide to assessing work activities*. San Francisco: Jossey-Bass.
- Gael, S. (1988). Subject matter expert conferences. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government: Volume I* (pp. 432-445). New York: John Wiley & Sons.
- Ghiselli, E.E., Campbell, J.P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: W.H. Freeman and Company.
- Goodwin, J.C. (2005). *Research in psychology: Methods and designs*. Hoboken: John Wiley & Sons, Inc.
- Griggs v. Duke Power Co. 401 U.S. 424 (1971). Retrieved from [http://www.law.cornell.edu/supct/html/historics/USSC\\_CR\\_0401\\_0424\\_ZO.html](http://www.law.cornell.edu/supct/html/historics/USSC_CR_0401_0424_ZO.html)
- Harvey, R.J. & Wilson, M.A. (2000). Yes Virginia, there is an objective reality in job analysis. *Journal of Organizational Behavior*, 21, 829-854.

- Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- International Organization for Standardization/International Electrotechnical Committee, Conformity Assessment Committee. (2003). *ISO/IEC 17024*.
- Joint Committee on Testing Practices, American Psychological Association (2004). *Code of fair testing practices in education*. Retrieved from <http://www.apa.org/science/jctpweb.html>
- Kane, M.T. (1982). The validity of licensure examinations. *American Psychologist*, 37, 911-918.
- Kane, M.T., Kingsbury, C., Colton, D., & Estes, C. (1989). Combining data on criticality and frequency in developing test plans for licensure and certification examinations. *Journal of Educational Measurement*, 26, 17-27.
- Kelman, M. (1991). Concepts of discrimination in “general ability” job testing. *Harvard Law Review*, 104(6), 1157-1247.
- Knapp, J., & Knapp, L. (1995). Practice analysis: Building the foundation for validity. In J.C. Impara (Ed.), *Licensure Testing: Purposes, procedures, and practices* (pp. 93-116).
- LaDuca, T. (2006). Commentary: A closer look at task analysis: Reactions to Wang, Schnipke, and Witt. *Educational Measurement: Issues and Practice*, 25(2), 31-33.
- Lipsey, M.W. & Wilson, D.B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Levine, E. L., Ash, R. A., Hall, H., & Sistrunk, F. (1983). Evaluation of job analysis methods by experienced job analysts. *Academy of Management Journal*, 26, 339-348.
- Manson, T.M., Levine, E.L., & Brannick, M.T. (2000). The construct validity of task inventory ratings: A multitrait-multimethod analysis. *Human Performance*, 13(1), 1-22.
- McCormick, E. J. (1976). Job and task analysis. In M. D. Dunnette (Ed.) *Handbook of industrial and organizational psychology* (pp. 651-697). Chicago: Rand McNally.
- National Commission for Certifying Agencies, Institute for Credentialing Excellence. (2004). *Standards for the accreditation of certification programs*. Retrieved from <http://www.credentialingexcellence.org/PublicationsandResources/Publications/Standards/tabid/390/Default.aspx>



- Nelson, D.S. (1994). Job analysis for licensure and certification exams: Science or politics?. *Educational Measurement: Issues and Practice*, 13(3), 29-35.
- Nelson, E.C., Jacobs, A.R., & Breer, P.E. (1975). The study of the validity of the task inventory method of job analysis. *Medical Care*, 13(2), 104-113.
- Newman, L.S., Slaughter, R.C., & Taranath, S.N. (1999, April). *The selection and use of rating scales in task surveys: A review of current job analysis practice*. Paper presented at the annual meeting of the National Council of Measurement in Education, Montreal, Canada.
- Norton, R.E. (2008). *DACUM handbook*. Columbus, OH: Ohio State University National Center for Research in Vocational Education.
- Owens, C.M. (2011). *Meta-analysis of single-case data: A Monte Carlo investigation of a three level model* (Doctoral dissertation). Retrieved from ProQuest database. (UMI No. 3449468).
- Pass, J.J., & Robertson, D.W. (1980). *Methods to evaluate scales and sample sizes for stable task inventory information*, San Diego, CA: Navy Personnel Research and Development Center.
- Primoff, E. S., & Eyde, L. D. (1988). Job element analysis. In S. Gael (Ed.) *The job analysis handbook for business, industry, and government* (Vol. II) pp. 807-824. New York: Wiley.
- Raymond, M.R. (1996). Establishing weights for test plans for licensure and certification examinations. *Applied Measurement in Education*, 9(3), 237-256.
- Raymond, M.R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education* 14(4), 369-415.
- Raymond, M.R. (2002). A practical guide to practice analysis for credentialing examinations. *Educational Measurement: Issues and Practice*, 21, 25-37.
- Raymond, M.R. (2005). An NCME instructional module on developing and administering practice analysis questionnaires. *Educational Measurement: Issues and Practice*, 24, 29-42. doi: 10.1111/j.1745-3992.2005.00009
- Raymond, M.R., & Neustel, S. (2006). Determining the content of credentialing examinations. In S.M. Downing, & T.M. Haladyna (Eds.), *Handbook of Test Development* (pp.181-223).
- \Ranyner, P. & Hermann, G. (1988). The relative effectiveness of tree occupational analysis models. *The Vocational Aspect of Education*, XL(106), 47-55.

- Ricci et al. v. DeStefano et al., 557 U.S. \_\_\_\_ (2009).
- Sanchez, J.I. & Fraser, S.L. (1992). On the choice of scales for task analysis. *Journal of Applied Psychology*, 77, 545-553.
- Sanchez, J. I., & Levine, E. L. (1989). Determining important tasks within jobs: A policy-capturing approach. *Journal of Applied Psychology*, 74, 336-342.
- Sanchez, J.I. & Levine, E.L. (2000). Accuracy or consequential validity: Which is the better standard for job analysis data?. *Journal of Organizational Behavior*, 21, 809-818.
- Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist*, 36(10), 1138-1146.
- Smith, I.L. & Hmbleton, R.K. (1990). Content validity studies of licensing examinations. *Educational Measurement: Issues and Practice*, 9(4), 7-10.
- Spray, J.A. & Huang, C. (2000). Obtaining test blueprint weights from job analysis surveys. *Journal of Educational Measurement*, 37, 187-201.
- Taber, T.D. & Peters, T.D. (1991). Assessing the completeness of a job analysis procedure. *Journal of Organizational Behavior*, 12(7), 581-593.
- Tannenbaum, R.J. & Wesley, S. (1993). Agreement between committee-based and field-based job analyses: A study in the context of licensure testing. *Journal of Applied Psychology*. 78(6), 975-980.
- Thompson, D.E. & Thompson, T.A. (1982). Court standards for job analysis in test validation. *Personnel Psychology*, 35, 65-874.
- Tourangeau, R., Couper, M.P., & Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, 68(3), 368-393.
- U.S. Equal Employment Opportunity Commission, EEOC Guidelines. (2010). *Employment tests and selection procedures*. Retrieved from [http://www.eeoc.gov/policy/docs/factemployment\\_procedures.html](http://www.eeoc.gov/policy/docs/factemployment_procedures.html)
- Van Cott, H.P., & Paramore, B. (1988). Task Analysis. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government: Volume I* (pp. 651-671). New York: John Wiley & Sons.
- Wang, N., Schnipke, D., & Witt, E.A. (2005). Use of knowledge, skill, and ability statements in developing licensure and certification examinations. *Educational Measurement: Issues and Practice*, 24(1), 15-22.

- Wang, N., Witt, E.A., & Schnipke, D. (2006). Rejoinder: A further discussion of job analysis and use of KSAs in developing licensure and certification examinations: A response to LaDuca. *Educational Measurement: Issues and Practice*, 25(2), 34-37.
- Willett, J. & Hermann, G. (1989). Which occupational analysis technique: Critical incident, DACUM, and/or information search?. *The Vocational Aspect of Education*, XLI(110), 79-88.
- Wilson, H.S. (1972). A second look at griggs v. duke power company: Ruminations on job testing, discrimination, and the role of the federal courts. *Virginia Law Review*, 58(5), 844-874.
- Wilson, M.A. (1997). The validity of task coverage ratings by incumbents and supervisors: Bad news. *Journal of Business and Psychology*, 12(1), 85-95.
- Wilson, M.A. & Harvey, R.J. (1990). The value of relative-time-spent ratings in task-oriented job analysis. *Journal of Business and Psychology*, 4(4), 453-461.

## APPENDICES

## Appendix A: Additional Detail Regarding Sample Studies

Table 48.  
*Coding of Sample Studies*

Studies	Industry	Sample Size	Percentage of Eliminated Respondents	N Tasks	Purpose of Study				Scales								Presentation Order		
					New Credential	Revalidation	Licensure	Certification	N Scales	Frequency	Importance	Criticality	Need at Entry	Comp1	Comp2	Comp3	Comp4	Scale	Task
1	Education	1639	26%	47	x			x	3	x	x		x	x			x	x	
2	Healthcare	195	32%	51	x			x	2	x	x			x					x
3	Healthcare	512	13%	33	x			x	2	x	x			x					x
4	Healthcare	400		34	x			x	3	x	x	x		x	x	x		x	
5	Healthcare	67	4%	37		x		x	2	x	x			x					x
6	Prof, Sci, Tech	116	42%	30	x			x	3	x	x	x		x	x	x		x	
7	Prof, Sci, Tech	3185	47%	190		x	x		3	x	x		x	x			x		x
8	Construction	149	17%	59	x			x	3	x	x	x		x	x	x		x	
9	Food	716	24%	32		x		x	3	x	x	x		x	x	x		x	
10	Prof, Sci, Tech	3043	37%	87		x		x	2	x		x			x				x
11	Construction	65	31%	60		x	x		2	x	x			x					x
12	Construction	37	16%	50		x	x		2	x	x			x					x
13	Utilities	481	11%	36	x			x	2	x	x			x					x
14	Utilities	186	12%	18	x			x	2	x	x			x					x

Table 48.  
Coding of Sample Studies

Studies	Industry	Sample Size	Percentage of Eliminated Respondents	N Tasks	Purpose of Study				Scales								Presentation Order			
					New Credential	Revalidation	Licensure	Certification	N Scales	Frequency	Importance	Criticality	Need at Entry	Comp1	Comp2	Comp3	Comp4	Scale	Task	
15	Information	204	21%	19	x				2	x	x			x					x	
16	Healthcare	1798	13%	123		x	x		3	x	x		x	x			x			x
17	Utilities	1033	57%	305		x	x		2	x		x			x					x
18	Utilities	621	51%	222		x	x		2	x		x			x					x
19	Utilities	1226	62%	331		x	x		2	x		x			x					x
20	Utilities	212	48%	180		x	x		2	x		x			x					x
Totals		794*	30%*	97*	9	11	3	17		20	15	9	3	15	9	4	3	6		14

\*Denotes average rather than total.

The sample studies presented in Table 47 are representative of the kinds of studies one would see if this analysis were to be repeated. There were several studies from the healthcare industry, which is not surprising as there are countless certifications in the healthcare industry. There were also a lot of studies from the construction and/or utilities industries, which is not uncommon as there are many licenses and several certifications related to the construction and utilities industries. There were was only one study from the education industry, which again, is not surprising. There are fewer certifications related to the education industry than there are in other industries. Additionally, there were some industries listed on the O\*Net list of industries that were not represented at all in this analysis. For example, there were no sample studies from the “government” industry. This is due to the fact that there are few, if any, “government” based licenses or certifications.

The sample sizes of these studies ranged from a low of 37 to a high of 3,185. This finding would be expected if this analysis was to be repeated. The number of respondents to any survey validation study for a job analysis is dependent on so many factors. For example, is the job analysis being developed for a new credential, in which the “job” doesn’t exist? If so, the sample sizes may be much larger, because a wide net would have to be cast to get anyone who could potentially desire to obtain the future credential. Is the credential national or state-specific? Obviously we would expect to see a very different sample size for a credential whose target audience is anyone living in North America compared to a credential for individuals working within one county.

The percentage of eliminated respondents is based upon the number of respondents who responded to less than 75% of the survey. Anyone who completed less

than 75% of the survey was removed from the final analysis. In some cases, that was a small amount of individuals (i.e., in the fifth study, 4% of respondents were eliminated) and in other cases that was a large amount of individuals (in the seventh study, 47% of respondents were eliminated). As indicated in Table 48 in Appendix B, the greater the number of tasks to rate, the greater the number of survey respondents that were eliminated. This finding suggests what we had already assumed, the longer the survey, the greater the attrition rate.

The number of tasks on the initial job analysis ranged from a low of 18 to a high of 331. While the range may seem uncharacteristically large, this job analyst does not believe that this finding is that unusual. Job analysts tend to fall into two categories, “lumpers” and “splitters”. Lumpers tend to lump tasks together. They may argue that if several tasks all have the same underlying KSAs, there is no reason to split them apart. Lumpers may also argue that if two tasks are similar, even if they have different KSAs, they could be justifiably lumped together. Splitters, on the other hand, tend to split tasks apart. Splitters argue that for someone reviewing the job analysis in the future, the duties, tasks, and corresponding KSAs will make infinitely more sense if they are segregated. Splitters argue that more detail is better. As such, a “lumper” and a “splitter” may end up with completely different numbers of task statements for the exact same job, hence the wide range of the number of tasks observed on the 20 sample validation studies.

In terms of the purpose of the study, there was almost a 50/50 split between survey validation studies for new credentials versus revalidations of existing credentials. This is not surprising, as ISO 17024 states that job analysis for credentialing exams shall be revalidated a minimum of every five years (ISO/IEC, 2003) and new credentials are



being developed daily for jobs that currently exist as well as new professions. If this study were to be repeated, it is possible that there would be more survey validation studies for existing credentials, as the need for revalidations will continue to increase as more and more organizations develop credentialing exams.

The majority of the studies included in this analysis were for Certification exams (17 studies) rather than licensure exams (three studies). This finding is not surprising as licensure exams tend to be regulated by some government agency (i.e., a state department) or regulatory authority, both of which tend to do their exam development work in-house. Additionally, these organizations are less likely to share their exam development data (job analysis and survey validation data) with a psychometrician doing research.

As previously mentioned in the body of this paper, the choice of scales used in the 20 sample studies is common. Frequency and Importance/Criticality are the two most common rating scales used in survey validation studies. Some job analysts prefer Importance over Criticality, others prefer Criticality over Importance. Most use some sort of Frequency rating scale in their survey validation work.

Lastly, in terms of presentation order, there were most studies that presented the rating scales in task order (asking survey respondents to use all scales to rate one task at a time) rather than scale order (rating all tasks on one scale before moving onto the next scale). Although survey methodology research would advise against presenting rating scales in task order, this seems to be the norm in survey validation work. When surveys are presented in task order, they appear shorter than when they are presented in scale order. As such, many organizations prefer to have the appearance of shorter surveys.

## Appendix B: Correlation of Survey Variables

Table 49.  
*Correlations of Sample Study Variables*

		Industry	Sample Size	Eliminated	N Tasks	New/ Reval <sup>1</sup>	Lic/ Cert <sup>2</sup>	N Scales	Freq	Imp	Crit	Need	Comp 1	Comp 2	Comp 3	Comp 4	Present <sup>4</sup>
Industry	Corr	1															
	Sig.																
SampleSize	N	20															
	Corr	-.249	1														
SampleSize	Sig.	.289															
	N	20	20														
Eliminated	Corr	.277	.309	1													
	Sig.	.236	.184														
Eliminated	N	20	20	20													
	Corr	.372	.363	.791**	1												
NTasks	Sig.	.106	.115	.000													
	N	20	20	20	20												
NewReval	Corr	.166	.356	.361	.587**	1											
	Sig.	.484	.124	.118	.007												
NewReval	N	20	20	20	20	20											
	Corr	.077	-.137	-.033	-.013	-.380	1										
LicCert	Sig.	.747	.564	.892	.958	.098											
	N	20	20	20	20	20	20										
NScales	Corr	-.462*	.278	-.028	-.185	-.179	.015	1									
	Sig.	.040	.236	.908	.435	.450	.951										

Table 49.  
*Correlations of Sample Study Variables*

	Industry	Sample Size	Eliminated	N Tasks	New/ Reval <sup>1</sup>	Lic/ Cert <sup>2</sup>	N Scales	Freq	Imp	Crit	Need	Comp 1	Comp 2	Comp 3	Comp 4	Present <sup>4</sup>
Freq	N	20	20	20	20	20	20									
	Corr	.b	.b	.b	.b	.b	.b	.b								
	Sig.															
Imp	N	20	20	20	20	20	20	20								
	Corr	.b	.b	.b	.b	.b	.b	.b	.b							
	Sig.															
Crit	N	15	15	15	15	15	15	15	15							
	Corr	.b	.b	.b	.b	.b	.b	.b	.b	.b						
	Sig.															
Need	N	9	9	9	9	9	9	9	4	9						
	Corr	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b					
	Sig.															
Comp1	N	3	3	3	3	3	3	3	3	0	3					
	Corr	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b				
	Sig.															
Comp2	N	15	15	15	15	15	15	15	15	4	3	15				
	Corr	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b			
	Sig.															
Comp3	N	9	9	9	9	9	9	9	4	9	0	4	9			
	Corr	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b		
	Sig.															
Comp4	N	4	4	4	4	4	4	4	4	4	0	4	4	4		
	Corr	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b	.b	
	Sig.															
N	3	3	3	3	3	3	3	3	3	0	3	3	0	0	3	

Table 49.

*Correlations of Sample Study Variables*

	Industry	Sample Size	Eliminated	N Tasks	New/ Reval <sup>1</sup>	Lic/ Cert <sup>2</sup>	N Scales	Freq	Imp	Crit	Need	Comp 1	Comp 2	Comp 3	Comp 4	Present <sup>4</sup>
Presentation	Corr	.120	.182	.088	.421	.504*	-.663**	.b	.b	.b	.b	.b	.b	.b	.b	1
	Sig.	.614	.442	.714	.064	.023	.241	.275								
	N	20	20	20	20	20	20	20	20	15	9	3	15	9	4	3

\*. Correlation is significant at the 0.05 level (2-tailed).

\*\* . Correlation is significant at the 0.01 level (2-tailed).

b. Cannot be computed because at least one of the variables is constant.

1. Dummy coded: 0=New, 1=Revalidation

2. Dummy coded: 0=Certification, 1=Licensure

3. Dummy coded: 0=By Task, 1=By Scale

## ABOUT THE AUTHOR

Adrienne Woodley Cadle has been working in and around licensure and certification testing for as long as she can remember. Her parents work in licensure and certification testing, so as a teenager she would sharpen pencils and seal examination booklets after school for upcoming test administrations. At the age of 19 she took a job at Professional Testing, Inc., and quickly found herself performing a number of activities related to testing. These activities included printing and shipping exams, administering paper-and-pencil and computer-based exams, scoring exams, and assisting in the facilitation of item development workshops. She enjoyed the industry so much that she made a decision to complete her education and become a psychometrician.

In December 2005, Adrienne graduated from the University of South Florida with a B.A. in Psychology. The following January, Adrienne began taking classes as a graduate student in the Educational Measurement and Research department at the University of South Florida. She completed her M.Ed. in the summer of 2006. During her time at the University, she worked as a student and then graduate assistant at the Institute for Instructional Research and Practice (IIRP). While at IIRP, she learned how to score and equate exams using SAS and administer both paper-and-pencil and CBT exams to ADA candidates.

In August 2006, Adrienne began work on her doctoral degree in the Educational Measurement and Research department at the University of South Florida. While taking classes full-time, she worked part-time as a graduate assistant in the Educational

Measurement and Research department and part-time as a psychometrician at Professional Testing. As a graduate assistant, Adrienne taught an undergraduate tests and measurement class for future teachers, co-taught graduate statistics, and worked on a number of research projects. As a psychometrician, she spends a great deal of her time facilitating job analysis meetings, item development and review meeting, and passing score study meetings. While she continues to work in exam development, she has found a passion for job analysis and the survey validation work association with job analyses. She has been certified as a DACUM Job Analysis Facilitator by Ohio State University, and has had the great privilege of co-authoring a chapter in *The Oxford Handbook of Personnel Assessment and Selection* titled “Job Analysis for KSAOs, Predictor Measures, and Performance Outcomes”. She continues her work in job analysis today, and hopes to publish more on the topic in the coming years.