2019

Extent, Correlates, and Consequences of Careless and Inattentive Responding in Certification Job Analysis Surveys

Patricia M. Muenzen
Walden University

Walden University

College of Social and Behavioral Sciences

This is to certify that the doctoral dissertation by

Patricia Muenzen

has been found to be complete and satisfactory in all respects,
and that any and all revisions required by
the review committee have been made.

Review Committee
Dr. James Herndon, Committee Chairperson, Psychology Faculty
Dr. Derek Rohde, Committee Member, Psychology Faculty
Dr. Brian Cesario, University Reviewer, Psychology Faculty

Chief Academic Officer
Eric Riedel, Ph.D.

Walden University
2019

Abstract

Extent, Correlates, and Consequences of Careless and Inattentive Responding in Certification Job Analysis Surveys

by

Patricia M. Muenzen

MA, Stony Brook University, 1988

Sc.B., Brown University, 1983

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Industrial/Organizational Psychology

Walden University

May 2019

Abstract

Survey data quality is influenced by the care and attention that respondents take in answering questions. Careless and inattentive (CI) responding is a confound in survey data that can distort findings and lead to incorrect conclusions. This quantitative study explored CI responding in job analysis studies supporting occupational certification programs and its relationship to survey features, data quality measures, and test content validity. Satisficing theory served as the framework, and secondary analysis of 3 job analysis surveys was undertaken. Results indicated that 9-33% of respondents engaged in CI responding, with the rate differing by CI index used (Mahalanobis distance, long string analysis, or person-total correlation) and by occupation. Each index detected a distinct pattern of carelessness, supporting the use of multiple indices. The indices performed best detecting carelessness in frequency ratings and may not be useful for all job analysis rating scales. Partial support was found for relationships between carelessness and survey features. CI responding had a minimal impact on mean ratings, correlations, and interrater reliability, and had no impact on certification test content outlines. By providing guidance and caution on the use of CI response detection methods with job analysis survey data, this study produced two potential avenues for social change. For practitioners conducting occupational job analyses, the use of CI detection methods can enhance the validity of data used to make certification decisions. For researchers, follow-up studies can yield a more nuanced understanding of the most appropriate use of these methods in the job analysis context.

Extent, Correlates, and Consequences of Careless and Inattentive Responding in Certification Job Analysis Surveys

by

Patricia M. Muenzen

MA, Stony Brook University, 1988

Sc.B., Brown University, 1983

Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Industrial/Organizational Psychology

Walden University

May 2019

Acknowledgments

I could not have done this alone. My family and friends supported me in so many ways during this journey and I wish to thank them all for their contributions to my success. First, my amazing family: Chuck, Kelly, Steven, Kathleen, and Elizabeth; Chase, Carla, Barry, Jessica, Shana, and Jesse; Mollie and Kaarlo; and Faith. I have an abundance of blessings with you in my life. You've all helped me persevere whether you know it or not. I love you all dearly. I wish my Mom and Dad could have been here for this. They would have been so proud. Thanks go to my treasured friends Shelley, Pattie, Janet, Shawna, Staci, Jeanne, Susan, Claire, and Daryl. They all believed in me and gave me the strength and confidence to push through the tough times.

My dissertation chair, Dr. James Herndon, generously shared his wisdom through this dissertation process. When I began this journey, I had no idea where it would take me. Dr. Herndon knew I would get where I needed to go if I just persevered. My second committee member, Dr. Derek Rohde, was tremendously supportive and encouraging. I thank them both for their invaluable assistance. Thanks also to URR Dr. Brian Cesario for providing an outsider view of my work.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction to the Study

Surveys are a widely used method to collect data in social science research (Breitsohl & Steidelmüller, 2018; de Vaus, 2013; Fulton, 2016). In addition to their widespread use, data collected through surveys play an important role in drawing inferences on social issues (Barge & Gehlbach, 2012; Dillman, n.d.; Thomas, 2014). Given the role that surveys play in investigating social issues, the need for high quality data is paramount. Researchers rely on survey takers to provide honest and accurate responses, yet evidence suggests this does not always occur (Krosnick, 1999). For example, respondents may engage in socially desirable responding by selecting response options they believe will convey a favorable impression of themselves, or deliberately falsify their answers (Krosnick & Presser, 2010). With the proliferation of Internet-delivered surveys, survey takers are less motivated and may be careless or inattentive (CI) when answering survey questions, drawing into question the accuracy of the information provided (Godinho, Kushnir, & Cunningham, 2016; Ward & Pond, 2015).

This study addresses survey data quality through the investigation of CI responding to job analysis surveys supporting national occupational certification programs. Occupational certification is the means by which independent organizations evaluate and award credentials to individuals demonstrating the requisite background, knowledge, skills, and abilities (Sireci & Hambleton, 2009). Typically, candidates for certification must meet eligibility requirements and pass a knowledge-based examination

(Raymond & Luecht, 2013). Examination specifications are structured around domains of practice, tasks performed, and optionally, knowledge used in practice.

The number of certifications being offered and the number of individuals seeking certification are both growing (Albert, 2017). The increasing popularity of certification may be attributed to several factors. Certification may provide a competitive advantage in the workplace as possession of the credential indicates a certain standard has been achieved through meeting education, experience, and examination requirements. For employers, certifications demonstrate that an applicant or incumbent possesses specific knowledge, skills, and experiences needed for specific positions. For those seeking to gain employment or make career changes, certification can enhance mobility by providing credentials that give entrée to new opportunities. Finally, certification may help address the skills gap by providing a means for training and verification of specific knowledge, skills, and abilities (Kochan, Finegold, & Osterman, n.d.; National Workforce Solutions Advisory Board, 2017).

The Bureau of Labor Statistics (2017) estimated that in 2017, 24.3% of the adult US population held occupational credentials, either licenses or certifications. Licenses are granted by US states to permit individuals to hold a title and practice in a profession. In contrast to licenses which are requirements for practice, certifications are voluntary, although some states may adopt certifications for licensure purposes. The Uniform Guidelines for Employee Selection Procedures (Equal Employment Opportunity Commission, 1978) mandate the use of job analysis in employee selection to establish a

link between assessment content and evidence of job-relatedness. The certification industry adheres to this guidance when developing examinations for voluntary certification programs. The credentialing industry also adheres to relevant standards such as the Standards for Educational and Psychological Testing (American Psychological Association [APA], American Educational Research Association [AERA], & National Council on Measurement in Education [NCME], 2014). The Standards for Educational and Psychological Testing specify that the content domain for a certification examination needs to be clearly defined, and that job analysis is an important method of defining the content domain. Accreditation standards for certification and licensure (Institute for Credentialing Excellence [ICE], 2014) require the conduct of occupational job analysis surveys in establishing a certification examination content outline, except in extenuating circumstances. Taken together, the legal environment, professional standards, and accreditation requirements all indicate the importance of job analysis as the means of establishing the content validity of certification examinations. Job analysis surveys conducted to support certification examinations play a key role in supporting the validation argument for certifications, and the data collected in credentialing program job analysis surveys impact testing content for examinations taken by nearly a quarter of the US adult population. Identification and removal of CI survey responses can potentially increase the reliability of job analysis survey data and better support the content validity argument for test content outlines drawn from the data.

In this chapter, I introduce the phenomenon of CI survey responding and discuss its relationship to data quality. I review research on CI responding both in general and in a job analysis context. Next, I describe a research program focused on job analysis studies for three occupations (literacy coach, patient care technician, and pharmacy technician) undertaken to support national certification programs for the respective occupations. Literacy coaches are consultants who provide professional development to teachers to improve teaching practices and student achievement in literacy. Pharmacy technicians work primarily in community and hospital pharmacies under pharmacists' direct supervision to fill prescriptions and support pharmacy operations. Patient care technicians work primarily under the supervision of nurses to provide basic care to patients, such as meals, toileting, and vital signs measurement. I describe researchable hypotheses, and outline steps to be undertaken to investigate the hypotheses. Assumptions, delimitations, and limitations of the study are also described.

Background

Beginning with an influential paper by Meade and Craig (2012) which reported an estimated 7 to 9% of survey data as careless, a growing body of research has focused on exploring CI survey responding (Godinho et al., 2016; Huang, Liu, & Bowling, 2015; Meade & Craig, 2012; Morgeson, Spitzmuller, Garza, & Campion, 2014; Steedle, 2018; Thomas, 2014; Ward & Pond, 2015). CI responding is differentiated from socially desirable responding, in which respondents select answers that create a favorable impression, or faking, in which respondents deliberately selecting false answers.

Response bias due to faking or socially desirable responding occurs when a respondent provides answers that are deliberate distortions based on the meaning of survey questions. In contrast, CI responses are unrelated to the questions posed. Response bias is related to the construct being measured, while CI responding is not (Meade and Craig, 2012). More recent estimates of CI responding are in the 8 to 2% range (Curran, 2016). The presence of CI responses in survey datasets has been found to decrease statistical power (Maniaci & Rogge, 2014) and reliability (Huang, Curran, Keeney, Poposki, & DeShon, 2012; Meade & Craig, 2012). It can attenuate or amply correlations depending on the characteristics of the valid data. Because of this, Huang et al. (2012) called CI responding an insidious confound. It distorts factor analysis structures (Huang et al., 2012), and decreases factor analysis model-data fit (Steedle, 2018) leading to different conclusions being drawn about the relationships among variables.

The Internet is commonly used to deliver occupational job analysis surveys to geographically dispersed national samples. CI responding occurs more frequently with Internet surveys than with their paper and pencil counterparts (Barge & Gehlbach, 2012; Thomas, 2014; Ward & Pond, 2015). Johnson (2005) found that the extent of inattentive responding to a personality inventory was greater when the inventory was administered online than when it was administered via paper and pencil. Internet-based surveys allow for distraction and multitasking (Carrier, Cheever, Rosen, Benitez, & Chang, 2009; Fang, Wen, & Prybutok, 2014; Hardré, Crowson, & Xie, 2012; Ward & Pond, 2015), making it easier to respond carelessly and inattentively. While there is no published data regarding

the relationship between Internet delivery and CI responding to occupational analysis surveys, extrapolation from the existing literature suggests it occurs and is an important issue.

Careless responding may be more prevalent in low stakes contexts, such as job analysis surveys that support human resource (HR) activities (Huang et al., 2012). Such surveys are the primary means for establishing content validity for national certification examinations (Raymond, 2001, 2002; Raymond & Luecht, 2013). Huang et al. (2012) argued that CI response detection is a quality control method useful to detect unmotivated responding to job analysis surveys. However, limited studies to date has focused on CI responding in job analysis surveys. Studies by Green & Stutzman (1986), Morgeson et al. (2014), and Stetz, Button, and Quist (2012) incorporated bogus items such as, *I was born on February 30*, into job analysis surveys to flag for respondents endorsing these items. Carelessness was operationalized as the number of bogus items endorsed. No cutoffs were applied to categorize respondents as careful or careless, so the rate of CI responding in job analysis surveys is not known.

Data from certification organizations is almost entirely lacking. Most certification organizations do not make their job analysis studies public. Of those that do, the majority do not mention data screening or data cleaning at all. I was able to identify only one organization, the American Registry of Radiologic Technologists (ARRT), that describes reviewing their collected job analysis data for aberrant responses. The ARRT used visual screening to identify unrealistic response patterns rather than mathematical

and statistical techniques. In addition, in 2017, I conducted a job analysis study of patient care technicians using bogus items and directed response items. Directed response items instruct respondents to select specific response options. This was the first time I used these techniques myself, and I believe my colleagues in the credentialing industry do not make these screening techniques part of their standard practice.

Outside of the job analysis context, several researchers have begun to study the utility of various post hoc statistical means of identifying and eliminating CI responses from datasets (Huang et al., 2012; Kam & Meyer, 2015; Maniaci & Rogge, 2014; Meade & Craig, 2012; Roivainen, Veijola, & Miettunen, 2016; Steedle, 2018; Ward & Meade, 2018; Ward & Pond, 2015). Numerous methods have been explored; however, no consistent pattern has emerged in the literature regarding which techniques to employ. DeSimone, Harms, and DeSimone (2015) and Huang et al. (2012) argued that a combination of techniques should be used, while Mancini and Rogge (2014) suggested that only a single detection technique can be sufficient. Meade and Craig (2012) identified two different types of CI responding, responding using a single response option and responding using a variety of response options, each identified by different carelessness indices. Curran (2016) recommended a multiple hurdles approach to data cleaning through the sequential use of different indices. In my study, the techniques and findings from these studies were applied to the certification job analysis context.

CI responding to self-administered surveys is not a new problem but it has received increased focus in the literature. A recent call for papers (Bowling & Huang,

2016) for a special issue of a measurement journal on the topic of "measurement, causes, and consequences" of CI responding highlights the relevance of the topic. The potential threat to reliability and validity posed by CI responding is stimulating progress toward better understanding of the phenomenon and efforts mitigation of the problem. The problem of careless responding is particularly acute for longer (Huang, Bowling, Liu, & Li, 2015) and Internet-based surveys (Hardré et al., 2012; Zhang & Conrad, 2014), and long, Internet-based surveys are characteristic of most occupational certification job analysis surveys. As "the topic of rating accuracy is a central yet unresolved issue in the job analysis literature" (Aguinis, Mazurkiewicz, & Heggestad, 2009, p. 433), it is essential to understanding the extent and impact of careless responding in this survey type.

Problem Statement

Limited research exists regarding CI responding in job analysis surveys, yet this type of responding is likely prevalent in job analysis (Huang et al., 2012). Careless responding exists and can distort survey results. Given the foundational role of job analysis data in determining examination content for occupational credentials, careless responding in occupational job analysis surveys is an important concern. Due to carelessness, findings from credentialing job analysis surveys may yield inaccurate certification test content outlines. If credentialing organizations award certifications to individuals who pass examinations based on content outlines of questionable validity, the meaning and value of credentials is compromised.

Purpose of the Study

The purpose of this quantitative study is to explore CI responding in self-administered Internet-based job analysis surveys that support occupational certification programs. A recommenced practice in certification program development is for sponsoring organizations to conduct two-phase job analysis studies. In the first phase, subject matter expert panels delineate the key attributes of the profession in terms of domains of practice, task performed within domains, and/or knowledge, skills and abilities employed. In the second phase, surveys are conducted to gather evidence to validate the elements of the delineation and create test content outlines (ICE, 2014; Raymond & Luecht, 2013). Credential sponsors administer Internet-based job analysis surveys to large samples of job holders across organizations and jurisdictions. The self-administered survey is a practical mode of data collection when large numbers of people perform the role or function under study (Van De Voort & Whelan, 2012), and the Internet provides a convenient way to deliver surveys to large samples of job holders who may be dispersed geographically.

This study draws upon two streams of prior research. The first is literature on job analysis quality. To date, studies regarding the quality of job analysis ratings have explored manipulations to encourage respondents to provide accurate responses (Lievens & Sanchez, 2007; Morgeson et al., 2014), as well as a limited set of techniques, typically the use of bogus items, to detect and eliminate poor quality responses (Green & Stutzman, 1986; Green & Veres, 1990; Stetz et al., 2012). None of the studies in the job

analysis arena have explored the potential use and utility of a broader set CI detection techniques. The second is the body of literature on the use of post hoc detection techniques to investigate careless responding. Because the extent of CI responding varies widely across studies (Curran, 2016), it is difficult to generalize findings across survey types. Thus, the extent of careless responding in job analysis surveys is not known. This study is the first to specifically explore post hoc detection of CI responding in job analysis surveys generally, and in national occupational job analysis studies for credentialing programs specifically.

In this study, the existing research on carelessness in job analysis surveys responding was extended to a larger set of detection methods than had been studied previously. The goals were to: (a) estimate the baseline rate of CI responding in job analysis questionnaires, (b) identify the types and extent of CI responding demonstrated with different survey questions and rating scales, (c) examine the psychometric implications of careless responses on job analysis results, and (d) explore optimal sets of CI response detection indices to use with job analysis survey data. Predictor variables were factors hypothesized to increase ratings complexity and decrease respondent motivation. The three predictor variables were survey length, job aspect rated (task versus knowledge), and type of rating scale (concrete versus abstract).

Regarding survey length, job analysts have long pointed out the tedious nature of lengthy job analysis surveys (Harvey & Wilson, 2000; Morgeson & Campion, 1997; Morgeson et al., 2014; Sanchez & Levine, 2001). In a study of job analysis ratings for

surveys supporting certification examinations, Wang, Wiser, and Newman (1999) found that for blocks of tasks appearing later in surveys, respondents used fewer response categories, a finding they attributed to fatigue. In studies of surveys other than job analysis, longer surveys were associated with poorer data quality (Galesic & Bosnjak, 2009; Hardré et al., 2012; Zhang & Conrad, 2014). These findings suggest the potential more CI responding in the longer surveys in my study.

Regarding job analysis elements and rating scales, Harvey and Wilson (2000) made a distinction between two continua relevant to job analysis collection data: the descriptor item metric and the rating scale metric. The descriptor item metric refers to the level of specificity or abstraction in job aspects rated. The most concrete job aspects are specific, observable, and verifiable job tasks. More abstract job aspects are knowledge, skills, abilities, and other attributes (KSAOs), with abilities and attributes being the most abstract. Morgeson and Campion (2000) made a similar distinction between ratings made for work activities/tasks and KSAOs, the "psychological constructs underlying job-related capabilities" (p. 823). Making KSAO ratings calls for an inferential leap from job tasks to judgments of the requisite KSAOs employed in task performance (Morgeson & Campion, 1997, 2017; Sanchez & Levine, 2001). Dierdorff and Morgeson (2009) found that task ratings had higher reliability than KSAO ratings and posited that making KSAO ratings was more cognitively demanding than rating tasks, leading to more idiosyncratic rater variance.

Harvey and Wilson's (2000) rating scale metric refers to the types of judgment a survey taker is required to make through the application of rating scales to tasks and KSAOs. A "behaviorally specific and easily verifiable" scale (Harvey & Wilson, 2000, p. 831), for example, an absolute frequency scale with response options such as annually, monthly, weekly, and daily is relatively straightforward, as it requires answers based in real time. Other job analysis scales are less concrete. For instance, a relativistic scale asking respondents to rate each task in relation to all other job tasks performed is complex, requiring the respondent to hold all job tasks in their minds simultaneously to make a judgment. Similarly, making yes/no task ratings for a hypothetical situation such as, *I would be expected to perform this task should the need arise*, is complex. In support of the distinction between abstract and concrete rating scales, Stetz et al. (2012) found that survey respondents endorsed bogus tasks more when rating whether they would be expected to perform the tasks if necessary than when rating the absolute frequency with which they performed the tasks. The implication is that having to determine whether they might be expected to perform a task was a more cognitively demanding and produced more careless responses.

The criterion variables in this study were three post hoc indices of CI responding. Long string analysis is an index indicting the number of consecutive identical responses in a response string (Meade & Craig, 2012). The Mahalanobis distance index compares each respondent's patterns of ratings to those of all other respondents. It is an indication of the extent to which an individuals' response patterns are inconsistent with those of all

other respondents (De Maesschalck, Jouan-Rimbaud, & Massart, 2000; Mahalanobis, 1936). The person-total correlation is the correlation between an individual's survey responses and the mean responses of all other survey takers, averaged across items (Curran, 2016). Long string analysis is a within-person measure, while Mahalanobis distance and person-total correlation are between-person measures. Additional criterion measures focused on job analysis quality were rating scale reliability, average interitem correlations, and mean ratings. Findings from Huang, Liu, and Bowling (2015), Maniaci and Rogge (2014), and Meade and Craig (2012) suggest values on these criterion measures will be attenuated for careless responders.

Research Questions and Hypotheses

RQ1: Is there a relationship between survey length and extent of carelessness?

$H_0 1$: There is no relationship between survey length and extent of carelessness

$H_a 1$: There is a relationship between survey length and extent of carelessness.

Both variables are continuous and Pearson's correlation coefficients were planned to address this question.

RQ2: Is there a relationship between job aspect rated and extent of carelessness?

$H_0 2$: There is no relationship between job aspect rated and extent of carelessness.

$H_a 2$: There is a relationship between job aspect rated and extent of carelessness.

Job aspect rated refers to which aspect of the job was surveyed, tasks performed or knowledge bases used. Tasks are descriptors of the job and knowledge bases are characteristics of job holders. Each survey was versioned, so respondents were randomly

routed to either tasks or knowledge. Job aspect (tasks versus knowledge), is a categorical variable and point-biserial correlation coefficients were planned to address this question.

RQ3: Is there a relationship between rating scale used and extent of carelessness?

$H_0 3$: There is no relationship between rating scale used and extent of carelessness.

$H_a 3$: There is a relationship between rating scale used and extent of carelessness.

For each survey version, two rating scales were used. For tasks, the scales were frequency and importance. For knowledge, the scales were frequency and point of acquisition. Rating scale is a categorical variable and point-biserial correlation coefficients were planned to address this question.

RQ4: Is each post hoc detection index equally useful for identifying careless responding in job analysis surveys?

$H_0 4$: All indices will be equally useful in flagging careless responding.

$H_a 4$: All indices will not be equally useful in flagging careless responding.

Addressing this question involved exploring the relationships among the indices through correlation and factor analysis and identifying data patterns and numbers of individuals flagged by each index.

RQ5: Is there a relationship between careless responding and the psychometric characteristics of job analysis data?

$H_0 5$: There is no relationship between careless responding and the psychometric characteristics of job analysis data.

H_a5: There is a relationship between careless responding and the psychometric characteristics of job analysis data.

To address this question, rating scale reliability, interrater reliability, and correlations among mean task or knowledge ratings were calculated before and after removing responses flagged as careless.

RQ6: Are there differences in terms of selection of tasks for a certification test content outline?

H_06: There is no difference in terms of tasks and knowledge for a certification test content outline?

H_a6: There are differences in terms of tasks and knowledge selected a certification test content outline?

To address this question, thresholds used by credential sponsors organizations were applied before and after removing responses flagged as careless.

Theoretical Framework

Satisficing is a framework for understanding suboptimal survey responding (Krosnick, 1999). When responding to a survey question, the participant must undertake a four-step cognitive process. First, the participant must read and understand the meaning of the question. Second, they must perform a memory search to retrieve relevant information. Third, they must integrate the information into a judgment. Fourth, they must accurately convey that judgment in accordance with the required response options (Tourangeau, 1984). For optimal responding to occur, a respondent

must devote cognitive effort to all four steps. A respondent who is satisficing skips or shortcuts one or more steps in the process, thus providing less meaningful information (Vannette & Krosnick, 2014).

Satisficing is influenced by three factors: lack of motivation, lack of ability, and task difficulty (Krosnick & Presser, 2010). Responding to a survey may require considerable cognitive effort (Vannette & Krosnick, 2014). Respondents might have difficulty answering survey questions due to their inability to engage in the four-step cognitive process. For those with the ability to respond, questions posed and judgments required may be challenging. Finally, even respondents who can respond and do not find the task too difficult may experience a waning in motivation to continue to expend effort, particularly when responding to lengthy surveys. Satisficing is more likely to occur when any of these three factors is present (Krosnick, 1999).

Satisficing offers a theoretical explanation for CI responding. Job analysis surveys require repetitive ratings for lists of tasks or knowledge areas. Typically, more than one rating scale is employed, doubling the number of judgments required for the listed items. Based on satisficing theory, predictions regarding the level of carelessness in different job analysis ratings contexts can be made. For example, the theory specifies that motivation is the proximal cause of carelessness and factors such as survey length and cognitive difficulty making the ratings are demotivating, resulting in more satisficing. Satisficing and its relationship with CI responding are discussed in more detail in Chapter 2.

Nature of the Study

The nature of this study was quantitative. A quantitative approach is appropriate for exploring the relationship between the predictor variables (length of survey, type of rating scale, presence or absence of incentives, and type of incentive) and criterion variables (extent of CI responding using three different detection indices, job analysis psychometric properties, and tasks selected for a test content outline). Secondary data analysis was undertaken to address the research questions.

Secondary data analysis is a cost-effective and efficient means of comparing data collected in different contexts (Frankfort-Nachmias & Nachmias, 2008; Johnston, 2014). Curran, Kotroba, and Denison (2010), Huang, Liu, and Bowling (2015), and Johnson (2005) used archival data in their CI research. In the present study, job analysis survey data already collected for three different professions' certification programs were re-analyzed.

Definitions

The following operational definitions were adopted for this study.

Abstract rating scale: Error! Bookmark not defined.*A*bstract rating scales require survey takers to make subjective evaluations of job tasks or KSAOs, such as importance to the job overall or importance to public protection. Scale anchors for abstract scales are non-verifiable (Harvey & Wilson, 2000). An example of an abstract rating scale is an importance scale for which respondents indicate whether a task is not important, minimally important, moderately important, or highly important to protecting patient

health and safety. Relative to concrete rating scales that focus on observable aspects of the job, abstract scales pose a heavier cognitive burden to respondents due to their greater information-processing demands and required level of inference (DuVernet, Dierdorff, & Wilson, 2015).

Certification Test Content Outline: A certification test content outline is comprised of domains of practice, tasks performed within domains, and optionally knowledge and skills. A content outline is hierarchical; the highest level is typically comprised of four to seven broad domains that encompass all tasks performed on the job. The outline specifies the percentage of test questions to assess content related to tasks in each domain. The percentage weights are derived empirically from job analysis ratings. The second level of a content outline is a list of tasks performed in the domain. Only tasks validated based on job analysis survey ratings are included.

CI Responding: This is "a survey response set in which a person responds to items without sufficient regard to the content of the items and/or the survey instructions" (Huang, Bowling, et al., 2015, p. 828). CI responses are a confound in survey data and researchers advocate screening for and removing these responses. Several screening methods exist, including explicit instructions, bogus survey items, self-report data, response time analysis, and post-hoc detection indices (Curran, 2016; DeSimone et al., 2015).

Concrete Rating Scales: Concrete ratings scales require survey takers to report on aspects of a job, such as how frequently they perform specific tasks. Concrete scales use

verifiable anchors (Harvey & Wilson, 2000). An example of a concrete rating scale is a frequency scale for which respondents indicate whether a task was performed daily, weekly, monthly, quarterly, or annually. Relative to abstract rating scales, which require subjective evaluations, concrete scales focus on observable aspects of work. Little inference is required to respond to concrete rating scales.

Job Aspects: Job aspects describe work activities and worker characteristics that are delineated in a job analysis study. In certification job analysis, the job aspects studied most frequently are domains of work and associated tasks (Raymond, 2001, 2002). Job analysis studies in certification may also include the testable knowledge and skills required to perform tasks in each domain. Job aspects are descriptors of various aspects of a job, occupation, or profession that are delineated in a job analysis study. In certification job analysis, the job aspects studied most frequently are domains of work and associated tasks (Raymond, 2001, 2002). They may also include the testable knowledge and skills required to perform tasks in each domain.

Post-Hoc Detection Indices: Post-hoc detection indices are mathematical or statistical calculations applied to already-collected survey data. Three such indices are used in this study: long string analysis, Mahalanobis distance, and person-total correlation.

Assumptions

An assumption underlying this study was that most responses to the job analysis surveys being studied represented careful and thoughtful responding. Some of the

indices being used will produce spurious results if they are run on datasets in which the percentage of carelessness responders exceeds 20%. While Curran (2016) estimates that the 10% of respondents are careless, which should mitigate the risk associated with excessive careless data, the baseline rate in job analysis is not known. However, I am assuming that enough careless responding exists in the datasets I am studying to permit hypothesis testing involving comparisons between groups.

Scope and Delimitations

Job analysis informs all HR functions within organizations; however, this study focuses only on job analyses conducted outside of a specific organizational context. National job analyses surveys that support certification programs sample broadly from individuals representing a variety of employers. While survey takers will reflect their own organizations' ways of working when they make their survey ratings, the lack of systematic variance from a single organization will not dominate the results.

The study involves three jobs: literacy coach, patient care technician, and pharmacy technician. Findings from this study may generalize to occupations that have similar eligibility criteria for their certifications. Caution needs to be taken when generalizing to professions requiring significant postsecondary education, because carelessness is related to level of education, with more carelessness associated with lower levels of education (Anduiza & Galais, 2017; Bowling et al., 2016; Morgeson et al., 2014; Roivainen et al., 2016). Professions requiring a master's degree or higher, such as

nursing, pharmacy, law, and occupational therapy may be associated with higher levels of professionalism and engender greater sustained attention and hence less carelessness.

Limitations

An obvious limitation of the study is that it employs a nonexperimental design. The absence of experimental controls means that noncontrolled variables may have an impact on the variables of interest. This may mask true differences between variables and inflate type II errors. The study design does aim to identify sources of variance hypothesized to relate to carelessness, but other sources likely exist that are not addressed. Because the study employs secondary analysis, the data have already been collected. Manipulating variables or asking additional questions about variables believed to relate to carelessness cannot be done. Should significant results not be obtained, it may be because of confounding variables. In this study, the disadvantages of the nonexperimental design can be weighed against the advantage of undertaking secondary data analyses of multiple studies. When several job analyses are examined simultaneously, comparison across studies is possible. Should findings occur in multiple studies, this strengthens the generalizability of the results.

Significance

By examining the extent of CI responding in job analysis studies, factors that influence it, and best methods for detecting and eliminating it, the hope is that guidance for future job analysts using an Internet-based approach to survey delivery can be developed. Using this guidance, job analysts can design surveys to minimize careless

responding and employ optimal detection techniques to identify and mitigate it when it does occur. Improving data accuracy enhances the validity of decision making based on the data. Given the past and expected growth of occupational certification, enhancing data quality will lead to the highest quality test content outlines, which are used in all aspects of certification program development.

More broadly, the findings from this study may inform a larger audience of survey researchers. The use of surveys in data collection is widespread, and the rise of the Internet has increased the frequency of survey use (Hauser & Schwarz, 2015). Guidance regarding the relationships between survey length, rating scale choice, and rates of careless responding can lead to more accurate data for a variety of purposes and strengthen research practices across disciplines.

Summary

This chapter introduced the data quality issue of CI survey responding, or responding without sufficient consideration of question content, a topic which has recently received increased research scrutiny. CI survey responding distorts survey results, and methods to detect and mitigate it have been proposed. Carelessness is particularly a problem in lengthy Internet-based surveys such as job analysis surveys administered in support of professional certification programs. A gap in the literature related to carelessness in job analysis surveys was articulated. A study addressing this gap and the problem posed by it was proposed, and testable hypotheses were specified. Definitions were supplied for terminology used in the study. The scope, delimitations,

and limitations of the study were discussed, and the significance of the research was outlined.

In the next chapter, the existing literature related to CI responding is discussed. Chapter 2 addresses the emerging survey research literature on carelessness, as well as literature on job analysis quality. In addition, different types of careless survey responding are defined, and methods to screen for and detect carelessness are described. Finally, the research to date on post hoc detection indices is reviewed.

Chapter 2: Literature Review

Introduction

CI responding is a mode of survey responding in which less-motivated survey takers respond without due reference to the survey questions or instructions (Huang, Liu, and Bowling, 2015). As such, it is distinct from impression management and other systematic forms of bias. It can distort reliability of measurement and the validity of interpretations of results. Detection and elimination of such survey responses can produce data that better reflect careful responders' judgments regarding the variables under study. Although under certain conditions removal of CI can decrease reliability rather than increase it, score reliability will generally increase scale reliability, item intercorrelations, and validity (Huang et al., 2012).

The purpose of this study is to employ post-hoc analysis to determine causes, types, and extent of CI responding in job analysis surveys used to determine the content of licensure and certification examinations by exploring the effects of CI data on scale reliability, correlations among tasks and knowledge, and decisions regarding testable content. Satisficing theory provides the theoretical basis for the study, as it offers an explanation for the relationships between survey length, job aspect rated, and type of rating scale used and the extent and type of CI responding.

Job analysis surveys were selected for study because of the importance of job analysis data in HR. Data from job analysis surveys are used to create position descriptions, candidate assessments, incumbent education and training material, and

promotion and performance evaluation systems (Morgeson & Campion, 2017; Sanchez & Levine, 2012; Singh, 2008). Job analysis to support licensure and certification testing was chosen because of the high-stakes nature of decisions made based on test scores. In addition, because 24% of the US population holds licenses or certifications, the potential reach of inaccurate job analysis data and test content outlines of questionable validity is wide. Finally, job analysis surveys are of a type likely to induce carelessness due to their length and numerous judgments required (Huang et al., 2012; Morgeson & Campion, 1997).

In this chapter, I review literature related to the topic. First, I review the job analysis carelessness literature. Second, I review the recent literature on CI responding. The review encompasses deterrence and detection methods, and findings from studies using post-hoc detection methods. Third, I discuss theoretical explanations of survey responding in general and CI responding specifically, including Krosnick's influential theory of satisficing. Fourth, I discuss survey design features likely to influence CI responding.

Literature Search Strategy

The following databases were searched: PsycINFO, PsycArticles, Academic Source Complete, Business Source Complete, and SAGE Journals. The following search terms were used: *satisficing*, *survey*, *random responding*, *inconsistent responding*, *careless responding*, *inattentive responding*, *insufficient effort responding*, *job analysis*, *work analysis*, *ratings*, *rating scales*, *practice analysis*, and *data quality*. Reference

sections of all articles obtained during the initial search were examined for further relevant references. As the literature review progressed, historical perspectives regarding survey responding from 1970 onward were included.

Job Analysis Literature

Job analysis is a foundation for virtually all HR functions, including selection, evaluation, and promotion (Morgeson & Campion, 1997; Siddique, 2004; Van Iddekinge, Putka, Raymark, & Eidson, 2005). Given the changing nature of work in the 21st century, the term work analysis has been more recently adopted to refer to the set of techniques used to identify key tasks and KSAOs required for a position (Sanchez & Levine, 2012). In addition to its role within organizations, job analysis plays a central role in the development of test content outlines for licensure and certification examinations (Raymond & Luecht, 2013; Wang et al., 1999). Accreditation requirements for credentialing programs specify that a job analysis must be conducted and used as the basis for examination development (ICE, 2014; International Standards Organization/ International Electrochemical Commission, 2012).

Classical test theory underlies much of the research on job analysis accuracy (Morgeson & Campion, 2000), with the assumption that true scores exist and that variation represents noise in the data. Inconsistency among survey takers' is not necessarily indicative of inaccuracy (Lievens, Sanchez, Bartram, & Brown, 2010). Lievens et al. pointed out that, to some extent, differences between raters can be job-related and logical. For example, level of autonomy, cognitive ability, and job skill

predicted differences in the number of job tasks performed by administrative professionals (Morgeson, Delaney-Klinger, & Hemingway, 2005). Occupational complexity predicted variance in competency ratings (Lievens et al., 2010), suggesting that greater professional autonomy in complex occupations permits individualized work patterns. However, not all variance is explainable. Van Iddekinge et al. (2005) found that the majority of variance in KSAO importance and needed at entry ratings made by customer service managers was unexplained. Job and organization-level factors could not account for it, and age and gender effects were nonsignificant. Level of experience may influence job analysis ratings. Richman and Quinones (1996) found that undergraduate participants in a laboratory study in which they rated relative and absolute frequency of task performance after either building or observing the building of a toy model that those who had less experience building or observing model building gave more accurate frequency ratings.

Potential sources of inaccuracy in job analysis ratings include both social and cognitive factors (Morgeson & Campion, 1997, 2012). While these sources have been proposed, they have not yet been studied exhaustively. Social causes that might affect responding to Internet-based job analysis surveys include distance from the researcher and anonymity of responses. Cognitive causes include limitations in information processing, fatigue, and the adoption of heuristics. Morgeson and Campion (1997) posited that rating inaccuracies in job analysis would be more likely when more-subjective inferences were required, for example, when rating KSAOs rather than tasks.

Morgeson and Campion (2012) outlined six potential manifestations of inaccuracy that affect reliability and validity: interrater reliability, interrater agreement, discriminability, dimensionality, mean ratings, and completeness.

Studies employing a bogus item approach to inaccurate responding found high rates of endorsement of bogus items. Green and Stutzman (1986) had mental health workers complete a 115-item task inventory, making ratings of time spent relative to other tasks and importance to the job. For each rating scale, they calculated the number of bogus items endorsed. Fifty-seven percent of job incumbents indicated they spent time performing at least one of the bogus tasks and 72% rated at least one bogus task at least somewhat important to their jobs. Pine (1995) found that 45% of corrections officers indicated they performed at least one of five bogus items in a job analysis survey. Stetz et al. (2012) found that phrasing of questions influenced level of endorsement of bogus tasks. Between 85% and 97% correctly indicated they never performed a bogus task in the past 12 months, yet 39% to 64% incorrectly indicated they were expected to perform the bogus task. Stetz et al. concluded that "as a scale becomes less specific, less observable, and more ambiguous, there is a corresponding increase in rating inaccuracy" (p. 105). Green and Veres (1990) found that 70% of police corporals and 13% of mental health workers endorsed bogus items. Additionally, they found that the effect of bogus item endorsement on reliability was inconsistent and differed by profession. For mental health workers, reliability of a scale asking whether tasks were performed at entry level was higher for respondents who were accurate in their ratings. However, this relationship

was not found for an importance scale, nor was it found on either scale for police corporals or clerical workers. In addition, mean task ratings were higher for inattentive respondents, suggesting a pattern of greater endorsement of items in general for respondents who had at least one incorrectly endorsed a bogus task. Finally, in a job analysis survey of government employees in an agency for international economic development that included bogus items, Morgeson et al. (2014) found large correlations between the endorsement of bogus and legitimate tasks.

Wilson, Harvey, and Macy (1990) used repeat items in a single task inventory to explore short-interval test-retest reliability. Although not treated as such by Wilson et al., one could consider lack of consistent endorsement or non-endorsement as in indicator of CI responding. For city municipal workers, 10 of 41 (24%) responded carelessly, as did 6 of 34 (18%) hospital foodservice employees and 5 of 20 (25%) manufacturing workers. Together with the findings related to bogus tasks, these findings suggest relatively high rates of carelessness on job analysis inventories.

Generalizability studies suggest that 5% to 9% of variance in job analysis ratings can be attributed to raters. In a study of competencies across 64 occupations employing a Q sort methodology (Lievens et al., 2010), the raters factor accounted for an average of 5% of the variance in individual competency ratings. In a study of job analysis surveys for two professions undertaken to create licensure examination content outlines (Wang et al., 1999), raters accounted for 7 to 9% of variance in task ratings. Based on IRT infit

and outfit statistics, Wang et al. were able to identify raters who consistently selected the middle or extreme categories.

Two recent studies in the job analysis literature examined ratings carelessness. Dierdorff and Morgeson (2009) examined O*Net ratings made by 47,137 incumbents in over 300 jobs based on work characteristics (i.e., tasks and responsibilities) and worker characteristics (i.e., knowledge, skills, and traits). The smallest amount of rater variance was for tasks (11.7%) and the largest percentage (34.8%) was for trait ratings. Knowledge ratings exhibited slightly more variance (14.6%) than task ratings. Task ratings were more reliable (.80) than knowledge ratings (.70), and trait ratings had the lowest reliability of all work descriptors studied (.45). Dierdorff and Morgeson hypothesized that these differences were due to differing levels of inference required when making ratings. Morgeson et al. (2014) explored the relationship between holistic ratings of major job components and decomposed ratings of specific tasks in each component in a job analysis of government aid workers. Three bogus tasks were added to the survey and the number of bogus tasks endorsed was the measure of carelessness. Morgeson et al. found that the number of careless responses was negatively related to the consistency between respondents' holistic and decomposed ratings. They hypothesized that inconsistency in ratings and endorsement of bogus tasks were both indicative of respondent carelessness. Because they did not employ cutoffs to classify raters as careless or not, the rate of carelessness in this study is not known. Neither study used post hoc detection methods.

Historical Perspective on Survey Response Accuracy

The quality of survey data has long been a focus of research interest (Alwin, 2016). Recognizing the potential impact on reliability of data and the validity of inferences drawn from data, researchers have studied response rates and response bias for decades (Johnson & Wislar, 2012). Response rate research primarily focuses on the adequacy of the respondent group as a representation of the population (Groves et al., 2009). To the extent that the respondents share characteristics of the population, generalizations of findings can be made with greater confidence. If respondents are not representative, results may provide a distorted view of the population (Anseel, Lievens, Schollaert, & Choragwicka, 2010). Response bias research focuses on conscious or unconscious response distortions, including socially desirable responding, faking, agreeableness, and acquiescence (McGrath, Mitchell, Kim, & Hough, 2010). These have been studied extensively in the context of personality assessments such as the Minnesota Multiphasic Personality Inventory and the NEO-PI, where specific scales have been developed to flag incongruous data (Baer, Ballenger, Berry, & Wetter, 1997; Berry et al., 1992, 1991; Kelley et al., 2016; McCrae, Kurtz, Yamagata, & Terracciano, 2011; Piedmont, McCrae, Riemann, & Angleitner, 2000).

Rates of CI Responding

The exact extent of CI responding in survey research is difficult to determine. Studies have identified widely differing estimates of the rate of CI responding (Curran, 2016). Estimates outside the job analysis arena range from a low of 1% (Johnson, 2005)

to a high of 50.5% (Curran et al., 2010). In their seminal work on careless responding, Meade and Craig (2012) estimated that between 10% and 12% of survey responses could be identified as careless. Maniaci and Rogge (2012) identified between 3 and 9% of their respondents as highly inattentive. A more recent estimate derived from a review of existing literature estimated the rate of CI responding at 8% to 12% (Curran, 2016). Most recently, Steedle (2018) found that 43% of responses to a college readiness measure could be classified as careless by at least one of nine detection methods studied. Some of the variance in these estimates can be attributed to the different survey instruments studied, indices used, and cutoff thresholds applied.

Deterrence and Detection Methods

Two general classes of strategies exist for mitigating CI responding: deterrence and detection. Deterrence strategies focus on encouraging respondent to be careful and attentive throughout the survey process. Detection strategies focus on identifying CI responding after responses have been collected. The latter strategies employ mathematical and logical analyses.

Deterrence Strategies

The three most commonly used deterrence strategies are instructional manipulation checks (IMCs), instructed response questions, and the infrequency approach. The IMC technique was developed by Oppenheimer, Meyvis, and Davidenko (2009). In this approach, a lengthy paragraph of text is provided along with appropriate answer choices. Embedded in the paragraph are instructions to respond in a specific way

unrelated to the response choices provided, for example, to select, *I have read these instructions,* a message placed in a different screen location than the response options. IMCs are designed to measure attentiveness in reading instructions. The assumption is that failing an IMC implies a lack of attention to survey instructions in general (Hauser & Schwarz, 2015). A limitation in this approach is that a respondent's level of attention may not be consistent throughout a survey (DeSimone et al., 2015). Instructed response questions explicitly instruct survey takers to answer questions in a specific way, e.g., *for this question, select strongly agree.* The infrequency approach seeds surveys with highly improbable questions for which the answer is obvious, for example, *I have been to every country in the world* (Maniaci & Rogge, 2014). Incorrect answers imply that attention to such questions was minimal.

Detection Strategies

Detection strategies are post-hoc logical and mathematical processes for identifying CI responding in collected survey data. Because deterrence methods are intrusive and may not be well received by survey takers (Curran, 2016), and are not 100% successful in curbing CI responding, post-hoc measures provide an alternate and complementary means to identify careless responses. CI detection methods have been described in detail by Curran (2016), Huang et al. (2012), and Meade & Craig (2012). They can be grouped into conceptually related approaches: inconsistency, invariance, and outlier.

Inconsistency approach. The within-person inconsistency approach focuses on a single individual's responses and their level of internal consistency (Curran, 2016). Several CI detection methods fall within this category. The first is the repeated items approach. When an item is repeated at different points in a survey, inconsistent responses can indicate inattention and lack of effort (DeSimone et al., 2015). In the semantic synonym and antonym approach (Goldberg & Kilkowski, 1985), items with near identical or near opposite meanings form item pairs. Inconsistent responding to these pairs similarly can indicate inattention. The premise of this approach is that careful respondents should correctly give identical or opposite answers to the pairs. Psychometric synonyms and antonyms (Johnson, 2005) are pairs formed on the basis of high positive or negative intercorrelations, irrespective of item meaning. Odd-even consistency (Jackson, 1977), which Huang et al. (2012) refer to as individual reliability, is the correlation between odd and even numbered items in a scale. This approach is useful when applied to unidimensional scales. Newer methods of examining intraindividual consistency, including polytomous Guttman errors and the inter-item standard deviation, are beyond the scope of this study. The interested reader is referred to Curran (2016) for a discussion of these methods.

The person-total correlation can be conceptualized as a measure of between-person consistency. This index is derived by inverting the item-total correlation matrix familiar in item analysis work to correlate an individual's consistency in responding with the consistency of all other survey takers' responses. The person-total correlation is

relatively unstudied to date. The only study I was able to locate (Dupuis, Meier, & Cuneo, 2018) found it to be a good predictor of simulated, non-human random responses.

Invariance approach. The invariance approach to CI response detection assumes that sequential identical responses to numerous items in a response sequence suggests a lack of attention to nuances in the items rated. Long string analysis (Herzog & Bachman, 1981; Johnson, 2005) is the primary means of exploring data for invariant patterns. It simply involves identifying the longest string of identical responses. Curran (2016) refers to cases identified through long string analysis as the "low-hanging fruit of CI responders" (p. 8).

Outlier approach. The outlier approach is exemplified by the Mahalanobis distance technique (Mahalanobis, 1936). Mahalanobis distance is a multivariate outlier detection technique, calculated by using the inverse of the variance-covariance matrix of the survey data (De Maesschalck et al., 2000). In recent studies, Mahalanobis distance has shown promise in detecting CI responding (Curran, 2016; Meade & Craig, 2012; Ward & Pond, 2015).

Other methods. Other methods not easily categorized as deterrence or detection approaches are response time and self-report measures. Response time is frequently used as a proxy for survey attention, with the premise that quicker responding implies a more superficial level of processing. Self-report measures are single-item survey questions included at the end of the survey addressing level of effort expended or attention devoted (DeSimone et al., 2015).

Types of CI Responding

In their research on the MMPI, Nichols, Greene and Schmolck (1989) drew a distinction between two types of problematic responses: content responsive and content non-responsive. Content responsivity occurs when respondents deliberately select answers to create an impression, either by faking good or faking bad. Content non-responsivity occurs when the test taker's response is unrelated to the item content. Data indicative of content non-responsivity includes patterned responding (e.g., selecting 1 to all questions on one page and selecting 2 to all questions on the next), random responding, and invariant responding.

Meade and Craig (2012) were the first to identify the latter as careless or inattentive responding in a two-study exploration of the phenomenon. In the first study, undergraduates took an Internet survey of 300 items from the International Personality Item Pool (IPIP; (Goldberg, 1999). Participants answered questions under one of three conditions: anonymity, required name identification, or warning regarding response integrity. Multiple indicators of CI responding were used. Within the survey, indicators included 10 infrequency items; self-report questions regarding levels of engagement, attention, and effort; and a final question regarding whether the respondent's survey data should be used. Post-hoc measures were time to complete the survey, number of infrequency items answered incorrectly, correlations between psychometric synonyms and antonyms, odd-even consistency, average and maximum length of long strings of invariant responses, and Mahalanobis distance.

Meade and Craig (2012) found that the different experimental conditions generated modestly differing amounts of CI responding, indicating that survey administration factors can influence rates of carelessness. The rate across conditions and indices ranged between 10% and 12% of total responses. Regarding relationships among different indices, exploratory factor analysis yielded three factors. The first was comprised of the consistency measures, Mahalanobis distance, and infrequency. The second was comprised of the self-report measures, and the third consisted of the two long string measures. This result suggested that there were different types of CI responding. Latent profile analysis of the post-hoc measures revealed that the measures were tapping into two different classes of responders. One was characterized by inconsistent responding and the other by patterned responding. Independent replication confirmed this distinction (Huang et al., 2012; Maniaci & Rogge, 2014).

Meade and Craig's second study was a simulation in which level of carelessness (full versus partial), type of carelessness (random versus patterned), and extent of carelessness (5%, 10%, or 15%) were manipulated. The goal was to explore how well psychometric synonyms, psychometric antonyms, odd-even consistency, and Mahalanobis distance detected CI responding under these different conditions. The outlier index Mahalanobis distance had the highest sensitivity and specificity for uniformly distributed random data across all levels of carelessness, but was worst for detecting responses with partially random, normally distributed data. Under conditions of uniformly distributed careless data, the odd-even consistency index performed better

than psychometric synonym and antonym conditions. For normally distributed random data, the odd-even consistency index worked well under conditions of total carelessness but poorly under conditions of partial carelessness. An important implication of Meade and Craig's finding, corroborated by others (Curran, 2016; DeSimone et al., 2015; Huang et al., 2012), was that survey researchers should employ multiple indices tapping into these different response types.

In two studies, Huang et al. (2012) employed both deterrence and detection approaches to examining CI responding. In the first study, they had a group of undergraduates take 300 items from the International Personality Item Pool. Responses were made using a 5-point Likert scale. Survey instructions were manipulated such that half the respondents received instructions to "describe yourself honestly" and the other half were warned that their data would be checked and that bad data would result in loss of research credit. In the second half of the survey, respondents experienced one of three conditions. Respondents were instructed either to continue as in the first half, to respond as if lazy, or to respond without effort. Five CI responding measures were employed: odd-even consistency, response time, long string analysis, psychometric antonyms, and individual reliability (referred to by other researchers as odd-even consistency). The indices demonstrated convergent validity in both correlational analysis and factor analysis. Respondents who scored high on one CI index tended to score higher on the others. In addition, levels of careless responding were higher for the lazy responding and responding without effort groups than for the group instructed to respond honestly. Study

2 was a replication later in the semester, a point at which Huang et al. posited that CI responding would be higher as students rushed to complete their research credits. In addition, self-report questions were added to the end of the survey inquiring about level of effort. As in Study 1, manipulations of levels of warning had a direct effect on levels of CI responding. Correlations among the five detection indices ranged from .18 to .69. All five indices loaded on a single factor suggesting they measured a single construct. These findings contradicted Meade and Craig (2012), but results may not be directly comparable because different sets of detection methods were used.

The work of Meade and Craig (2012) and Huang et al. (2012) spurred numerous studies examining the utility of post-hoc indices to detect different types of CI responding (Bowling et al., 2016, Huang, Bowling et al., 2015; Huang, Liu, & Bowling, 2015; Maniaci & Rogge, 2014; McKay, Garcia, Clapper, & Shultz, 2018; Steedle, 2018; Thomas, 2014; Ward & Pond, 2015; Zijlstra, Van der Ark, & Sijtsma, 2011). Each employed different detection indices and applied them to different types of data, making it difficult to generalize regarding the overall utility of each index. One finding that has emerged is that different post hoc detection measures are sensitive to different types of CI responding. The measures are summarized well by Meade and Craig (2012) and Curran (2016) and are described here based on their excellent work. Long string analysis detects straightlining or invariant responding and is most useful for lengthy surveys employing the same rating scale or thematically similar items to rate. Mahalanobis distance is used to detecting data patterns that differ from those of most respondents. This is a more

complex type of careless responding that cannot be identified by other means (Meade & Craig, 2012). The person-total correlation similarly compares the individual with the rest of the sample and requires low rates of CI responding in the total sample to produce meaningful results (Curran, 2016). The odd-even correlation is useful to detect illogical response patterns when applied to unidimensional scales. Semantic and psychometric synonym and antonym pairs can perform well in detecting carelessness but require judgment in determining how high a correlation is sufficient to warrant the pairing of items.

Applicability of Detection Methods to Job Analysis Studies

To date, no studies have examined the use of post-hoc analyses in the job analysis context. Not all indices are applicable to job analysis. Long string analysis is applicable, given the lengthy and repetitive nature of job analysis ratings. Mahalanobis distance and person-total correlation could help identify respondents whose response patterns do not relate to those of most other respondents. Odd-even correlations would not be useful, due to the multidimensionality of job analysis structures. Semantic synonyms or antonyms are an unlikely choice, given that each ratable item in a job analysis inventory is intended to describe a unique task, responsibility, or KSAO. The use of psychometric synonyms or antonyms is a possibility but given that all items on the inventory are intended to describe a single job, the existence of antonyms is unlikely. The measures most applicable and selected to be explored in this study are long string analysis, to detect

invariant responding, and Mahalanobis distance and person-total correlation, to detect more subtle patterns of CI responding.

Theoretical Foundations

Taking a classical test theory approach to survey responding, accurate survey responses can be thought of as representing true scores for the variables being studied. Responding is the result of a four-stage cognitive process (Tourangeau, 1984; Tourangeau, Rips, & Rasinski, 2000). The respondent first decodes the question text to infer the survey sponsor's meaning. Next, the respondent conducts a mental search to identify stored information related to the question. Third, the respondent integrates the information, and fourth, the respondent "maps their judgment onto a response category" (Tourangeau et al., 2000, p.8) or generates a response if the question is open-ended. Errors can occur at any of the four stages (Krosnick, 1991; Tourangeau et al., 2000), resulting in inaccuracy.

Due to the cognitive demands of survey response process, there is "considerable room for error" (Tourangeau, 1984, p. 73). The effort of responding is such that respondents may not attend equally to all four stages. According to Krosnick (1991, 1996), this lack of attention yields satisficing, or suboptimal responding. The term satisficing, a combination of *satisfy* and *suffice* (Daniel, 2012), originated with Simons (1957), who was studying decision making in general. Simon's proposed that rather than conducting an exhaustive evaluative process when engaged in decision-making, "people expend only the effort necessary to make a satisfactory or acceptable decision"

(Krosnick, 1996, p.30). Barge and Gehlbach (2012) first introduced the idea of applying satisficing theory to data quality issues. More recently, it has been specifically applied to careless and inattentive responding (Steedle, 2018).

Krosnick draws a distinction between two types of cognitive shortcuts: weak and strong satisficing. Weak satisficing occurs when respondents engage in all four phases but are not fully committed cognitively. Strong satisficing occurs when stages are "the retrieval and judgment states are skipped entirely" (Krosnick, 1996, p. 31). Instead, "the answer is selected without referring to any internal psychological cues specifically relevant to the attitude, belief or event of interest" (Vannette & Krosnick, 2014, p. 315). Response strategies associated with weak satisficing include searching for the first plausible answer and acquiescence. Responses associated with strong satisficing include endorsing no opinion and arbitrary responses. Optimizing and strong satisficing represent two ends of a continuum.

Three factors influence the extent of survey satisficing: the difficulty of the task, the ability of the respondent, and the level of motivation to optimize (Krosnick, 1999). Task difficulty encompasses factors including the level of complexity of the questions and response options, respondent challenges in retrieval, and the level of environmental distraction. Ability is influenced by cognitive adeptness at the steps in responding and familiarity with the topic. Motivation to optimize is influenced by the need for cognition, importance of the topic to the individual, perceived value of the survey, and the length of the survey. Support for all three factors has been found (Krosnick, 1987, 1991, 1996).

There are various ways a survey task can be difficult. The first is the complexity of the questions and answer options. In the job analysis literature, greater carelessness has been found for ability ratings, which are more abstract, than for task, knowledge, or skill ratings, which are more concrete (Morgeson et al., 2004). In addition, carelessness may greater for relativistic rating scales, which require the respondent to consider a task/KSAO in relation to all other tasks performed/KSAOs employed. Distraction can also increase the survey difficulty. The majority of surveys currently administered via the Internet are likely taken under conditions of environmental distraction (Hardré et al., 2012; McKay et al., 2018) and multitasking (Carrier et al., 2009).

Cognitive ability appears related to the tendency to satisfice. Using data from the European Social Survey and measures including semantic inconsistency, straightlining, and percentage of "don't know" responses, higher cognitive ability was found to be associated with lower levels of satisficing (Kaminska, McCutcheon, & Billiet, 2010). Similar findings regarding level of education and measures of carelessness have been found in the job analysis literature (Green & Veres, 1990; Zhang & Conrad, 2014).

Motivation can influence satisficing through fatigue effects. Over the course of a lengthy survey, response fatigue may decrease motivation to respond accurately (Harvey & Wilson, 2000). Numerous studies have found that survey length has a negative effect on overall response rate (Cook, Heath, & Thompson, 2000; Fan & Yan, 2010; Guo, Kopec, Cibere, Li, & Goldsmith, 2016). For example, Sarraf & Tukibayeva (2014) found the number of survey pages was correlated with the level of item nonresponse.

Deutskens, De Ruyter, Wetzels, & Oosterveld (2004) found that the absolute number of ratings was correlated with the level of survey nonresponse.

One study examined all three factors that satisficing theory predicts as influencing the quality of responses (i.e., difficulty, cognitive ability, and motivation) simultaneously (Hamby & Taylor, 2016). Randomly assigning Amazon Mechanical Turk (MTurk) members and college students to surveys that varied by number of rating scale response options and labels, they found that satisficing behavior, which they defined as straightlining in one or more personality scales, was more prevalent for respondents with less than a college degree and those motivated by pay, that is, the MTurk sample. Contrary to expectation, there was less satisficing when rating scales had more options than when they had fewer.

Individual Differences in CI Responding

Literature outside of the job analysis context suggests that there are both state and trait components of CI responding. Recent studies examining the relationship between personality and CI responding suggest several relationships between personality and this response type. With respect to the big five personality factors, both conscientiousness and agreeableness have shown a consistent inverse relationship with CI responding (Kelley et al., 2016; McKay et al., 2018; Meade & Pappalardo, 2013; Ward, Meade, Allred, Pappalardo, & Stoughton, 2017). Extroversion was found to have a positive relationship with CI responding (Meade & Pappalardo, 2013), while malevolent personality traits showed an even stronger relationship to CI responding than benevolent

ones (McKay et al., 2018) for long string analysis and instructed response items, but not for Mahalanobis distance or length of time to respond to a survey. In summary, the evidence to date suggests that there is a trait-based component to CI responding.

In addition to personality traits, other individual differences may play a role in careless responding. Level of education has been found to relate to invariant responding, with less educated survey takers having higher rates of straightlining (Zhang & Conrad, 2014). Using an infrequency approach to detecting carelessness, Roivainen et al (2016) found that males and respondents with less than a high school education had higher rates of endorsement of bogus items. In contrast, Oppenheimer et al (2009) found no relationship between failure at an instructional manipulation check and respondent age or gender. The preponderance of evidence suggest that individual differences play a role in CI responding.

In the job analysis literature, the role of individual differences in survey responding paints an inconclusive picture of the relationship between demographic factors and survey responses. Green and Veres (1990) found a small but significant negative correlation between education level and scores on an infrequency scale for mental health workers. Morgeson et al. (2016) found that task, job, organizational, and career experience variables had no relationship to carelessness, and Van Iddekinge et al (2005) found that the rank ordering of KSAO ratings was not influenced by a customer service manager's organization, position level, age, or gender. Race does not appear to be related to job analysis carelessness (Landy & Vasey, 1991; Prien, Prien, & Wooten,

2003). These disparate findings leave the role of individual differences in job analysis responding open.

Design Characteristics

Design characteristics such as survey length and cognitive complexity of questions were described earlier in the context of general survey research (Herzog & Bachman, 1981). Additional studies of job analysis surveys found that survey length and rating scales used affected CI responding rates (Green & Veres, 1990; Wang et al., 1999). Dierdorff and Morgeson (2009) found differences in rating scale reliability based on the level of concreteness versus abstraction in job aspects rated.

Summary and Transition

The literature reviewed above suggest several pertinent observations regarding CI responding. First, the baseline rate of CI responding is unclear given that different studies using different types of surveys, rating scales, and deterrence and detection methods yielded widely varying estimates. The estimated base rate of 8% to 12% carelessness was derived from non-job analysis surveys. In job analysis research employing the bogus item technique, rates were much higher, ranging from 45% to 73% (Green & Stutzman, 1986; Green & Veres, 1990; Stetz et al., 2012). Higher rates in job analysis is consistent with the proposition that job analysis surveys are particularly prone to CI responding, given the lengthy lists of tasks and/or KSAOs to be rated (Morgeson & Campion, 1997, 2017). Second, the rate of CI responding is likely to vary by rating scale used (Stetz et al., 2012), job aspect rated, length of the data collection instrument

(Morgeson & Campion, 2017), and in the case of job analysis, the profession studied (Green & Veres, 1990; Lievens et al., 2010). Third, substantive survey ratings have typically found to be affected by CI responding. Fourth, indices to detect carelessness may be useful in identifying different patterns of CI responding. The utility of these indices with job analysis data is unknown and is the major focus of this study. In Chapter 3, a research design and a methodology to test hypotheses related to CI responding in job analysis surveys are discussed.

Chapter 3: Research Method

Introduction

The purpose of my research was to explore the extent, correlates, and consequences of CI responding (alternately referred to as careless responding) in occupational job analysis surveys. The extent was explored by calculating post-hoc detection indices and setting cut scores. The correlates were explored by examining the relationship between job analysis survey features and type and amount of CI responding. The consequences were explored by examining the effects of removing survey records with careless responses on reliability of measurement, intercorrelations, mean ratings, and tasks validated for inclusion on certification test content outlines developed from survey ratings. To my knowledge, this study represented the first regarding CI responding in job analysis employing post-hoc detection indices.

In this chapter, I describe the method to address CI responding. I describe and provide a rationale for the quantitative research design. Next, predictor and criterion variables are defined and operationalized. Hypothesized relationships among the variables are described along with the data analysis procedures to be used to test the hypotheses. Finally, ethical considerations and procedures are discussed and threats to the validity of interpretation of the study results are outlined.

Research Design and Rationale

A quantitative approach was used to explore the relationship of the predictor variables (length of survey, job aspect, and rating scale) and outcome variables (extent of

carelessness based on three different detection indices, relationship with survey features, psychometric characteristics of the data, and tasks included in test content outlines pre- and post-removal of CI data). Quantitative methodology was appropriate given that numerical indices were used to characterize extent of carelessness. All studies to date on CI responding have employed quantitative methods and my study was designed to build on the existing literature.

The study was archival in nature; secondary data analysis was undertaken to address the research questions. Secondary data analysis is an efficient and cost-effective way to the use of existing data for purposes other than those for which the data were originally collected (Vartanian, 2010). There is precedence for using archival data in research on careless responding (Huang, Lui, and Bowling, 2015; Johnson, 2005).

Secondary Data Sources

Data sources were job analysis datasets collected by current and former employers on behalf of sponsors of high-stakes certification examinations. The use of multiple datasets allowed for examination of survey responses within and across professions. The occupations represented by the job analysis surveys included in this study appear in Table 1. The year of data collection and number of tasks and knowledge areas in each survey are specified. The number of tasks ranged from 58 to 96 and the number of knowledge areas ranged for 54 to 170. All surveys include frequency and importance scales for tasks and frequency and point of acquisition scales for knowledge.

Table 1

Data Sources

Occupational Job Analysis	Year Data Collected	Number of Tasks	Number of Knowledge Areas
Literacy coach	2017	58	54
Patient care technician	2017	86	115
Pharmacy technician	2016	96	170

Conditions of Data Collection

Each survey was conducted to update an existing certification program and administered nationally to members of the profession. Respondents were routed to either a survey version containing tasks or one containing knowledge. This was the primary reason for selecting these studies. Because each dataset included ratings for both tasks and knowledge, comparison of CI responding for the two job aspects could be examined. A second rationale for selecting these studies was that each included two rating scales per job aspect. The two scales used to rate tasks were frequency of task performance in the past 12 months, measured on a 5-point scale with response options ranging from never to daily, and importance of the task to health/safety/outcomes, measured on a 4-point scale with response options ranging from not important to highly important. The two scales used to rate knowledge were frequency of knowledge use during the past 12 months, measured on a 5-point scale with response options ranging from never to daily, and point at which the knowledge should be acquired by members of the profession, measured on a 3-point scale with response options of never, before certification, and after certification. The latter scale is of importance in the certification context, where examination content

must be targeted to the just eligible candidate. Because each survey version employed two scales, comparison of CI responses to different scales for the same job aspect could be examined. Finally, each survey had different numbers of tasks and knowledge statements, with content determined by subject matter experts in the profession, permitting the exploration of survey length in relation to CI responding.

For each survey, email invitations were sent to either a random sample or the entire population of certified individuals, using the email of record maintained by the survey sponsor, either individual or organizational. Survey invitations provided a description of the purpose of the study, provided an assurance of confidentiality, and included a link to the survey. The estimated time required to complete the survey was specified. Participants were instructed that they could complete the survey over multiple sessions if desired. Incentives for responding were described. For the study of patient care technicians, Amazon Mechanical Turk (MTurk) was used to obtain a sample of noncertificants for comparison purposes. MTurk pays subjects for each survey taken, so the volume of surveys taken serves as a monetary incentive. The MTurk responses were excluded from data analysis because they were collected under different conditions from all other data and were not directly comparable.

Procedures for Obtaining the Data

By contract provisions, clients are the owners of the job analysis data I reanalyzed. Recruitment involved contacting each study sponsor and requesting permission to use their datasets. Prior to this, I obtained written permission from my

current and former employers to conduct this study using client data and to contact the study sponsors. Appendix A contains a letter of permission from my current employer to contact clients and request use of their data in my dissertation. Appendix B contains a similar letter from my prior employer.

Upon completion of Institutional Review Board (IRB) Form A, I obtained and submitted to the IRB signed approvals from study sponsors indicating that the sponsors agreed to release their data for my project. Subsequently, IRB approval was granted and IRB approval number 09-11-18-0479490 was assigned to the project.

Operationalizing the Study Variables

CI Responding

For all analyses, the amount of CI responding per subject was operationalized using three techniques: long string analysis, Mahalanobis distance, and person-total correlation. Long string analysis detects invariant responding, and Mahalanobis distance detects random, pseudorandom, and extreme responding (DeSimone et al., 2015). Person-total correlation is a newer detection method that compares the entirety of an individual's responses to all other survey takers' responses. Its inclusion in this study was exploratory. Other post-hoc analysis methods were excluded from consideration because they were better suited for use with Likert-type ratings scales, they required data elements not appropriate for job analysis surveys such as semantic synonyms, or they assumed scale unidimensionality. Job analysis data are organized within domains of practice, each representing a different dimension of the job.

Long String Analysis

Long string analysis flags ratings invariance in responses, which is indicated by long sequences of identical ratings. The assumption behind this analysis is that identical responses indicate a lack of sufficient consideration of the nuances of individual statements being rated. Long string analysis has been shown to be effective in identifying invariant response patterns (Meade & Craig, 2012; Ward & Pond, 2015). In this study, it was operationalized as the longest string of identical responses per aspect (i.e., tasks or knowledge) for each rating scale. The maximum possible long string value is survey-specific and was expected to be longer for lengthier surveys. Regarding cutoff values, the original recommendations for cutoffs for this method, based on use of a 5-point Likert scale, were established based on research in personality assessment (Costa & McCrae, 1997). In the absence of recommended cutoffs for job analysis, a scree approach was used (Johnson, 2005; Steedle, 2018) to identify an appropriate cutoff to categorize respondents as careless or not. In the scree approach, a frequency distribution of long strings is produced and "the last substantial decrease in the distribution before it becomes more uniform" (Steedle, 2018, p.12) is selected as the cutoff.

Job Aspect

The job analysis literature makes a distinction between the work performed, i.e., tasks and responsibilities, and the individual performing the work, i.e., KSAOs[1] (Harvey & Wilson, 2000; Sanchez & Levine, 2012). This distinction was adopted to categorize job analysis survey aspects thematically as either work- or worker-oriented, with tasks the work-oriented aspect and knowledge the worker-oriented aspect. Harvey and Wilson proposed that making worker-oriented knowledge ratings is more difficult than making work-oriented task ratings due to the level of inference required.

Length of Survey

Survey length is the number of items to be rated in a survey multiplied by the number of rating scales used. For example, a survey with 63 items and 2 rating scales has a length of 126. A survey with 85 job items and 3 rating scales has a length of 255.

Mahalanobis Distance

Mahalanobis distance flags outliers in survey ratings based on comparing the overall pattern of ratings to that of other survey takers. It has been used in multiple studies of CI responding (Bowling & Huang, 2018; McKay et al., 2018; Meade & Craig, 2012; Steedle, 2018; Ward et al., 2017). The following equation was used to calculate the value of Mahalanobis distance (De Maesschalck et al., 2000):

---

[1] Note that while KSAOs can all be elements of a worker-oriented job analysis, job analysis studies for certification primarily delineate testable knowledge only.

$$\text{Mahalanobis distance} = MD = \sqrt{(xi - \bar{x})C_x^{-1}(x_i - \bar{x})}$$

Mahalanobis distance is particularly sensitive to skewed data (Meade & Craig, 2012), as is typically found in job analysis studies. It is intended to flag a different type of response pattern than long string analysis, although research is contradictory as to whether the two measures are positively or negatively correlated. While Meade and Craig (2012) and Huang et al. (2016) found the two were moderately to highly positively correlated, McKay et al. (2018) and Ward and Pond (2015) found the two were weakly negatively correlated. There is no universally adopted cutoff for Mahalanobis distance. In this study, I used the square of the Mahalanobis distance value. Mahalanobis distance^2 is distributed as a chi^2 variable (DeSimone et al., 2015) and values exceeding a critical value $\left(MD^2 > \chi_j^2, \alpha\right)$ were flagged (Steedle, 2018).

Person-Total Correlation

This measure identifies how consistently a respondent's answers are to those of all other survey takers. Person-total correlation is an extension of the point-biserial correlation (Donlon & Fischer, 1968) that is used to examine test item performance. It was proposed by Karabatsos (2003) as an index of item difficulty. The person-total correlation is determined by transposing the item by total matrix prior to calculating the correlation coefficient (Curran, 2016). Karabatsos found the person-total correlation to be one of the most useful in detecting random responding. There is little research on the use of as this index as a post-hoc detection method.

Type of Rating Scale

Rating scales were categorized based on their concreteness versus abstractness. Concrete rating scales have a shared meaning regarding their applicability to a job (Harvey & Wilson, 2000). A highly concrete absolute frequency scale was used in each job analyses in this study. The importance and knowledge acquisition scales used in the three job analysis studies require considerably more inference and judgment about the profession (Harvey & Wilson, 2000; Morgeson & Campion, 1997) and were categorized as abstract.

Data Analysis Plan

All analyses were performed using SPSS Version 25. Treatment of missing data was as follows. Prior to analysis, cases with 30% or more missing data were deleted. For all remaining cases, missing values were replaced through multiple imputation (Dong & Peng, 2013; Graham, 2009; Newman, 2014). Because long string analysis compares whole number responses in rating scales, imputed values were rounded to whole numbers before conducting long string analysis. Rounding introduces some lack of numerical precision, but the alternative was to delete all records with partially missing data. The danger of this approach was that if records with partially missing data were missing due to correlations with unobserved variables, deletion could introduce unintended bias into the retained dataset (Graham, 2009).

There were three predictor variables: survey length, job aspect rated, and type of rating scale. All were categorical. Criterion variables related to careless responding

detection were long string analysis, Mahalanobis distance, and person-total correlation. I treated these variables as both continuous and categorical (Krosnick, 1999): continuous when calculating values on the indices and dichotomous when applying cutoffs (Ran, Liu, Marchiondo, & Huang, 2015).

Analyses were conducted separately for each survey. To permit visual comparison of findings across surveys, tables and figures summarizing results were created. To explore research questions 1 through 3, descriptive statistics were produced, and non-parametric tests were conducted to examine the relationship between each predictor variable and extent of carelessness on each index. The following conditions were expected to produce more careless responses: longer surveys, surveys containing knowledge statements rather than task statements, and abstract rather than concrete rating scales. To explore research question 4, correlational and factor analysis were undertaken. Rules for categorizing respondents as careful or careless were created based on guidance from the literature, and the numbers and types of records flagged as careless by each index were calculated.

For research question 5, three aspects of psychometric quality were calculated pre- and post-CI response removal: rating scale reliability, mean task or knowledge ratings, and average item intercorrelations. Reliability reflects consistency among raters in the selection of scale points. High reliability suggests that different survey respondents judged aspects of the job similarly. Rating scale reliability was calculated using the interclass correlation coefficient (Shrout & Fleiss, 1979), a measure commonly used in

job analysis studies (Sanchez & Fraser, 1992; Voskuijl & van Sliedregt, 2002). For tasks, frequency and importance correlations and mean ratings were calculated in each domain. For knowledge, the same calculations were undertaken for frequency and acquisition. The magnitude of differences between the psychometric characteristics of the data pre- and post-CI response removal was tabled and inspected visually.

Finally, to explore research question 6, the relationship between carelessness and test content outlines, the decision rules adopted by the sponsoring organizations were applied to datasets with records containing CI responses removed. Decision rules are thresholds used to identify which tasks and knowledge should be included in a certification test content outline (Raymond, 2001, 2002; Raymond & Luecht, 2013). The statements selected for inclusion before and after removal of CI data were compared for substantive differences.

Research Questions and Hypotheses

The research questions and hypotheses laid out in Chapter 1 are repeated here.

RQ1: Is there a relationship between survey length and extent of carelessness?

$H_0 1$: There is no relationship between survey length and extent of carelessness.

$H_a 1$: There is a relationship between survey length and extent of carelessness.

Satisficing theory predicts a loss of motivation in lengthier surveys (Krosnick, 1996) and an increase in satisficing. There is some evidence to suggest that longer surveys are associated with speeded and straightlined responding (Hardré et al., 2012; Zhang & Conrad, 2014). RQ1 explores the possibility of that carelessness occur more in

longer surveys. It was anticipated that longer surveys would be associated with more CI, particularly in the form of straightlining, which can be detected by long string analysis.

RQ2: Is there a relationship between job aspect and extent of carelessness?

H_02: There is no relationship between job aspect and extent of carelessness.

H_a2: There is a relationship between job aspect and extent of carelessness.

Satisficing theory predicts that more complex ratings will produce greater inaccuracy due to greater cognitive demands. In support of this, job analysis theory (Morgeson & Campion, 1997) and research (Dierdorff & Morgeson, 2009; Morgeson et al., 2004) suggest there will be more CI responding for knowledge ratings than for task ratings.

RQ3: Is there a relationship between rating scale and extent of carelessness?

H_03: There is no relationship between rating scale and extent of carelessness.

H_a3: There is a relationship between rating scale and extent of carelessness.

This analysis was restricted to the task-based surveys, which contained both concrete and abstract rating scales, permitting a direct within-subjects comparison. Based on predictions from the job analysis literature (Harvey & Wilson, 2000), it was expected that importance ratings would be associated more carelessness than frequency ratings.

RQ4: Is each post hoc detection index equally useful for identifying careless responding in job analysis surveys?

H_04: All indices will be equally useful in flagging careless responding in job analysis surveys.

H_a4: All indices will not be equally useful in flagging careless responding in job analysis surveys.

This was an exploratory question since that is the first known application of carelessness research to job analysis data. Based on the purposes of the Mahalanobis distance and long string analyses indices, it was hypothesized that each index would be useful for assessing different response characteristics. It was anticipated that long string analysis would be most useful to detect invariance (Curran, 2016; Steedle, 2018) and that Mahalanobis distance would be most useful to detect outliers (Curran, 2016; Meade & Craig, 2012; Zijlstra et al., 2011). No hypotheses were made regarding the person-total correlation, as its inclusion in this study was exploratory.

RQ5: Is there a relationship between careless responding and the psychometric characteristics of job analysis data?

H_05: There is no relationship between careless responding and the psychometric characteristics of job analysis data.

H_a5: There is a relationship between careless responding and the psychometric characteristics of job analysis data.

Based on prior research (Huang et al., 2015; Meade & Craig, 2012; Zijlstra et al., 2011), it was expected that the presence of careless responses would decrease reliability and attenuate correlations and mean ratings. Note that while removal of careless responses can either increase or decrease reliability (Huang et al., 2015), the latter has only been found with Likert-type scales, which were not used in this study.

RQ6: Are there differences in terms of selection of tasks for a certification test content outline?

$H_0$6: There is no difference in terms of tasks for a certification test content outline?

H_a6: There are differences in terms of tasks selected for a certification test content outline?

Tasks included in certification test content outlines are eligible for assessment if they exceed inclusion thresholds. If any tasks selected for a test content outline change after removing careless responders, this threatens the content validity argument for the certification. If differences are found, the need to screen and eliminate carelessness is strongly indicated for all future studies.

Threats to Validity

Threats to validity can be categorized as internal and external (Campbell & Stanley, 1967). Internal validity threats relate to potential weaknesses in study design that limit the attribution of causality, while external validity threats relate to the ability to generalize the study findings to the larger population (Onwuegbuzie, 2000). Campbell and Stanley identified eight internal validity threats: history, maturation, testing, instrumentation, statistical regression, selection, experimental mortality, and selection-maturation interaction. All are features of experimental design. Because this study employed a non-experimental, cross-sectional research design, the eight factors outlined by Campbell and Stanley did not apply directly (Onwuegbuzie, 2000). However, the

inability to attribute causality was an internal validity threat directly related to the correlational nature of the study. With a correlational design, I could characterize the strength of relationships among variables statistically but could not attribute causality.

The study lacked the strengths associated with experimental designs because there were no experimental manipulations or experimental controls that might mitigate extraneous sources of variation. There is a rich literature on factors influencing job analysis responding, and a growing literature on factors influencing CI responding. Only a subset of each was explored in this study. Lack of experimental controls compromises the ability to isolate and detect true differences in the variables under study. For example, each job analysis study was administered at a different timeframe to different professions under different circumstances. Also, there were slight differences in the wording of rating scales based on the needs of the sponsors. Finally, individual differences that may influence carelessness, such as gender, education and personality, were not controlled.

External validity is the ability to generalize findings across timeframes, locations, settings and entities (Bainbridge, Sanders, Cogin, & Lin, 2017). This study looked at only a small set of job analysis studies conducted for a specific purpose. While one potential strength of the study is that it includes multiple job analysis surveys of different professions, it is important to realize that job analysis for certification programs differs from job analysis within an organization, in that respondents in the former represent practitioners in a variety of settings. Motivation to participate may be greater for

individuals holding certification than for employees within an organization, because certificants have invested already in obtaining the credential and may wish to provide ongoing support of the credentialing program. Higher motivation may produce less careless responding than might occur in a within-organization job analysis.

Ethical Procedures

The job analysis datasets provided by current and former clients were used only for my dissertation; no other use was made of the information. All respondent identifying information was stripped from the datasets prior to analysis. The data were stored on a password-protected computer and only I had access to the data. The datasets used for analysis will be destroyed after the dissertation is approved.

Summary and Transition

This chapter described and operationalized the study variables, as well as the research design and methods to be employed to explore the relationships among them. The secondary data sources used and the methods for obtaining them were described, and the ethical considerations around their use were outlined. The hypotheses to be tested were specified and threats to internal and external validity were discussed.

Chapter 4: Results

Restatement of Research Questions

The purpose of this study was to explore the extent to which careless responding occurs in job analysis surveys, the relationship between carelessness and job analysis design features, and the consequences of carelessness relative to the psychometric properties of job analysis ratings and decisions made based on job analysis data. In this chapter, I describe analyses undertaken to address the following six research questions:

RQ1: Is there a relationship between survey length and extent of carelessness?

RQ2: Is there a relationship between job aspect rated and extent of carelessness?

RQ3: Is there a relationship between rating scale used and extent of carelessness?

RQ4: Is each post hoc detection index equally useful for identifying careless responding in job analysis surveys?

RQ5: Is there a relationship between careless responding and the psychometric characteristics of job analysis data?

RQ6: Are there differences in terms of selection of tasks a certification test content outline?

The first three research questions explored correlates of carelessness, the fourth explored extent of carelessness, and the fifth and sixth explored the consequences of carelessness.

Tests for Assumptions

The SPSS Explore procedure was run to analyze properties of the CI variables. Visual inspection of histograms and Q-Q plots suggested the data were non-normal. This was confirmed by running the Shapiro-Wilks test. All six task indices and five knowledge indices were non-normally distributed (see Table 2).

Table 2

Shapiro-Wilk Tests of Normality

CI Index	Literacy Coach	Patient Care Technician	Pharmacy Technician
Tasks			
Mahalanobis distance–Frequency	.961**	.988**	.955**
Mahalanobis distance–Importance	.976**	.883**	.873**
Long string–Frequency	.608**	.772**	.714**
Long string–Importance	.774**	.856**	.835**
Person-total correlation–Frequency	.927**	.932**	.905**
Person-total correlation–Importance	.964**	.922**	.920**
Knowledge			
Mahalanobis distance–Frequency	.961**	.988**	.955**
Long string–Frequency	.672**	.766**	.596**
Long string–Acquisition	.802**	.903**	.856**
Person-total correlation–Frequency	.924**	.958**	.894**
Person-total correlation–Acquisition	.964**	.970**	.992*

*p < .05
**p < .01

Inspection of boxplots for CI index data revealed a large number of outliers. The outliers could not be removed as they may represent instances of careless responding. Therefore, in later analyses Spearman's rank order correlation coefficients were calculated instead of Pearson's correlations. Spearman's correlations are less sensitive to outliers than Pearson's correlation (de Winter, Gosling, & Potter, 2016) and are appropriate for ordinal or higher levels of measurement (de Winter et al., 2016).

Characteristics of Datasets

All analyses were undertaken on three job analysis datasets. Each job analysis had two versions, and each respondent was randomly routed to one. One version contained task statements and the other contained knowledge statements. The occupations, and the number of tasks and knowledge statements, the rating scales used, and number of respondents to each version are presented in Table 3.

Table 3

Characteristics of Job Analysis Datasets

Occupation	Job Aspect Rated	Number of Elements	Length Category	Rating Scales	N Respondents
Literacy coach					
Version A	Tasks	58	Short	Frequency and Importance	406
Version B	Knowledge	54	Short	Frequency and Acquisition	401
Patient care technician					
Version A	Tasks	86	Medium	Frequency and Importance	344
Version B	Knowledge	115	Medium	Frequency and Acquisition	390
Pharmacy technician					
Version A	Tasks	96	Long	Frequency and Importance	513
Version B	Knowledge	170	Long	Frequency and Acquisition	429

Results

Research Question 1

RQ1: Is there a relationship between survey length and extent of carelessness?

The question concerns whether surveys with different numbers of elements rated exhibit different levels of carelessness on the three indices.

Task-based surveys. A Kruskal-Wallace *H* test was conducted to examine for differences among three survey lengths (short, medium, and long) and median values for

the six CI indices. There were statistically significant differences for all six CI values: Mahalanobis distance - frequency ($H = 188.07$, $p < .001$), Mahalanobis distance - importance ($H = 7.33$, $p < .05$), person-total correlation - frequency ($H = 22.94$, $p < .001$), person-total correlation - importance ($H = 164.90$, $p < .001$), long string - frequency ($H = 412.69$, $p < .001$), and long string - importance ($H = 293.55$, $p < .001$).

Given these significant results, follow-up pairwise comparisons were undertaken using the non-parametric Mann-Whitney *U* test. The Bonferroni correction was made to control for Type I error across the comparisons. Descriptive statistics for the CI indices at each task survey length and the results of the pairwise comparisons are shown in Table 4.

For Mahalanobis distance–frequency, median values were higher for medium length surveys than for short surveys for both rating scales and were higher for long than for short surveys for the importance rating scale. Differences between medium and long length surveys were in the expected direction but were not significant. For long string analysis, all pairwise comparisons were significant, with larger CI index values at longer survey lengths for both frequency and importance scales. For the person-total correlation, the lowest index values were associated with the medium length value.

Table 4

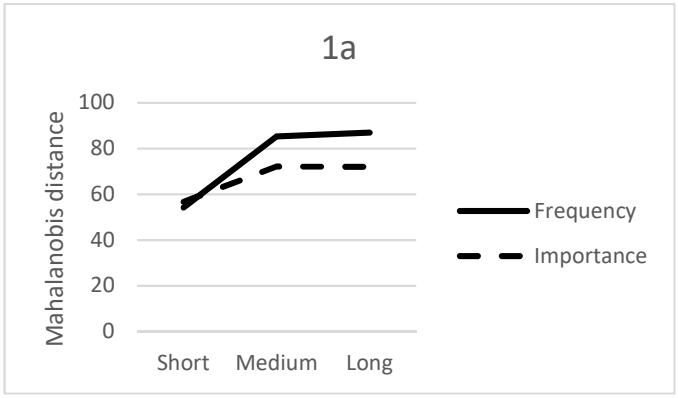
CI Index Descriptive Statistics for Task Ratings by Survey Length

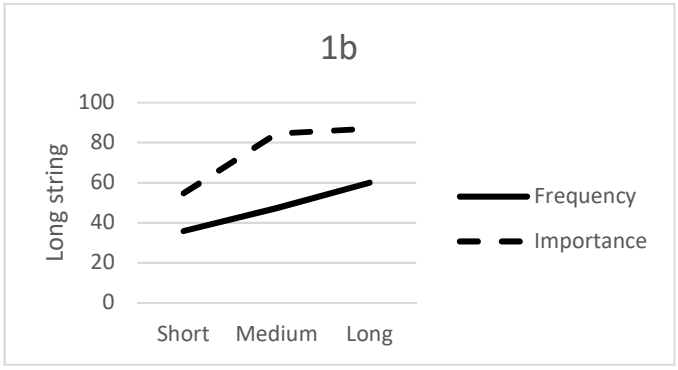
Survey Length	Mean	SD	25th Percentile	Median	75th Percentile	Mann-Whitney *U* Test	Directionality of Significant Differences
Mahalanobis distance–Frequency							
Short	57.86	22.67	41.29	54.22	72.90	-291.42**	S<M
Medium	87.12	37.22	59.47	85.37	110.04	-307.51**	S<L
Long	95.81	54.11	56.62	86.98	124.95	-0.63	
Mahalanobis distance–Importance							
Short	57.86	33.14	32.89	56.74	78.11	-23.44	
Medium	87.23	82.18	1.09	72.15	153.99	-64.23*	S<L
Long	95.81	96.53	1.10	71.97	166.52	-40.79	
Long string–Frequency							
Short	7.94	5.75	5.00	7.00	9.00	-316.38**	S<M
Medium	16.42	13.97	8.00	12.00	21.00	-489.46**	S<L
Long	23.41	20.40	12.00	15.00	26.00	-173.08**	M<L
Long string–Importance							
Short	17.36	14.04	8.00	12.00	22.00	-298.16**	S<M
Medium	45.53	31.71	13.50	41.00	86.00	-408.00**	S<L
Long	52.51	34.78	21.00	43.00	96.00	-109.82**	M<L
Person-total correlation–Frequency							
Short	.50	.19	.41	.53	.63	55.68	
Medium	.44	.25	.27	.49	.63	-64.24*	S<L
Long	.50	.28	.26	.60	.74	-119.91**	M<L
Person-total correlation–Importance							
Short	.36	.21	.22	.39	.52	324.81**	S>M
Medium	.16	.20	.00	.11	.32	232.15**	S>L
Long	.22	.20	.00	.23	.38	-92.71**	M<L

Note. Order of presentation of *U* test paired comparisons down column is short versus medium, short versus long, and medium versus long survey length.
*p < .05.
**p < .01.

Figure 1 illustrates the median task CI index values at different survey lengths. Mahalanobis distance values were higher the more statements there were to rate but only for short and medium surveys, which is the finding depicted in Figure 1a. Long string values were consistently higher with longer survey lengths, as shown in Figure 1b.

Person-total correlation values were not higher at longer survey lengths, as can be seen in see Figure 1c.





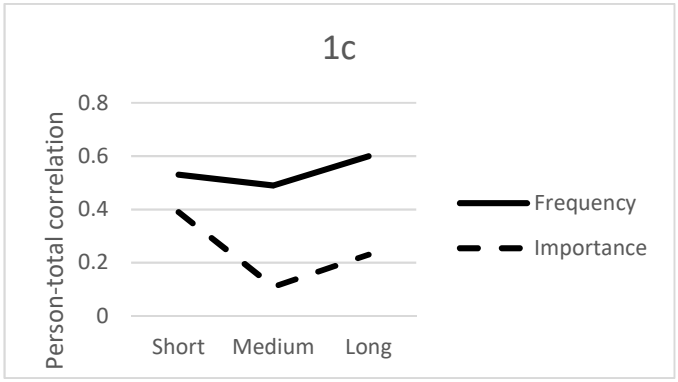


Figure 1. Median task CI index values for frequency and importance ratings by survey length. Figure 1a displays Mahalanobis distance values, 1b displays long string values, and 1c displays person-total correlations.

Knowledge-based surveys. Similar analyses were undertaken to explore the relationship between the number of knowledge statements rated across all three job analysis surveys and the values for five CI indices. Mahalanobis distance could not be computed for knowledge acquisition because it was a nominal variable. There were statistically significant differences for all five CI indices: Mahalanobis distance–frequency(K-W=582.29), person-total correlation–frequency (K-W=303.14), person-total correlation–acquisition (K-W = 451.92), long string–frequency (K-W=301.78), and long string–acquisition (K-W=451.92), with $p < .001$ for all analyses.

Following up on these significant results, pairwise comparisons were undertaken using the Mann-Whitney *U* test. Bonferroni corrections were employed to correct for Type 1 error. Descriptive statistics for the CI indices at each knowledge survey length and the results of the paired comparisons are displayed in Table 5.

There were significant differences for all pairwise comparisons for Mahalanobis distance and person-total correlation. For Mahalanobis distance, median CI values were higher at longer survey lengths. Long string values were higher for medium than short length surveys and higher for long than medium length surveys. Long string–frequency values were highest for the medium length surveys. For person-total correlation, all paired comparisons were significant, but the values did not increase systematically with longer survey lengths.

Table 5

CI Index Descriptive Statistics for Knowledge Ratings by Survey Length

Survey Length	Mean	SD	25th Percentile	Median	75th Percentile	Mann-Whitney U Test	Directionality of Significant Differences
Mahalanobis distance–Frequency							
Short	53.87	23.64	37.15	49.59	66.03	-13.03*	S<M
Medium	114.71	61.65	69.87	117.66	158.20	-24.11**	S<L
Long	169.60	55.74	139.10	171.82	205.22	-10.69**	M<L
Long string–Frequency							
Short	9.15	7.01	5.00	7.00	10.00	-8.18**	S<M
Medium	28.48	29.12	9.00	19.00	37.00	-16.77**	S<L
Long	25.93	29.66	11.00	16.00	27.00	-8.33**	M>L
Long string–Acquisition							
Short	14.80	9.62	8.00	12.00	17.00	-447.02**	S<M
Medium	56.82	37.90	24.00	51.00	93.00	-464.39**	S<L
Long	60.62	47.46	23.00	44.00	87.00	-17.37	
Person-total correlation–Frequency							
Short	.49	.19	.40	.52	.62	12.56**	S>M
Medium	.29	.20	.13	.33	.44	-16.82**	S>L
Long	.54	.22	.41	.61	.70	-4.06**	M<L
Person-total correlation–Acquisition							
Short	.25	.22	.14	.26	.42	16.16**	S>M
Medium	.03	.12	-.03	.00	.10	8.06**	S<L
Long	.14	.18	.00	.14	.26	-8.42**	M<L

Note. In the *Directionality of Significant Differences* column, the order of presentation of comparisons down the column is short versus medium, short versus long, and medium versus long.
*$p < .05$
**$p < .01$

Figure 2 illustrates the median knowledge CI index values at different survey lengths. Mahalanobis distance was larger at longer survey lengths, as depicted in Figure 2a. Long string analysis (Figure 2b) and person-total correlation (Figure 2c) values showed no systematic increase at longer survey lengths. Each occupation showed a different pattern of results. Literacy coaches (short survey) had less invariance in responses based on long string analysis and more inconsistency based on person-total

correlation. Patient care technicians (medium survey) and pharmacy technicians (long survey) exhibited the opposite pattern: more invariance and less inconsistency.

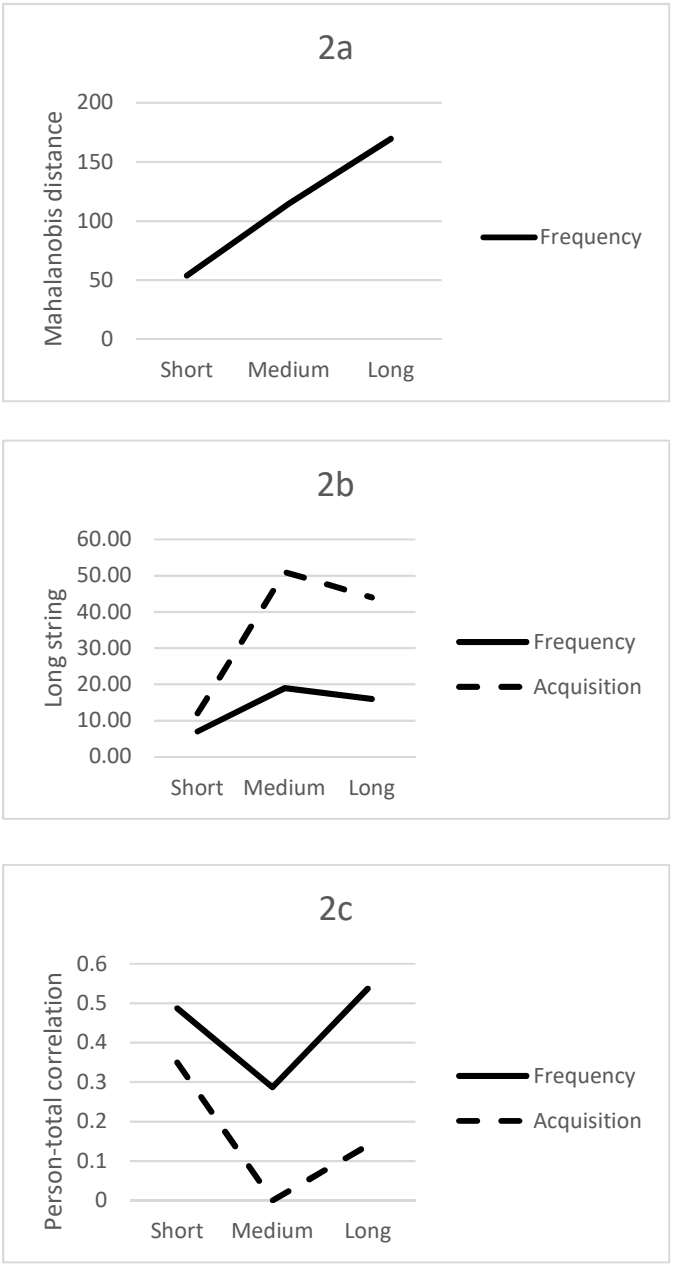


Figure 2. Median knowledge CI index values for frequency and acquisition by survey length. Figure 2a displays Mahalanobis distance values, 2b displays long string values, and 2c displays person-total correlations.

Research Question 2

RQ2: Is there a relationship between job aspect rated and extent of carelessness?

This within-subjects analysis was conducted on frequency ratings for task and knowledge survey versions. Mann-Whitney U tests were conducted to compare median CI index values for Mahalanobis distance–frequency, person-total correlation–frequency and long string–frequency across survey versions. Descriptive statistics and the results of the comparisons between versions are displayed in Table 6.

Table 6

CI Index Descriptive Statistics for Task and Knowledge Frequency Ratings

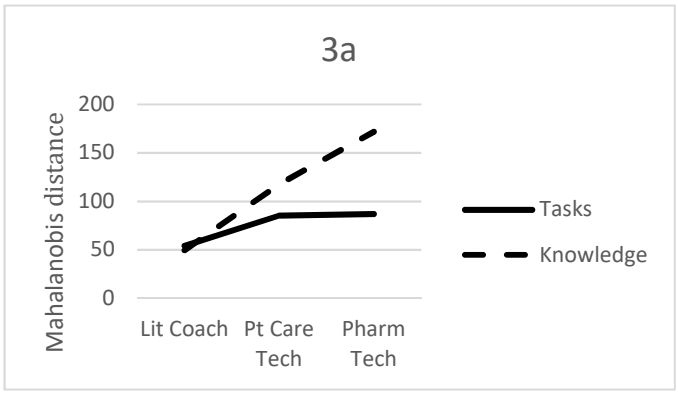
CI Index	Job Aspect	Mean	SD	25th Percentile	Median	75th Percentile	Mann-Whitney *U* Test	Directionality of Significant Differences
Mahalanobis Distance								
Lit Coach	Tasks	57.86	22.67	41.29	54.22	72.90	-2.84 **	T>K
	Knowledge	53.87	23.64	37.15	49.59	66.03		
Pt Care Tech	Tasks	87.12	37.22	59.47	85.37	110.04	6.63**	K>T
	Knowledge	114.71	61.65	69.87	117.66	158.20		
Pharm Tech	Tasks	95.81	54.11	56.62	86.98	124.95	17.56**	K>T
	Knowledge	169.60	55.74	139.10	171.82	205.22		
Long String Value								
Lit Coach	Tasks	7.94	5.75	5.00	7.00	9.00	2.74**	K>T
	Knowledge	9.15	7.01	5.00	7.00	10.00		
Pt Care Tech	Tasks	16.42	13.97	8.00	12.00	21.00	5.17**	K>T
	Knowledge	28.48	29.12	9.00	19.00	37.00		
Pharm Tech	Tasks	23.41	20.40	12.00	15.00	26.00	-0.40	
	Knowledge	25.93	29.66	11.00	16.00	27.00		
Person Total Correlation								
Lit Coach	Tasks	.50	.19	.41	.53	.63	-1.08	
	Knowledge	.49	.19	.40	.52	.62		
Pt Care Tech	Tasks	.44	.25	.27	.49	.63	-9.50**	T>K
	Knowledge	.29	.20	.13	.33	.44		
Pharm Tech	Tasks	.50	.28	.26	.60	.74	0.27	
	Knowledge	.54	.22	.41	.61	.70		

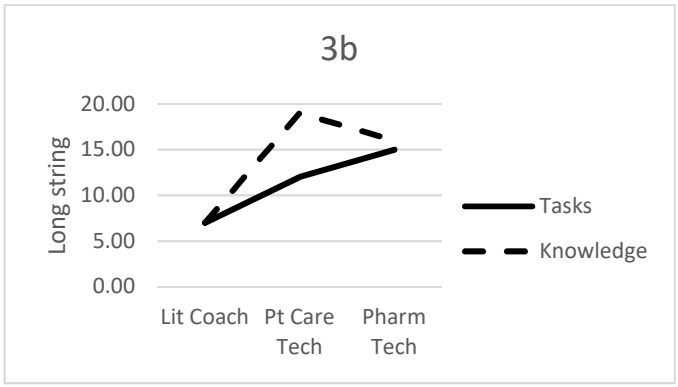
Note. Lit Coach = Literacy coach; Pt Care Tech = Patient care technician; Pharm tech=Pharmacy technician.

***p < .01*

Index values were expected to be higher for knowledge-based survey ratings than task-based survey ratings. For Mahalanobis distance, index values were higher for knowledge than tasks for patient care technicians and pharmacy technicians but not for literacy coaches. For the latter, the difference was in the opposite direction, with task CI values higher than knowledge CI values. For long string analysis, index values were significantly higher for knowledge than tasks for literacy coaches and patient care technicians but not pharmacy technicians. Finally, person-total correlations for tasks and knowledge were nearly identical for literacy coaches and pharmacy technicians but were higher for tasks than knowledge for the patient care technicians.

Median careless ratings for the two scales are displayed in Figure 3. Looking across Figures 3a, 3b, and 3c, minimal differences were found in the task and knowledge frequency ratings for literacy coaches. For patient care technicians, Mahalanobis distance (3a) and person-total correlation (3c) index values were higher for tasks, but long string index values (3b) were higher for knowledge. For pharmacy technicians, Mahalanobis distance was higher for knowledge than for tasks (3a); however, there were no substantive differences in long string (3b) or person-total correlation (3c) values. Based on long string and person-total correlation values, patient care technicians exhibited more invariance and less inconsistency in their ratings.





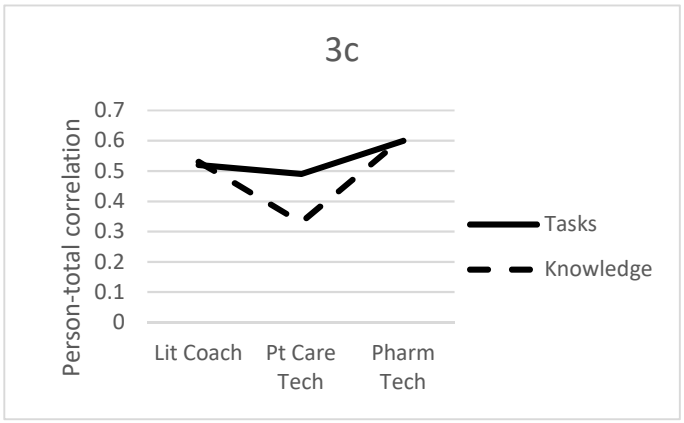


Figure 3. Median CI index values for task and knowledge frequency ratings by survey length. Figure 3a displays Mahalanobis distance values, 3b displays long string values, and 3c displays person-total correlations.

Research Question 3

RQ3: Is there a relationship between rating scale and extent of carelessness?

This analysis was conducted for both task-based and knowledge-based survey versions, as each contained both concrete and abstract rating scales.

Task-based surveys. Wilcoxon Signed Ranks tests (Rey & Neuhäuser, 2011) were conducted between frequency and importance index pairs for each occupation. The Wilcoxon is a non-parametric alternative to the t-test used when the assumptions of the t-test are not met. It was selected because preliminary data exploration revealed that the assumption of normality was not met for any CI indices. Results of the analysis are shown in Table 7.

There were no significant differences for the Mahalanobis distance indices. Long string and person-total correlation signed ranks were significantly different for all three surveys. For long strings, carelessness values were higher for the importance scale than the frequency scale. For person-total correlation, the opposite result was found.

Table 7

CI Index Descriptive Statistics for Task Ratings by Occupation

CI Index	Rating Scale	Mean	SD	25th Percentile	Median	75th Percentile	Wilcoxon Signed Rank Test	Directionality of Significant Differences
Mahalanobis Distance								
Lit Coach	Frequency	57.86	22.67	41.29	54.22	72.90	0.37	
	Importance	57.86	33.14	32.89	56.74	78.11		
Pt Care Tech	Frequency	87.12	37.22	59.47	85.37	110.04	0.15	
	Importance	87.23	82.18	1.09	72.15	153.99		
Pharm Tech	Frequency	95.81	54.11	56.62	86.98	124.95	-0.28	
	Importance	95.81	96.53	1.10	71.97	166.52		

(table continued)

Table 7, continued

CI Index	Rating Scale	Mean	SD	25th Percentile	Median	75th Percentile	Wilcoxon Signed Rank Test	Directionality of Significant Differences
Long String Value								
Lit Coach	Frequency	7.94	5.75	5.00	7.00	9.00	14.78**	F < I
	Importance	17.36	14.04	8.00	12.00	22.00		
Pt Care Tech	Frequency	16.42	13.97	8.00	12.00	21.00	12.01**	F < I
	Importance	45.53	31.71	13.50	41.00	86.00		
Pharm Tech	Frequency	23.41	20.40	12.00	15.00	26.00	16.58**	F < I
	Importance	52.51	34.78	21.00	43.00	96.00		
Person-Total Correlation								
Lit Coach	Frequency	.50	.19	.41	.53	.63	-11.49**	F > I
	Importance	.36	.21	.22	.39	.52		
Pt Care Tech	Frequency	.44	.25	.27	.49	.63	-13.87**	F > I
	Importance	.16	.20	.00	.11	.32		
Pharm Tech	Frequency	.50	.28	.26	.60	.74	-15.73**	F > I
	Importance	.22	.20	.00	.23	.38		

*** p < .01.*

Median ranks for tasks and knowledge for each occupation are displayed in Figure 4. Visually, difference in CI values for the frequency and importance rating scales are apparent, as are differences between occupations. Within each occupation, Mahalanobis distance index values (4a) were similar for frequency and importance and were lower for literacy coaches than for the other two occupations. Large differences were observed in long string analysis (4b) for patient care technicians and pharmacy technicians, with smaller differences for literacy coaches. Person total correlations (4c) values were lower for importance than frequency. For patient care technicians and pharmacy technicians, frequency ratings had high inconsistency and low invariance and importance ratings had the opposite: higher invariance and low inconsistency.

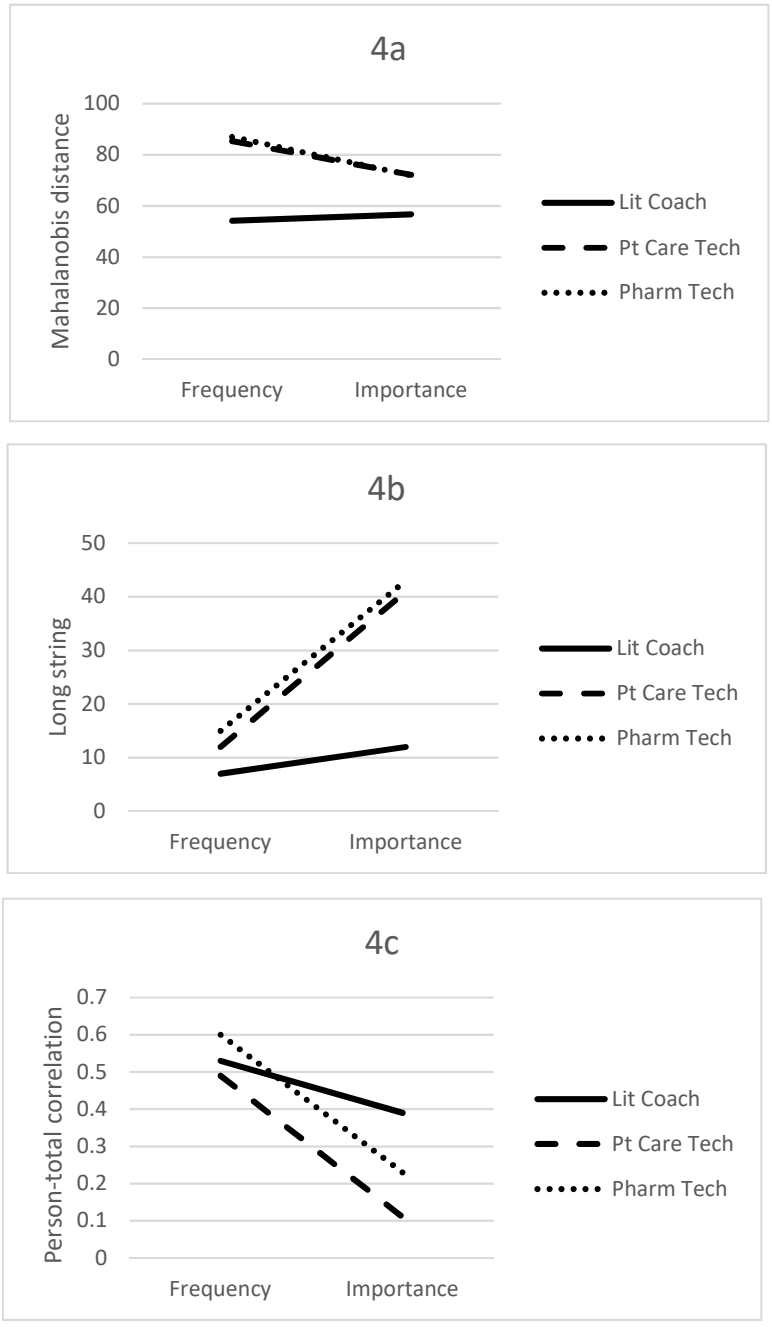


Figure 4. Median CI index values for task frequency and importance ratings by occupation surveyed. Figure 4a displays Mahalanobis distance values, 4b displays long string values, and 4c displays person-total correlations.

Knowledge-based surveys. As with task-based surveys, Wilcoxon Signed Ranks tests were conducted between long string–frequency and long string–acquisition ratings and between the person-total correlation–frequency and person-total correlation–acquisition values for each occupation's knowledge surveys. There were no paired data values for Mahalanobis distance. Results are displayed in Table 8. All three tests yielded significant differences associated with the index values. For long strings, acquisition carelessness values were higher than frequency carelessness values. In contrast, for person-total correlation, frequency carelessness values were greater than the acquisition values.

Table 8

CI Index Descriptive Statistics for Knowledge Ratings by Occupation

CI Index	Rating Scale	Mean	SD	25th Percentile	Median	75th Percentile	Wilcoxon Signed Rank	Directionality of Significant Differences
Long String Value								
Lit Coach	Frequency	9.15	7.01	5.00	7.00	10.00	11.29**	F < A
	Acquisition	14.80	9.62	8.00	12.00	17.00		
Pt Care Tech	Frequency	28.48	29.12	9.00	19.00	37.00	12.98**	F < A
	Acquisition	56.82	37.90	24.00	51.00	93.00		
Pharm Tech	Frequency	25.93	29.66	11.00	16.00	27.00	14.91**	F < A
	Acquisition	60.62	47.46	23.00	44.00	87.00		
Person-Total Correlation								
Lit Coach	Frequency	.49	.19	.40	.52	.62	-14.26**	F > A
	Acquisition	.25	.22	.14	.26	.42		
Pt Care Tech	Frequency	.29	.20	.13	.33	.44	-14.99**	F > A
	Acquisition	.03	.12	-.03	.00	.10		
Pharm Tech	Frequency	.54	.22	.41	.61	.70	-17.15**	F > A
	Acquisition	.14	.18	.00	.14	.26		

*** p < .01.*

Median differences in index values for the two scales are displayed in Figure 5. Long string values were consistently higher for acquisition than frequency and larger for patient care technicians and pharmacy technicians than literacy coaches. Person-total correlation values were values were higher for frequency than acquisition. For acquisition, all occupations demonstrated less invariance and more inconsistency in their frequency ratings, and less inconsistency and more invariance in their acquisition ratings.

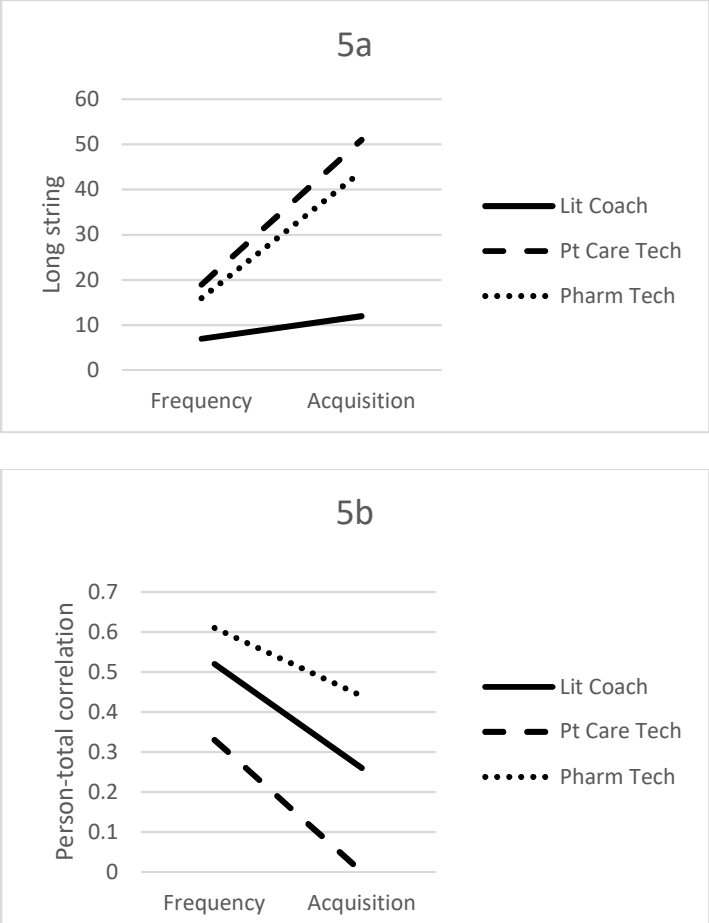


Figure 5. Median CI index values for knowledge frequency and acquisition scales. Figure 5a displays Mahalanobis distance values, 4b displays long string values, and 4c displays person-total correlation.

Research Question 4

RQ4: Is each post hoc detection index equally useful for identifying careless responding in job analysis surveys?

This question concerned how the carelessness indices related to each other, and the number and types of responses flagged by each method. As a first step in exploring this question, correlations coefficients were calculated among all CI indices for each survey length and each survey version. Because the long string value is dependent on the number of items in the survey, prior to calculating correlation coefficients to be compared across surveys, long string values were converted to z-scores to place them on a common metric. All instances in which z-scores were used in analysis are indicated with the labels z-long string–frequency and z-long string–importance.

Correlations among CI indices. Spearman's correlation coefficients were calculated among CI indices. First, correlations were calculated separately for each survey. Positive correlations for paired CI indices (e.g., long string–frequency and long string–importance) in the same survey would suggest the indices capture the same type of responding regardless of rating scale. Negative correlations for cross-index comparisons (e.g., long string–frequency versus Mahalanobis distance–frequency) would be expected if the indices were detecting different types of careless responding. Second, to examine the relationship among the CI indices across all surveys, partial correlations were calculated controlling for occupation/survey length. In all analyses, effect sizes were

evaluated using Cohen's recommendations of .10 indicating a small effect, .30 indicating a medium effect, and .50 indicating a large effect (Cohen, 1992).

Task Rating Scales. Spearman correlation coefficients for the carelessness indices for the task-based survey versions are presented in Table 9. Regarding within-index correlations, Mahalanobis distance values for frequency and importance rating scales were positively correlated for all three surveys, with a medium effect for literacy coaches (r_s=.405, $p < .01$) and a small effect for patient care technicians (r_s =.296, $p <$ .01) and pharmacy technicians (r_s =.241, $p < .01$). Long string values for frequency and importance were positively correlated for literacy coaches and pharmacy technicians, with a large-sized positive correlation for pharmacy technicians (r_s =.513, $p <.01$) and a medium-sized positive correlation for literacy coaches (r_s =.397, $p < .01$), but were uncorrelated for patient care technicians. Finally, person-total correlations for frequency and importance were positively correlated for all three surveys, with a moderate effect size for literacy coaches (r_s = .393, $p < .01$) and patient care technicians (r_s =.269, $p <.01$) and a small effect size for pharmacy technicians (r_s =.256, $p <.01$). In summary, eight of nine within-index correlations were significant. Of the 36 cross-index correlations, 15 were significant.

Table 9

Spearman's Correlation Coefficients for Task CI Measures

CI Index	MD Frequency	MD Importance	z-LS Frequency	z-LS Importance	PTC Frequency
Literacy Coach					
MD Frequency					
MD Importance	.405**				
z-LS Frequency	-.484**	-.304**			
z-LS Importance	-.302**	-.787**	.397**		
PTC Frequency	.144**	.148**	-.089	-.118*	
PTC Importance	-.008	.003	-.022	-.054	.393**
Patient Care Technician					
MD Frequency					
MD Importance	.296**				
z-LS Frequency	-.475**	-.462**			
z-LS Importance	-.086	.000	-.045		
PTC Frequency	.104*	-.259**	-.039	.011	
PTC Importance	.165**	.401**	-.116*	-.006	.269**
Pharmacy Technician					
MD Frequency					
MD Importance	.241**				
z-LS Frequency	-.441**	-.426**			
z-LS Importance	-.215**	-.906**	.513**		
PTC Frequency	-.164**	.112*	-.342**	-.145**	
PTC Importance	.167**	.646**	-.450**	-.744**	.256**

Note. MD = Mahalanobis distance, z-long string = z-score of long string value; PTC = person total correlation.

*$p < .05$.

**p < .01

Knowledge Rating Scales. Spearman's correlations coefficients between CI indices for each occupation/survey length for the knowledge-based survey versions are presented in Table 10. Regarding within-index comparisons, long string z-scores for frequency and acquisition ratings were positively correlated across all three surveys, with a small effect size for literacy coaches and a medium effect size for patient care technicians and pharmacy technicians. Person-total correlation values for frequency and

acquisition were positively correlated for literacy coaches and pharmacy technicians. Of the 24 cross-index correlations, 15 were significant were significant.

Table 10

Spearman's Correlation Coefficients for Knowledge CI Measures

CI Index	MD Frequency	z-LS Frequency	z-LS Acquisition	PTC frequency
Literacy Coach				
MD Frequency				
z-LS Frequency	-.538**			
z-LS Acquisition	-.102*	.240**		
PTC Frequency	-.075	-.197**	-.041	
PTC Acquisition	-.120*	-.060	-.036	.220**
Patient Care Technician				
MD Frequency				
z-LS Frequency	-.663**			
z-LS Acquisition	-.450**	.479**		
PTC Frequency	.267**	-.156**	.021	
PTC Acquisition	-.007	-.036	-.226**	.045
Pharmacy Technician				
MD Frequency				
z-LS Frequency	-.362**			
z-LS Acquisition	-.308**	.411**		
PTC Frequency	.104*	-.405**	-.265**	
PTC Acquisition	.077	-.178**	-.200**	.139**

*p < .05
**p < .01

Careless responding patterns. Most of the correlational results suggest the different CI indices flag different types of carelessness. Using the literacy coach survey as an example, sample rating strings flagged by the different indices are displayed in Table 11. The data patterns flagged by Mahalanobis distance showed that many response options were selected in both task- and knowledge-based survey versions. The patterns flagged by person-total correlation index are difficult to describe in isolation, as the value represents the difference between the respondent's entire set of responses and the sets of

responses of all other respondents. Finally, as expected, long string analysis flagged response strings that were either invariant or nearly invariant.

Table 11

Sample Response Strings Flagged by Each Method

CI Index	Response Pattern
MD Frequency	1144444453223323323132121133332112155545553543535 23
MD Importance	434332224333331233423212234322211213333111312111223223433 2
LS Frequency	424444444444444444444434344444444444444444544444444444444
LS Importance	42343444444444444444444444444444444444444444444444444444
LS Acquisition	233333333333333333333333333333333333333333333333333333
PTC Frequency	435433444443343333334434325543333434443222333322222
PTC Importance	424444444432333333234332233323243344443334443333334443334
PTC Acquisition	222222232232233223223233211223223111233322123322122222

Exploratory factor analysis. An exploratory factor analysis of the six indices was conducted for the task-based survey versions to further examine the relationships among the CI indices for tasks. Examination of diagnostic statistics indicated the factor analysis met required assumptions. The Kaiser-Meyer-Olkin measure of sampling adequacy was .560 which is adequate for analysis purposes (Field, 2013, p. 684). Bartlett's test of sphericity, which evaluates whether the correlation matric has non-zero off-diagonal components, was significant; however, this is true for any large sample size (Field, 2013, p 685). The determinant of the correlation matric was .340, indicating lack of singularity in the matric. Inspection of the correlation coefficients revealed that no correlation exceeded .445, indicating the absence of multicollinearity.

Principal factor analysis with Varimax rotation with Kaiser normalization was used. A two-factor solution was yielded, with the eigenvalue for the first factor of 2.16

and the eigenvalue for the second factor of 1.34. Factor loadings are displayed in Table 12. The rotated loadings accounted for 37.8% of the variance. Factor 1 loaded positively on the person-total correlation indices and negatively on the long string indices. Factor 2 loaded positively on the Mahalanobis distance indices. These results support distinctions between Mahalanobis distance and long string analysis, and between long string and person-total correlation.

Table 12

Rotated Factor Loadings and Factor Correlations for Task CI Indices

Variable	Factor 1	Factor 2
MD Frequency		**.608**
MD Importance	.268	**.588**
z-LS Frequency	**-.517**	-.314
z-LS Importance	**-.642**	-.216
PTC Frequency	**.394**	-.154
PTC Importance	**.683**	

Note. Bolding indicates largest loading for each variable

Factor analysis was also conducted for the five knowledge-based carelessness indices. Examination of diagnostic statistics indicated the factor analysis met required assumptions. The Kaiser-Meyer-Olkin Measure of sampling adequacy was .559 which is adequate for analysis purposes (Field, 2013, p. 684). Bartlett's test of sphericity, which evaluates whether the correlation matric has non-zero off-diagonal components, was significant (Field, 2013, p. 685). However, the determinant of the correlation matric was .476, indicating lack of singularity in the matric. Inspection of the correlation coefficients revealed that no correlation exceeded .460, indicating the absence of multicollinearity.

A two-factor solution was yielded, with the eigenvalue for the first factor of 1.93 and for the second factor of 1.26. Factor loadings are displayed in Table 13. Cumulatively, the rotated loadings accounted for 42.4% of the variance. Factor 1 loaded negatively on the person-total correlation indices and positively on the long string indices. Factor 2 loaded positively on Mahalanobis distance–frequency and negatively on long string–frequency. As was the case for tasks, these results support the correlational findings of distinctions between Mahalanobis distance and long string analysis, and clearly elucidate a distinction between person-total correlation and long string analysis.

Table 13

Rotated Factor Loadings and Factor Correlations for Knowledge CI Indices

Variable	Factor 1	Factor 2
MD Frequency		**.733**
z-LS Frequency	**.705**	**-.407**
z-LS Acquisition	**.638**	
PTC Frequency	**-.409**	.264
PTC Acquisition	**-.483**	-.163

Note. Bolding indicates the largest factor loadings for each variable.

Initial decision rules. In all analyses conducted to this point, the carelessness indices were treated as continuous variables. In the next set of analyses, the variables were recategorized into nominal variables with two values: careful and careless. In order to categorize responses, cut points for each index were established rationally based on the recent recommendations proposed by Curran (2016), considerations based on the results of their application, and other recommendations from the literature. Setting cut points

requires substantial judgment (Meade & Craig, 2012) due to the "lack of well-defined or empirically justified cutoff values for the various screening techniques" (DeSimone et al., 2015, p. 179). First, as suggested by Steedle (2018), I examined scree plots for all indices to search for logical break points; however, the plots exhibited no clear demarcations. I moved next to examine Curran's suggested cutoffs, which are: (a) significant Mahalanobis distance values at $p < .05$, (b) negative person-total correlations, and (c) long strings equal to or greater than 50% of the items rated. Curran's recommended cutoff for Mahalanobis distance flags a greater number of values than does the more common practice of flagging values significant at $p < .001$ (Tabachnick & Fidell, 2007). In this study, applications Curran's recommended cut point for Mahalanobis yielded extremely high percentages of responses flagged as carelessness for the three surveys. Therefore, I adopted the more conservative approach of Tabachnick and Fidell, using a cutoff for Mahalanobis distance of $p < .001$. For person-total correlation, I also adopted Curran's recommendation of flagging negative person-total correlation values as careless, as it is the only recommendation that exists in the literature. Finally, for long string analysis, a survey dependent rating, I adopted a length value corresponding to 75% of items rated. Other recommendations in the literature based on Likert scale data did not work well with the job analysis data, in that they over-selected values as careless. Table 14 summarizes the original decision rules applied to categorize responses as careful or careless.

Table 14

Initial Decision Rules for Flagging CI Values as Careful or Careless

Index	Cut Point
Mahalanobis distance	All values significant at $p < .001$
Long string	All values equal to or greater than 75% of statements rated
Person-total correlation	All values less than 0.00

Table 15 contains the number and percentage of records flagged by each carelessness index for each survey and The highest percentage of knowledge survey records was flagged based on person-total correlation–acquisition and the lowest percentage based on long string–frequency. As was the case with task surveys, patient care technicians and pharmacy technicians had higher percentage rates of flagging than literacy coaches, although the magnitude of the differences was less for knowledge surveys than for task surveys. These findings indicate that the application of a consistent set of rules led to different extents of flagging both within surveys (i.e., between index type) and between professions.

Table 16 contains the number and percentage of knowledge statements flagged. In eight of nine comparisons for the task-based surveys, importance ratings were flagged at a higher rate than frequency ratings. In all six comparisons for the knowledge-based surveys, acquisition ratings were flagged at a higher rate than frequency ratings. Higher percentages of knowledge survey records were flagged for patient care technicians and pharmacy technicians than for literacy coaches.

The highest percentages of task survey responses were flagged based on Mahalanobis distance–importance and long string–importance, and lowest percentage based on long string–frequency. For literacy coaches and patient care technicians, the lowest percentage of task survey records was flagged by long string–frequency. For pharmacy technicians, the lowest percentage was flagged based on person-total correlation–importance. Comparing across occupations in the task surveys, patient care technicians and pharmacy technicians had much higher rates of flagging than literacy coaches.

Table 15

Number and Percentage of Task Survey Records Flagged by CI Index

	MD Frequency		MD Importance		LS Frequency		LS Importance		PTC Frequency		PTC Importance	
Survey	N	%	N	%	N	%	N	%	N	%	N	%
Lit Coach	25	6.2	52	12.8	2	0.5	34	8.4	9	2.2	21	5.2
Pt Care Tech	55	16.0	117	34.0	7	2.0	127	36.1	25	7.3	40	11.6
Pharm Tech	97	18.9	159	31.0	30	5.8	193	37.6	31	6.0	19	3.7

The highest percentage of knowledge survey records was flagged based on person-total correlation–acquisition and the lowest percentage based on long string–frequency. As was the case with task surveys, patient care technicians and pharmacy technicians had higher percentage rates of flagging than literacy coaches, although the magnitude of the differences was less for knowledge surveys than for task surveys. These findings indicate that the application of a consistent set of rules led to different extents of flagging both within surveys (i.e., between index type) and between professions.

Table 16

Number and Percentage of Knowledge Survey Records Flagged by CI Index

	MD Frequency		LS Frequency		LS Acquisition		PTC Frequency		PTC Acquisition	
Survey	N	%	N	%	N	%	N	%	N	%
Lit Coach	33	8.2	4	1.0	15	3.7	12	3.0	49	12.2
Pt Care Tech	84	21.5	29	7.4	105	26.9	35	9.0	126	32.3
Pharm Tech	47	11.0	11	2.6	57	13.3	5	1.2	96	22.3

The percentage of respondents to each survey and version who were flagged on between zero and five indices (none were flagged by six) is shown in Table 17. Several pertinent observations can be made with respect to these data. First, if carelessness is defined as having at least one index value below its cutoff, the result is a large reduction in sample size available for analysis, ranging from 29% for the literacy coach knowledge-based survey to 79% for the pharmacy technician task-based survey. Second, as was the case with the results in Table 15 and The highest percentage of knowledge survey records was flagged based on person-total correlation–acquisition and the lowest percentage based on long string–frequency. As was the case with task surveys, patient care technicians and pharmacy technicians had higher percentage rates of flagging than literacy coaches, although the magnitude of the differences was less for knowledge surveys than for task surveys. These findings indicate that the application of a consistent set of rules led to different extents of flagging both within surveys (i.e., between index type) and between professions.

Table 16, flagging did not produce not consistent results across occupations. The amount of carelessness was lower for literacy coaches than for patient care technicians

and pharmacy technicians. Third, the highest percentage of flagged responses occurred on the basis of a single index, indicating that careless respondents were more likely to demonstrate a single type rather than multiple types of careless responding.

Table 17

Number and Percentage of CI Indices for which Respondents Flagged

	None		One		Two		Three		Four		Five	
Survey	N	%	N	%	N	%	N	%	N	%	N	%
Tasks												
Lit Coach	263	64.8	109	26.8	30	7.4	3	0.7	1	0.2	0	0.0
Pt Care Tech	105	30.5	143	41.6	70	20.3	17	4.9	8	2.3	1	0.3
Pharm Tech	109	21.2	270	52.6	112	21.8	21	4.1	1	0.2	0	0.0
Knowledge												
Lit Coach	275	71.1	100	24.9	16	4.0	0	0.0	0	0.0	0	0.0
Pt Care Tech	113	29.0	176	45.1	86	22.1	13	3.3	2	0.5	0	0.0
Pharm Tech	239	55.7	143	33.3	42	9.8	5	1.2	0	0.0	0	0.0

Table 18 displays the number of records flagged for each index for either one or both rating scales, the total number of flagged values, and the number of unique records flagged. If all records with at least one flag were deleted based on the initial decision rules, different numbers of records would be eliminated for each survey. For literacy coaches, 32.0% of the task records and 25.2 % of the knowledge records would be eliminated. For patient care technicians, 82.6% of the task and 68.7% of the knowledge records would be eliminated. For pharmacy technicians, 75.4% of the tasks and 39.6% of the knowledge records would be eliminated. Such a large reduction in the number of responses reduces the overall representativeness of the respondent group and runs the risk of selecting out other, correlated variables.

Table 18

Number of Records Flagged using Initial Decision Rules

	Mahalanobis Distance			Long String Analysis			Person-Total Correlation			Total Number Flags	Unique Records Flagged
Survey	Neither	One	Both	Neither	One	Both	Neither	One	Both		
Tasks											
Lit Coach	342	51	13	372	42	6	379	24	3	139	130
Pt Care Tech	200	116	28	241	137	13	289	45	10	349	284
Pharm Tech	298	174	41	316	157	24	466	44	3	443	387
Knowledge											
Lit Coach	368	33	n/a	382	19	0	343	55	3	110	101
Pt Care Tech	306	84	n/a	271	104	15	241	137	12	352	268
Pharm Tech	382	47	n/a	369	52	8	329	99	1	207	170

Note. *One* denotes index flagged for single rating scale; *Both* denotes index flagged for both rating scales. Number of cases for tasks: literacy coach 406; patient care technician, 344; pharmacy technician, 513. Number of cases for knowledge: literacy coach, 401; patient care technician, 390; pharmacy technician, 429.

I next reviewed histograms for each index (see Appendix B) to investigate why such varying numbers of records were being flagged. For the task ratings, I made the following observations.

- The distribution of Mahalanobis distance–importance ratings had a large left tail but was relatively uniform for patient care technicians and pharmacy technicians.
- Large numbers of long strings values were found at the highest end of the distribution, representing respondents who did not vary from the "highly important" response option.
- There were many values of 0.0 for person-total correlation–importance, a finding particularly pronounced for patient care technicians and pharmacy technicians.

For the knowledge ratings, I made the following observations.

- Mahalanobis distance–frequency values for patient care technicians were more uniformly distributed than values for literacy coaches and pharmacy technicians.
- There was a large number of 0.00 values for person-total correlation–frequency for patient care technicians. Pharmacy technicians had a smaller number and literacy coaches had almost none.
- Patient care technicians had more 0.00 values than any other value for person-total correlation–importance.
- After tapering off toward the high end of the distribution, there was large number of long string–acquisition values at the highest point of the distribution for patient care technicians and pharmacy technicians. Inspection of the data revealed a large number of respondents who did not vary from the "acquisition before assuming job responsibilities" response option. A smaller number consistently selected the highest value on the frequency of knowledge use scale.

Based on the number of respondents excluded by the decision rules, the observations on response distribution anomalies, and the differences among the three job analysis surveys, I decided that it was not practical to apply the original cut points to the datasets. When respondents provided little or no differentiation in their ratings on importance and acquisition scales, as was the case for the patient care technician and pharmacy technician studies, CI indices flagged too many responses to be useful.

Revised decision rules. Studies have used a variety of methods to assign respondents to careful and careless categories, and there are no universally accepted rules

for setting cut scores (Curran, 2016). A logical process for establishing rules and a careful review of the results of decision rule application are essential. Based on review of the results of the initial rules, and the clear problems with the data and the indices applied to importance and acquisition rating scales, I decided to use only frequency scale-based decision rules. The revised decision rules are shown in Table 19.

Table 19

Revised Decision Rules for Flagging CI Values as Careful or Careless

Index	Cut Point
Mahalanobis distance–frequency	All values significant at $p < .001$
Long string–frequency	All values equal to or greater than 75% of statements rated
Person-total correlation–frequency	All values less than 0.00

The results of applying the revised decision rules are shown in Tables 20 and 21. Table 20 displays the number and percentage of indices for which each respondent was flagged. For the task-based surveys, the percentage of respondents flagged on at least one index ranged from 8.9% for literacy coaches to 27.1% for pharmacy technicians. For the knowledge-based surveys the percentage flagged ranged from 12.2% for pharmacy technicians to 33.3% for patient care technicians. Most respondents were flagged on only a single index.

Table 20

Revised Number and Percentage of CI Indices for which Respondents Flagged

	Neither		One		Both	
Survey	N	%	N	%	N	%
Tasks						
Lit Coach	370	91.1	36	8.9	0	0.0
Pt Care Tech	264	76.7	73	21.2	7	2.0
Pharm Tech	374	72.9	120	23.4	19	3.7
Knowledge						
Lit Coach	352	87.8	49	12.2	0	0.0
Pt Care Tech	260	66.7	112	28.7	18	4.6
Pharm Tech	367	85.5	61	14.2	1	0.2

Note. *One* denotes index flagged for single rating scale; *Both* denotes index flagged for both rating scales. Number of cases for tasks: literacy coach 406; patient care technician, 344; pharmacy technician, 513. Number of cases for knowledge: literacy coach, 401; patient care technician, 390; pharmacy technician, 429.

Table 21 shows the number of records flagged for each index, the total number of flagged values, and the number of unique records flagged using the revised decision rules. Relative to the results of applying the initial decision rules (see Table 18), the revised rules sharply decreased the number of flagged records relative to the original rules. The largest decreases were for the patient care technician and pharmacy technician task surveys.

Table 21

Number of Records Flagged using Revised Decision Rules

	Mahalanobis Distance	Long String Analysis	Person-Total Correlation	Total Number Flags	Unique Records Flagged	Reduction from Initial Decision Rules
Task-based Versions						
Lit Coach	25	2	9	36	36	94
Pt Care Tech	55	7	25	87	80	204
Pharm Tech	97	30	31	158	139	248
Knowledge-based Versions						
Lit Coach	33	4	12	49	49	52
Pt Care Tech	84	29	35	148	130	138
Pharm Tech	47	11	5	63	62	108

Note: Number of cases for tasks: literacy coach 406; patient care technician, 344; pharmacy technician, 513. Number of cases for knowledge: literacy coach, 401; patient care technician 390; pharmacy technician, 429.

Research Question 5

RQ5: Is there a relationship between careless responding and the psychometric characteristics of job analysis data?

Table 22 displays the psychometric properties of the task-based surveys before and after removing data for CI respondents. Two observations are pertinent. First, there are few differences between pre- and post-removal values of interitem correlations, mean frequency and importance ratings, or inter-class correlation measures of reliability. Where differences exist, they are small in magnitude, generally less than .05 scale points. Second, there is little uniformity in the directionality of differences. In some cases, the values are larger pre-exclusion and in other cases they are larger post-exclusion.

On the frequency scale, values for the intraclass correlation coefficient changed minimally (no more than 0.02) for literacy coaches and patient care technicians.

Differences were slightly larger for pharmacy technicians (between 0.05 and 0.07). Average item intercorrelations differed by less than 0.05 for literacy coaches and patient care technicians but differed up to 0.10 for pharmacy technicians. Finally mean ratings remained relatively similar for all three groups, with differences of no more than 0.02. The importance ratings, not selected for removal, did not differ by more than 0.03 except for patient care technicians. For Domain 3, there was a 0.06 difference, with the correlation lower after removal of careless data.

Table 22

Task Rating Scale Psychometrics Pre- and Post-Removal of CI Responses

		Frequency						Importance					
		CI Included			CI Removed			CI Included			CI Removed		
	# Items	ICC	AIC	M	ICC	AIC	M	ICC	AIC	M	ICC	AIC	M
Lit Coach													
Domain 1	8	.80	.39	3.8	.81	.40	3.8	.70	.26	3.7	.69	.27	3.7
Domain 2	27	.92	.36	3.7	.93	.38	3.7	.91	.31	3.5	.91	.31	3.6
Domain 3	14	.87	.40	3.7	.87	.41	3.7	.86	.35	3.5	.87	.35	3.6
Domain 4	9	.92	.60	2.9	.93	.61	2.9	.92	.57	3.3	.92	.57	3.3
Pt Care Tech													
Domain 1	40	.95	.37	3.7	.94	.37	3.8	.97	.50	3.6	.97	.48	3.6
Domain 2	17	.78	.29	3.7	.76	.30	3.8	.94	.52	3.7	.94	.50	3.7
Domain 3	6	.81	.46	4.4	.79	.46	4.5	.92	.65	3.8	.89	.57	3.9
Domain 4	14	.95	.60	3.5	.95	.63	3.6	.95	.59	3.6	.95	.60	3.7
Domain 5	9	.95	.70	3.3	.95	.74	3.4	.94	.64	3.6	.95	.67	3.7
Pharm Tech													
Domain 1	31	.91	.31	4.2	.84	.23	4.4	.93	.33	3.7	.93	.32	3.7
Domain 2	10	.88	.53	3.9	.83	.46	4.0	.94	.63	3.7	.94	.60	3.7
Domain 3	47	.94	.32	3.7	.89	.27	3.9	.97	.42	3.7	.96	.38	3.7
Domain 4	8	.84	.47	4.5	.78	.43	4.6	.86	.47	3.9	.87	.50	3.9

Note. ICC = Intraclass correlation coefficient. AIC = Average item intercorrelation. M = Mean rating across tasks in Domain.

Table 23 displays the psychometric properties of the knowledge-based surveys before and after removing data for CI respondents. Differences were generally small in

magnitude. For literacy coaches, reliability decreased slightly as did inter-item correlations, while mean task ratings generally remained the same. For patient care technicians, the magnitude of differences was larger than for the other two surveys and was greatest for Domain 3. Mean knowledge ratings increased very slightly (0.01) for frequency and decreased for acquisition. Reliability increased from .87 to .93 for Domain 3 (Infection Control). Task intercorrelations in this same domain increased from .55 to .66. Finally, for pharmacy technicians, the results were nearly identical pre- and post-removal of flagged data. In total, the results do not suggest enhanced psychometric properties of job analysis data after flagged, careless records are removed.

Table 23

Knowledge Rating Scale Psychometrics Pre- and Post-Removal of CI Responses

	# Items	Frequency						Acquisition					
		CI Included			CI Deleted			CI Included			CI Deleted		
		ICC	AIC	M	ICC	AIC	M	ICC	AIC	M	ICC	AIC	M
Lit Coach													
Foundational	14	.82	.30	4.2	.81	.29	4.2	.75	.18	2.2	.71	.16	2.3
Domain 1	9	.88	.49	3.9	.87	.48	3.9	.79	.31	2.4	.76	.28	2.4
Domain 2	15	.90	.43	4.1	.89	.41	4.1	.85	.29	2.3	.82	.26	2.4
Domain 3	7	.82	.46	3.8	.80	.45	3.8	.77	.32	2.3	.71	.26	2.4
Domain 4	9	.89	.52	3.3	.87	.48	3.3	.86	.41	2.4	.76	.27	2.6
Pt Care Tech													
Domain 1	57	.98	.44	3.8	.97	.41	3.9	.97	.38	2.2	.98	.42	2.1
Domain 2	16	.93	.49	3.9	.92	.51	4.0	.94	.51	2.2	.95	.56	2.1
Domain 3	7	.87	.55	3.8	.93	.66	4.4	.92	.61	2.2	.94	.68	2.1
Domain 4	23	.99	.75	3.8	.99	.80	3.9	.97	.61	2.2	.98	.67	2.1
Domain 5	11	.97	.72	3.6	.97	.76	3.7	.95	.63	2.2	.97	.74	2.1
Pharm Tech													
Domain 1	64	.96	.33	3.6	.95	.32	3.6	.96	.28	2.2	.96	.28	2.2
Domain 2	17	.92	.49	4.1	.93	.51	4.1	.92	.42	2.2	.91	.41	2.2
Domain 3	68	.95	.30	3.6	.95	.30	3.6	.97	.31	2.2	.97	.31	2.2
Domain 4	21	.94	.52	3.4	.94	.51	3.4	.95	.49	2.2	.95	.50	2.2

Note. ICC = Intraclass correlation coefficient. AIC = Average item intercorrelation. M = Mean rating across tasks in a domain.

RQ6: Are there differences in terms of selection of tasks for a certification test content outline?

Decision rules for including a task or knowledge base in a certification test content outline are made based on mean ratings for the rating scales used to validate the statements, as well as considerations of the purpose of the certification (e.g., entry to practice versus post-entry), employer hiring criteria, subgroup patterns in ratings, and other related factors. An overarching concern is to assure that important aspects of the job are represented. The creation of decision rules for a professional job analysis requires the judgment of a subject-matter expert committee. Each job analysis in this study employed a unique set of decision rules for considering whether a task base was validated or not. The decision rules are displayed in Table 24. In all instances, the validated elements did not change based on removal of the flagged careless responses. This finding is not surprising given the extremely modest nature of changes in psychometric properties of the datasets resulting from removal of careless data. Thus, there were no practical implications of removing the CI data.

Table 24

Results of Application of Validation Thresholds

Survey	Decision Rules	Number of Validated Items
Tasks		
Lit Coach	Mean rating ≥ 2.5 for frequency and ≥ 3.0 for importance	58/58
Pt Care Tech	Mean rating ≥ *3.0* for Frequency and ≥ 2.5 for Importance	78/86
Pharm Tech	Mean rating ≥ 3.0 for frequency and ≥ 3.0 for importance	83/96
Knowledge		
Lit Coach	Mean frequency rating ≥ 2.5 and acquisition before assuming job responsibilities ≥ 40% respondents	47/54
Pt Care Tech	Mean frequency rating ≥ 3.5 and acquisition before certification by ≥ 60% respondents	115/115
Pharm Tech	Mean frequency rating ≥ 2.5 and acquisition before assuming job responsibilities ≥ 50% respondents	160/170

Summary of Findings

In this study, two types of careless responding were identified: one based on long strings of identical responses and the other based on ratings patterns that differed from those of other respondents. The former was identified using long string analysis and the latter by Mahalanobis distance and person-total correlation. The extent of careless responding was found to widely depending on the detection index used and the occupation studied. Hypothesized relationships between carelessness and job analysis features were only partially supported due to differences within and between job analysis studies. The initial development and application of thresholds to categorize respondents into careful and careless groups overselected respondents to knowledge-based surveys, and overselected on both importance and acquisition ratings, in some case selecting more than half the survey respondents. Because of this, thresholds were ultimately applied only to the frequency ratings for the task-based survey versions. After the responses for careless responses were removed from the datasets, mean ratings, average item intercorrelations, and reliability values changed only minimally, and did not affect the tasks selected for inclusion in certification test content outlines.

In Chapter 5, the results of the study are discussed. Interpretations are provided in the context of theory and prior research. Implications, limitations to generalizability, and recommendations for further research are outlined and the impact of the study's findings on social change are examined.

Chapter 5: Conclusions

Introduction

This study was conducted to investigate the phenomenon of careless survey responding. The specific type of survey studied was the job analysis survey conducted to support certification program test development. Carelessness and its correlates in three different job analysis surveys were examined to investigate generalizability of findings across professions. The impact of careless data on the psychometric properties of job analysis data were investigated, as was the extent to which carelessness affected test content outlines derived from job analysis survey data.

The results of this study indicated that the extent of CI responding differs widely based on the index used and specific job analysis study conducted. Hypothesized relationships between carelessness and job analysis features were partially supported and dependent on the occupation and index. Factor analysis results confirmed that the three detection indices identified different patterns of carelessness. In general, all three indices were most useful when applied to concrete tasks rated on an absolute frequency scale. Finally, hypothesized relationships between carelessness and the psychometric properties of job analysis data were not supported, and there was no impact of carelessness on certification test content outlines.

Interpretation of the Findings

This study examining CI responding in job analysis surveys drew upon two bodies of research to develop testable hypotheses. The first is the limited body of

research on job analysis carelessness. Most of these studies are more than 10 years old and all employed only a single carelessness detection method, the inclusion of bogus survey items. The second is a larger body of knowledge related to post hoc CI detection methods. Both bodies of research as well as predictions based on satisficing theory are discussed in interpreting the findings.

Survey Length

Satisficing theory suggests that motivational factors influence survey ratings. Longer surveys are associated with decreased motivation to respond accurately due to survey fatigue based on sustained cognitive demands (Daniel, 2012; Krosnick, 1996). Based on satisficing theory, it was hypothesized that surveys with more items to rate would be associated with more carelessness.

While prominent job analysis researchers have long recommended studying the relationship between survey length and response characteristics, almost no research has been conducted to date in this arena. Wang et al. (1999) found that selective non-response to a single rating scale when multiple scales were used increased in frequency with survey length. They also found that in later portions of a job analysis survey, respondents increased their use of only a single scale when making ratings. A meta-analysis conducted by DuVernet et al. (2015) suggested a more-complex relationship between survey length and data quality. They found that interrater reliability and between-job discriminability increased with survey length and then diminished. Outside

the job analysis context, studies have demonstrated a relationship between survey length and data quality (Galesic & Bosnjak, 2009; Hardré et al., 2012; Zhang & Conrad, 2014).

In the present study, for task-based survey versions, longer survey length was weakly associated with a higher incidence of invariant responses as indicated by long string values. For knowledge-based survey versions, longer lengths were primarily associated with more response pattern outliers is indicated by greater Mahalanobis distance values. While the results were not entirely consistent, the findings suggest that some job analysis survey respondents respond more carelessly to longer surveys.

Job Aspect Rated

In this study, survey takers rated one of two job aspects, either tasks, which represent work-oriented activities performed on the job, or knowledge, which represents worker characteristics needed to perform the job. Respondents rating tasks evaluated statements that were "specific, concrete, and directly observable" (Stetz et al., 2012, p. 103). In contrast, respondents rating knowledge made judgments about statements that were not directly observable, necessitating more complex and subjective inferences. Based on satisficing theory as well as job analysis theory and research, it was hypothesized that individuals responding to knowledge-based surveys would exhibit more carelessness in their frequency ratings than individuals responding to task-based surveys.

Higher levels of carelessness for knowledge ratings than task ratings were not found consistently. Instead, CI indices were differentially sensitive to different patterns

of carelessness on the task and knowledge surveys. Literacy coaches and patient care technicians exhibited more inconsistent responding when rating tasks and more invariant responding when rating knowledge. Also, each occupation exhibited unique patterns of differences in CI index values. Taken together, the findings suggest a more complex relationship between survey length and carelessness than originally envisioned.

Rating Scale Used

Stetz et al. (2012) found more carelessness for abstract scales that required respondents to make inferences than for concrete scales. This does not mean that importance scales should not be used in job analysis, rather that removal of careless responses be undertaken. Although Christal and Weissmuller (1988) recommended against using importance ratings due to the complexity of inferences required, this is a somewhat extreme view. It is at odds with the typical practice in licensure and certification testing. To create test outlines for licensure and certification, job analysts typically make use of multiple rating scales (Cadle, 2012; Raymond, 2001). As Raymond (2005) said, the identification of important job tasks in creating assessments is consistent with the Standards for Educational and Psychological Testing (APA, AERA & NCME, 2014).

This study found consistent differences in CI index values based on rating scale used. For task ratings across all occupations, there were consistently higher long string values for importance than frequency, and consistently higher person-total correlation values for frequency than importance. For knowledge ratings across all occupations,

there were consistently higher long string values for acquisition than frequency, and consistently higher person-total correlation values for frequency than acquisition.

In certification job analysis, knowledge acquisition ratings are used to determine whether knowledge bases should be included on a certification examination. When a sufficient number of job analysis survey takers indicate a knowledge based should be acquired before certification, it becomes eligible for inclusion in a certification test content outline. Carelessness in responding to acquisition scales proved difficult to distinguish from normal responding, given that nearly all respondents selected the before certification option on this scale.

CI Index Performance

Findings from this study demonstrated that the three CI indices capture different sources of rater variance in certification job analysis surveys. The correlation and factor analysis results strongly support prior findings (McKay et al., 2018; Meade & Craig, 2012; Niessen, Meijer, & Tendeiro, 2016) that Mahalanobis distance and long string analysis identify two different types of CI responding. The results further suggest that person-total correlation identifies a third type of CI responding. Because each index captured a distinct pattern of CI responding, the results support recommendations by DeSimone et al. (2015) and Curran (2016) to use multiple indices.

Extent of Careless and Inattentive Ratings

As discussed in Chapter 2, the exact extent of careless and inattentive responding in survey data has proved difficult to establish given the wide range of simulation and

survey data studied and the different methods used to establish cutoffs. I found that depending on index used, job aspect, rating scale, and cutoffs applied, carelessness varied among professions, in some cases substantially. This argues against the generalizability of findings across certification job analyses and argues instead for the existence of idiosyncratic rater differences based on occupation. Literacy coach survey respondents had the lowest rates of careless responding. There is evidence that level of education is associated with job analysis ratings accuracy (Green & Veres, 1990; Zhang & Conrad, 2014), and literacy coaches require more education and training to be eligible for certification that patient care technicians and pharmacy technicians.

Establishing cutoffs to categorize respondents into careful and careless responders was challenging in the absence of well-established methods (DeSimone et al., 2015). The use of prior approaches in the literature resulted in identification of untenably large percentages of CI responses. Examination of histograms of index values revealed that importance and acquisition scales produced little response variation. These responses are rating-scale related, not careless, and should not be removed based on CI index values. When certification job analysis studies use these scales, deterrence approaches should be used as a replacement for or an adjunct to the post hoc methods. The indices appear more useful for flagging carelessness in frequency ratings.

Psychometrics

Huang et al. (2015), Maniaci and Rogge (2014), Morgeson and Campion (2017), and Wilson et al. (1990) found the removal of careless responses improved the

psychometric characteristics of survey data. However, Steedle (2018) found no impact of removing flagged data on item intercorrelations and mean ratings based on a survey of social and emotional learning rated using a Likert-type scale. Results of the current study are consistent with those of Steedle in that the removal of flagged data had little impact overall, except on ratings for isolated domains in individual surveys. It may be that the final categorization scheme I used was too lenient and failed to detect and eliminate additional records that may have represented carelessness.

Similar to findings on the psychometric qualities of the job analysis data, there was no substantive impact on the ultimate decisions of which tasks to include in test content outlines. Differences in mean ratings were well above the typical thresholds for elimination. This brings into question the issue of whether there is need for screening and eliminating careless and attentive data from job analysis surveys. Detection methods may not be appropriate for all surveys (Ran et al., 2015). It is premature at this time to dismiss the use of post hoc detection methods, particularly because this study represented only a beginning in examining the utility of such methods in job analysis, and limitations inherent in the secondary data analytic approach may have limited the ability to fully explore their potential.

Limitations of the Study

Because this study represented a secondary analysis of already completed job analysis studies, it was not possible to manipulate variables to test the hypotheses outlined in the research questions. Analyses were restricted to the available data, a

known limitation of secondary data analysis (Johnston, 2014). Experimental or quasi-experimental designs (Campbell & Stanley, 1967) would have permitted the manipulation and control of variables with a hypothesized relationship to carelessness. Examples of such studies are described in the next section.

Regarding generalizability, it was hoped that by using data from several job analysis studies, findings might generalize occupations not included in this study. This, however, was not the case. Results not only differed across professions, in some cases, they differed within professions. Based on the inconsistencies found, generalizability is not possible.

Results were consistent with the supposition that job analysis represents a type of survey likely to induce carelessness (Huang et al., 2012). Careless survey responses were flagged using all three indices. However, over-flagging occurred for rating scales that had little response variability. The acquisition and importance scales are examples of this problem. Use of cutoff thresholds for these scales had to be abandoned because too many valid ratings were flagged as careless and inattentive. This is a particular concern in licensure and certification job analysis. Tasks and knowledge bases in a credentialing job analysis survey have already been vetted by subject-matter experts who deemed them important at entry level to a job or occupation. Collection of survey ratings is a largely confirmatory process. In this study, even the long string method—perhaps the simplest and easiest to interpret of the indices—may have over-identified records as careless and

inattentive. Alternatively, it may be that these surveys contained more data quality issues than is typical in other types of surveys.

Recommendations

There are many sources of variance in job analysis ratings (Richman & Quiñones, 1996; Sanchez & Fraser, 1992; Van Iddekinge et al., 2005; Wang et al., 1999). This study suggests that careless responding is one of those sources. Yet the limitations of the study suggest clear avenues for further research. In terms of study design, a more-controlled study in which the respondents rated both tasks and knowledge, with the order of presentation of the two counterbalanced, would permit a more systematic exploration of within- and between-subject CI responding for both types of ratings. Another avenue of research might be examination of the optimum length for job analysis surveys, given survey fatigue and it hypothesized impact on careful responding. Splitting data collection into smaller subsets of job analysis elements and examining the impact on ratings accuracy would be useful. Capturing personality variables as part of data collection may also be useful as emerging research suggests individual differences as a systematic source of variance in CI responding.

Lack of correspondence with prior findings may because prior studies used Likert-type scales, which have different properties than job analysis rating scales. In this study, the behavior of post hoc detection indices was clearly influenced by the rating scales used. It would be helpful to explore whether other types of job analysis rating scales, for example, difficulty of acquisition or performance, exhibited more response

variance that would better permit detection of data issues. Perhaps adding explicit instructions to the survey instrument would enhance response accuracy and reduce carelessness. For example, warnings to be careful have shown to decrease long string responding (Ward & Pond, 2015). Indeed, it has been suggested that research using prevention methods will be a fruitful avenue for study (Morgeson et al., 2014). Mitigation of CI responding during the data collection process places less onus on post hoc methods of data cleaning.

Future research is needed to better understand the potential utility of post hoc CI indices for knowledge-based surveys job analysis surveys. The indices used in the present study, combined with a lack of variance in responses, identified too many cases as careless to be useful. In particular, the application of the results for the acquisition rating scale would have resulted in screening out more than 50% of the data for one of the surveys. Additional research on methods for establishing cut points is needed, as well as examination of how application of these cut points affects the psychometric properties of collected survey data.

Implications

Job analysis survey ratings contain a great deal of unexplained variance (Morgeson & Campion, 2017; Schmitt & Stuits, 1985; Wang et al., 1999). The three indices used in this study were able to detect three different types of variance unrelated to the job analysis constructs being surveyed. However, lacking well-established methods for setting cut points for categorizing job analysis responses as careful or careless,

practical use of these indices was limited in this study. The thresholds adopted did not result in improvements in job analysis data. It was hoped that the results of this study would inform job analysis practices more immediately, but future research will be needed to determine whether use of post hoc indices will improve the accuracy of data used to develop licensure and certification test content outlines.

Because 24% of the adult US employed population hold licenses or certifications (Bureau of Labor Statistics, 2017), the accuracy of job analysis survey data is essential to ensuring that test content outlines for licensure and certification programs are an accurate representation of practice. While inconclusive, findings from this study do suggest that not all survey takers who contribute to test content outline development give their sustained effort to the ratings process. The extent to which this may affect substantive aspects of job analysis data is an avenue for further research, as this study was hampered by challenges in establishing cut points. Regardless of whether post hoc indices of CI responding prove to influence substantive findings, I would argue that they should to be investigated and used in an informed way to clean job analysis data.

Conclusion

Job analysis surveys supporting licensure and certification programs yield important data for establishing the content validity argument for the programs. When test content outlines established from credentialing job analysis surveys are used to align all subsequent item and examination development activities, scores on examinations can be interpreted as accurate representations of the credentialing construct. The level of

accuracy of the job analysis survey data from which test outlines are derived can either support or undermine the content validity argument. This study investigated three post hoc methods for detecting inaccurate job analysis survey data, each of which appeared to identify a different type carelessness and inaccuracy. While the findings suggest that applying these methods to *all* job analysis rating scales may not be warranted, they do at least appear useful when applied to frequency rating scales. Further research is clearly warranted. Until then, particularly in the context of high-stakes credentialing assessment, judicious identification and removal inaccurate job analysis data will continue to be necessary to support to validity inferences. As argued by Harvey and Wilson (2000), "what matters is finding and fixing inaccuracies whatever their causes may have been" (p. 849). As the body of research on detection methods for careless and inattentive responding continues to evolve, a more sophisticated understanding of their appropriate use in job analysis will develop.

References

Aguinis, H., Mazurkiewicz, M. D., & Heggestad, E. D. (2009). Using web-based frame-of-reference training to decrease biases in personality-based job analysis: An experimental field study. *Personnel Psychology*, *62*(2), 405-438.

Albert, K. (2017). The certification earnings premium: An examination of young workers. *Social Science Research*, *63*, 138-149. https://doi.org/10.1016/j.ssresearch.2016.09.022

Alwin, D. (2016). Survey data quality and measurement precision. In C. Wolf, D. Joye, T. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 527-557). Thousand Oaks, CA: SAGE Publications, Inc.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Anduiza, E., & Galais, C. (2017). Answering without reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, *29*(3), 497-19. https://doi.org/10.1093/ijpor/edw007

Anseel, F., Lievens, F., Schollaert, E., & Choragwicka, B. (2010). Response rates in organizational science, 1995–2008: A meta-analytic review and guidelines for survey researchers. *Journal of Business and Psychology*, *25*(3), 335-349. https://doi.org/10.1007/s10869-010-9157-6

Baer, R. A., Ballenger, J., Berry, D. T., & Wetter, M. W. (1997). Detection of random responding on MMPI-A. *Journal of Personality Assessment*, *68*(1), 139-151.

Bainbridge, H. T. J., Sanders, K., Cogin, J. A., & Lin, C.-H. (2017). The pervasiveness and trajectory of methodological choices: A 20-year review of human resource management research. *Human Resource Management*, *56*(6), 887-913. https://doi.org/10.1002/hrm.21807

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education*, *53*(2), 182-200. https://doi.org/10.1007/s11162-011-9251-2

Berry, D. T., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4*(3), 340.

Bowling, N. A., & Huang, J. L. (2018). Your attention please! Toward a better understanding of research participant carelessness. *Applied Psychology*, *67*(2), 227-230. https://doi.org/10.1111/apps.12143

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Lui, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, *111*(2), 218-229. https://doi.org/10.1037/pspp0000085.supp

Breitsohl, H., & Steidelmüller, C. (2018). The impact of insufficient effort responding detection methods on substantive responses: Results from an experiment testing

parameter invariance. *Applied Psychology*, *67*(2), 284-308. https://doi.org/10.1111/apps.12121

Bureau of Labor Statistics. (2017). Certification and licensing status of employed persons 16 years and over. Retrieved from https://www.bls.gov/cps/cpsaat49.pdf

Cadle, A. W. (2012). *The Relationship between Rating Scales Used to Evaluate Tasks from Task Inventories for Licensure and Certification Examinations*. (Doctoral dissertation.) Retrieved from https://scholarcommons.usf.edu/etd/4296/

Campbell, D. T., & Stanley, J. C. (1967). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin Comp.

Carrier, L. M., Cheever, N. A., Rosen, L. D., Benitez, S., & Chang, J. (2009). Multitasking across generations: Multitasking choices and difficulty ratings in three generations of Americans. *Computers in Human Behavior*, *25*(2), 483-489. https://doi.org/10.1016/j.chb.2008.10.012

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159.

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web-or internet-based surveys. *Educational and Psychological Measurement*, *60*(6), 821-836.

Costa, P. T., & McCrae, R. R. (1997). Stability and change in personality assessment. *Journal of Personality Assessment*, *68*(1), 86-94.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4-19. https://doi.org/10.1016/j.jesp.2015.07.006

Curran, P. G., Kotroba, L., & Denison, D. (2010, April). *Careless responding in surveys: Applying traditional techniques to organizational settings.* Presented at the Society for Industrial and Organizational Psychology, Atlanta, GA.

Daniel, S. (2012). Satisficing in survey design. *Contemporary Approaches to Research in Mathematics, Science, Health and Environmental Education Symposium*, 29-30. Retrieved from http://www.academia.edu/download/32155638/Daniel_2012_Satisficing_in_survey_design.pdf

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*(1), 1-18.

de Vaus, D. (2013). *Surveys in Social Research*. New York, NY: Routledge/Taylor & Francis Group.

de Winter, J. C. F., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, *21*(3), 273-290. https://doi.org/10.1037/met0000079

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*(2), 171-181. https://doi.org/10.1002/job.1962

Deutskens, E., De Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters*, *15*(1), 21-36.

Dierdorff, E. C., & Morgeson, F. P. (2009). Effects of descriptor specificity and observability on incumbent work analysis ratings. *Personnel Psychology*, *62*(3), 601-628.

Dillman, D. (n.d.). Future surveys: Monthly Labor Review: U.S. Bureau of Labor Statistics. Retrieved from https://www.bls.gov/opub/mlr/2015/article/future-surveys.htm

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *Springer Plus*, *2*(1), 222. https://doi.org/10.1186/2193-1801-2-222

Donlon, T., & Fischer, F. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement*, *28*, 105-113.

Dupuis, M., Meier, E., & Cuneo, F. (2018). Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1103-y

DuVernet, A. M., Dierdorff, E. C., & Wilson, M. A. (2015). Exploring factors that influence work analysis data: A meta-analysis of design choices, purposes, and organizational context. *Journal of Applied Psychology*, *100*(5), 1603-1631. https://doi.org/10.1037/a0039084

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*, *26*(2), 132-139. https://doi.org/10.1016/j.chb.2009.10.015

Fang, J., Wen, C., & Prybutok, V. (2014). An assessment of equivalence between paper and social media surveys: The role of social desirability and satisficing. *Computers in Human Behavior*, *30*, 335-343. https://doi.org/10.1016/j.chb.2013.09.019

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). Thousand Oaks, CA: Sage Publications, Inc.

Frankfort-Nachmias, C., & Nachmias, D. (2008). *Research methods in the social sciences* (7th ed.). New York: NY: Worth Publishing.

Fulton, B. R. (2016). Organizations and survey research: Implementing response enhancing strategies and conducting nonresponse analyses. *Sociological Methods and Research*. https://doi.org/10.1177/0049124115626169

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349-360. https://doi.org/10.1093/poq/nfp031

Godinho, A., Kushnir, V., & Cunningham, J. A. (2016). Unfaithful findings: Identifying careless responding in addictions research: Editorial. *Addiction*, *111*(6), 955-956. https://doi.org/10.1111/add.13221

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7-28). Tilberg, The Netherlands: Tilberg University Press.

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology*, *48*(1), 82-98.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*(1), 549-576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Green, S. B., & Stutzman, T. (1986). An evaluation of methods to select respondents to structured job-analysis questionnaires. *Personnel Psychology*, *39*(3), 543-564.

Green, S. B., & Veres, J. G. (1990). Evaluation of an index to detect inaccurate respondents to a task analysis inventory. *Journal of Business and Psychology*, *5*(1), 47-61.

Groves, R., Fowler, F., Couper, M. P., Lepkowski, J. A., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken: NJ: John Wiley & Sons, Inc.

Guo, Y., Kopec, J. A., Cibere, J., Li, L. C., & Goldsmith, C. H. (2016). Population survey features and response rates: A randomized experiment. *American Journal of Public Health*, *106*(8), 1422-1426. https://doi.org/10.2105/AJPH.2016.303198

Hamby, T., & Taylor, W. (2016). Survey satisficing inflates reliability and validity measures: An experimental comparison of college and Amazon Mechanical Turk samples. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164415627349

Hardré, P. L., Crowson, H. M., & Xie, K. (2012). Examining contexts-of-use for web-based and paper-based questionnaires. *Educational and Psychological Measurement*, *72*(6), 1015-1038. https://doi.org/10.1177/0013164412451977

Harvey, R. J., & Wilson, M. A. (2000). Yes Virginia, there is an objective reality in job analysis. *Journal of Organizational Behavior*, 829-854.

Hauser, D. J., & Schwarz, N. (2015). It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *SAGE Open*, *5*(2), 1-6. https://doi.org/10.1177/2158244015584617

Herzog, A. R., & Bachman, J. G. (1981). Effect of questionnaire length on response quality. *Public Opinion Quarterly*, *45*, 549-559.

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*(2), 299-311. https://doi.org/10.1007/s10869-014-9357-6

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99-114. https://doi.org/10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828-845. https://doi.org/10.1037/a0038510

Jackson, D. (1977). *Jackson Vocational Interest Survey manual*. Port Huron: MI: Research Psychologists Press.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103-129. https://doi.org/10.1016/j.jrp.2004.09.009

Johnston, M. P. (2014). Secondary data analysis: A method of which the time has come. *Qualitative and Quantitative Methods in Libraries*, 8.

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, *18*(3), 512-541. https://doi.org/10.1177/1094428115571894

Kaminska, O., McCutcheon, A. L., & Billiet, J. (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly*, *74*(5), 956-984. https://doi.org/10.1093/poq/nfq062

Kochan, T., Finegold, D., & Osterman, P. (n.d.). Who can fix the "middle-skills" gap? *Harvard Business Review*, (2012). Retrieved from https://hbr.org/2012/12/who-can-fix-the-middle-skills-gap

Krosnick, J. A. (1987). Cognitive theory of response order effects. *Public Opinion Quarterly*, *51*(2), 201-219.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213-236.

Krosnick, J. A. (1996). Satisficing in surveys: Initial evidence. In *M.T. Braverman & J.K Slater (Eds.)*. *Advances in survey research* (pp. 29-44). San Francisco, CA: Jossey-Bass.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, (50), 537-567.

Krosnick, J. A., & Presser, S. (2010). Survey questions. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed.). Bingley: Emerald.

Landy, F. J., & Vasey, J. (1991). Job analysis: The composition of SME samples. *Personnel Psychology*, *44*(1), 27-50.

Lievens, F., & Sanchez, J. I. (2007). Can training improve the quality of inferences made by raters in competency modeling? A quasi-experiment. *Journal of Applied Psychology*, *92*(3), 812-819. https://doi.org/10.1037/0021-9010.92.3.812

Lievens, F., Sanchez, J. I., Bartram, D., & Brown, A. (2010). Lack of consensus among competency ratings of the same occupation: Noise or substance? *Journal of Applied Psychology*, *95*(3), 562.

Mahalanobis, P. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, *12*, 49-55.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61-83. https://doi.org/10.1016/j.jrp.2013.09.008

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28-50. https://doi.org/10.1177/1088868310366253

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*(3), 450-470. https://doi.org/10.1037/a0019216

McKay, A. S., Garcia, D. M., Clapper, J. P., & Shultz, K. S. (2018). The attentive and the careless: Examining the relationship between benevolent and malevolent personality traits with careless responding in online surveys. *Computers in Human Behavior*, *84*, 295-303. https://doi.org/10.1016/j.chb.2018.03.007

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437-455. https://doi.org/10.1037/a0028085

Meade, A. W., & Pappalardo, G. (2013, April). *Predicting careless responses and attrition in survey data with personality*. Presented at the meeting of the Society for Industrial and Organizational Psychology, Houston, TX.

Morgeson, F., & Campion, M. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology*, *82*(5), 627-655.

Morgeson, F., & Campion, M. A. (2000). Accuracy in job analysis: Toward in inference-based model. *Journal of Organizational Behavior*, *21*, 819-827.

Morgeson, F., & Campion, M. A. (2017). A framework of sources of inaccuracy in job analysis. In M. Wilson, W. Bennett Jr, S. Gibson, & G. Alliger (Eds.), *The handbook of work analysis: The methods, systems, applications, and science of work measurement in organizations.* New York: NY: Psychology Press/Taylor & Francis.

Morgeson, F., Delaney-Klinger, K., & Hemingway, M. A. (2005). The importance of job autonomy, cognitive ability, and job-related skill for predicting role breadth and job performance. *Journal of Applied Psychology*, *90*(2), 399-406. https://doi.org/10.1037/0021-9010.90.2.399

Morgeson, F., Spitzmuller, M., Garza, A. S., & Campion, M. A. (2014). Pay attention! The liabilities of respondent experience and carelessness when making job analysis judgments. *Journal of Management*, *42*(7), 1904-1933. https://doi.org/10.1177/0149206314522298

National Workforce Solutions Advisory Board. (2017). *Understanding and Solving the Skills Gap: A Call to Action* (p. 16). Iowa City, IA: ACT.

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*(4), 372-411. https://doi.org/10.1177/1094428114548590

Onwuegbuzie, A. (2000, November). *Expanding the framework of internal and external validity in quantitative research*. Presented at the meeting of the Association for the Advancement of Educational Research (AAER), Ponte Vedra, FL.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867-872. https://doi.org/10.1016/j.jesp.2009.03.009

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology*, *78*(3), 582-593. https://doi.org/10.1037//0022-3514.78.3.582

Prien, K. O., Prien, E. P., & Wooten, W. (2003). Interrater reliability in job analysis: Differences in strategy and perspective. *Public Personnel Management*, *32*(1), 125-141.

Ran, S., Liu, M., Marchiondo, L. A., & Huang, J. L. (2015). Difference in response effort across sample types: Perception or reality. *Industrial and Organizational Psychology*, *8*(02), 202-208.

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*(3), 25-37.

Raymond, M. R., & Luecht, R. M. (2013). Licensure and certification testing. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education.* (pp. 391-414). https://doi.org/10.1037/14049-019

Rey, D., & Neuhäuser, M. (2011). Wilcoxon-Signed-Rank Test. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 1658-1659). Germany: Berlin: Springer Berlin.

Richman, W. L., & Quiñones, M. A. (1996). Task frequency rating accuracy: The effect of task engagement and experience. *Journal of Applied Psychology*, *81*(5), 512.

Roivainen, E., Veijola, J., & Miettunen, J. (2016). Careless responses in survey data and the validity of a screening instrument. *Nordic Psychology*, *68*(2), 114-123. https://doi.org/10.1080/19012276.2015.1071202

Sanchez, J. I., & Fraser, S. L. (1992). On the choice of scales for task analysis. *Journal of Applied Psychology*, *77*(4), 545.

Sanchez, J. I., & Levine, E. L. (2001). The analysis of work in the 20th and 21st centuries. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work & organizational psychology: Volume 1: Personnel psychology*. Thousand Oaks, CA: Sage Publications, Inc.

Sanchez, J. I., & Levine, E. L. (2012). The rise and fall of job analysis and the future of work analysis. *Annual Review of Psychology*, *63*(1), 397-425. https://doi.org/10.1146/annurev-psych-120710-100401

Sarraf, S., & Tukibayeva, M. (2014). Survey page length and progress indicators: What are their relationships to item nonresponse? *New Directions for Institutional Research*, *2014*(161), 83-97. https://doi.org/10.1002/ir.20069

Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*(4), 367-373.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420.

Siddique, C. M. (2004). Job analysis: a strategic human resource management practice. *The International Journal of Human Resource Management*, *15*(1), 219-244. https://doi.org/10.1080/0958519032000157438

Singh, P. (2008). Job analysis for a changing workplace. *Human Resource Management Review*, *18*(2), 87-99. https://doi.org/10.1016/j.hrmr.2008.03.004

Sireci, S. G., & Hambleton, R. K. (2009). Mission—protect the public: Licensure and certification testing in the 21st century. *Correcting Fallacies about Educational and Psychological Testing*, 199-217.

Steedle, J. T. (2018, April). *Detecting inattentive responding on a psychosocial measure of college readiness*. Presented at the meeting of the American Educational Research Association, New York: NY.

Stetz, T. A., Button, S. B., & Quist, J. (2012). Rethinking carelessness on job analysis surveys: Not all questions are created equal. *Journal of Personnel Psychology*, *11*(2), 103-106. https://doi.org/10.1027/1866-5888/a000061

Tabachnick, B., & Fidell, L. (2007). Using multivariate statistics. In *Using Multivariate Statistics* (5th ed., p. 74). Boston, MA: Pearson.

The R Foundation. (n.d.). R: What is R? Retrieved June 25, 2018, from https://www.r-project.org/about.html

Thomas, R. K. (2014). Fast and furious … … or much ado about nothing?: Sub-optimal respondent behavior and data quality. *Journal of Advertising Research*, *54*(1), 17-31. https://doi.org/10.2501/JAR-54-1-017-031

Tourangeau, R. (1984). Cognitive aspects of survey methodology: Building a bridge between disciplines. In J. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive science and survey methods* (pp. 73-100). Washington, DC: National Academy Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Van De Voort, D., & Whelan, T. (2012). Work analysis questionnaires and app interviews. In M. A. Wilson, W. Bennett Jr, S. Gibson, & G. Alliger (Eds.), *The*

handbook of work analysis: Methods, systems, applications and science of work measurement in organizations (pp. 41-79). Taylor & Francis Group.

Van Iddekinge, C. H., Putka, D. J., Raymark, P. H., & Eidson Jr, C. E. (2005). Modeling error variance in job specification ratings: The influence of rater, job, and organization-level factors. *Journal of Applied Psychology*, *90*(2), 323.

Vannette, D. L., & Krosnick, J. A. (2014). A comparison of survey satisficing and mindlessness. In A. Ie, C.T. Ngnoumen, & E.J. Langer (Eds.), *The Wiley Blackwell Handbook of Mindfulness* (pp. 312-327). Hoboken: NJ: John Wiley & Sons, Ltd.

Vartanian, T. (2010). *Secondary data analysis*. New York, NY: Oxford University Press.

Voskuijl, O. F., & van Sliedregt, T. (2002). Determinants of interrater reliability of job analysis: A meta-analysis. *European Journal of Psychological Assessment*, *18*(1), 52-62. https://doi.org/10.1027//1015-5759.18.1.52

Wang, N., Wiser, R. F., & Newman, L. S. (1999, April). *Examining reliability and validity of job analysis survey data*. Presented at the meeting of National Council on Measurement in Education, Montreal, QC.

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology*, *67*(2), 231-263. https://doi.org/10.1111/apps.12118

Ward, M. K., Meade, A. W., Allred, C. M., Pappalardo, G., & Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of

personality traits and performance. *Computers in Human Behavior*, *76*, 417-430. https://doi.org/10.1016/j.chb.2017.06.032

Ward, M. K., & Pond, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior*, *48*, 554-568. https://doi.org/10.1016/j.chb.2015.01.070

Wilson, M. A., Harvey, R. J., & Macy, B. A. (1990). Repeating items to estimate the test-retest reliability of task inventory ratings. *Journal of Applied Psychology*, *75*(2), 158.

Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*, 127-135.

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics*, *36*(2), 186-212. https://doi.org/10.3102/1076998610366263

Appendix A: Letters of Permission to Contact Clients

Professional
Examination Service
Credentialing Insight

475 Riverside Drive
Suite 600
New York, NY 10115
212.367.4200

June 21, 2018

To Whom It May Concern:

Patricia Muenzen was employed by Professional Examination Service (ProExam) from 1992 to 2017, at which time the organization was acquired by ACT. As ProExam's Director of Research Programs, she conducted job analysis studies of professions to support the development and maintenance of our clients' licensure and certification programs.

Patricia is currently working on a dissertation in partial fulfillment of the requirements for a PhD in Industrial/Organizational Psychology at Walden University. She proposes to perform secondary data analysis on job analysis survey datasets she previously collected for ProExam clients. These datasets are the property of ProExam's client organizations. I hereby give my permission for Patricia to contact ProExam's clients and request their data for use in her dissertation research.

Please do not hesitate to contact me if you require any further information.

Sincerely,

Sandra Logorda
Sandra Logorda
Executive Director

June 25, 2018

To whom it may concern:

Patricia Muenzen is currently employed at ACT as a Director in the Credentialing Advisory Services unit of the Research Department.

Patricia is currently working on a dissertation in partial fulfillment of the requirements for a PhD in Industrial/Organizational Psychology at Walden University. She proposes to perform secondary data analysis on job analysis survey datasets previously collected for ACT clients. These datasets are the property of ACT's client organizations. I hereby give my permission for Patricia to contact ACT's clients and request their data for use in her dissertation research.

Please do not hesitate to contact me if you require any further information.

Sincerely,

Vice President, Credentialing Advisory Services
Research, ACT
475 Riverside Drive, 6th Floor, New York, NY 10115
212-367-4271
sandra.greenberg@act.org

Appendix B: Histograms of Carelessness Index Values

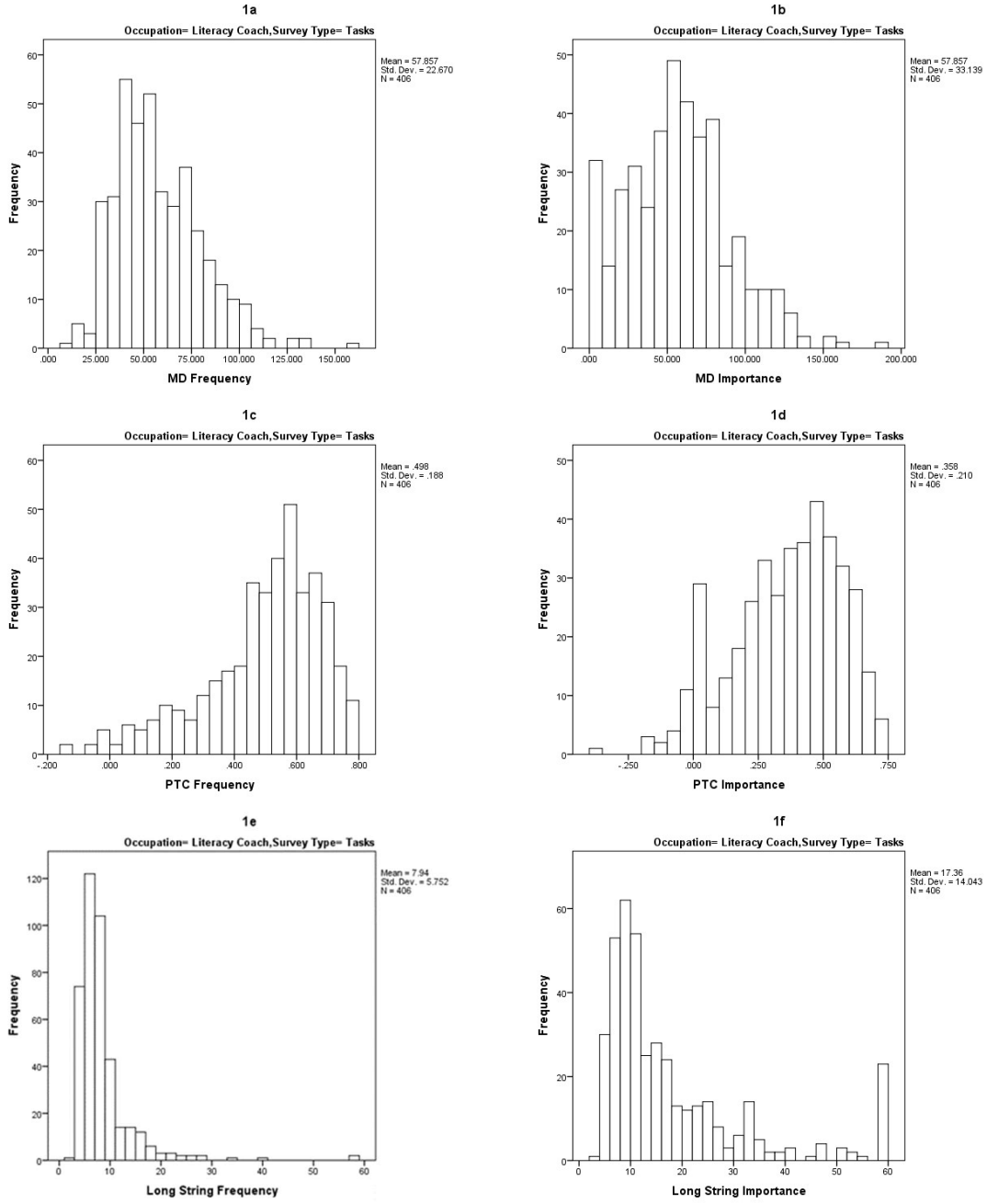


Figure B1. Histograms of task CI index values for literacy coaches. B1a shows Mahalanobis distance–frequency values, B1b shows Mahalanobis distance–importance values, B1c shows person-total correlation–frequency values, B1d shows person-total correlation–importance values, B1e shows long string–frequency values, and B1f shows long string–importance values.

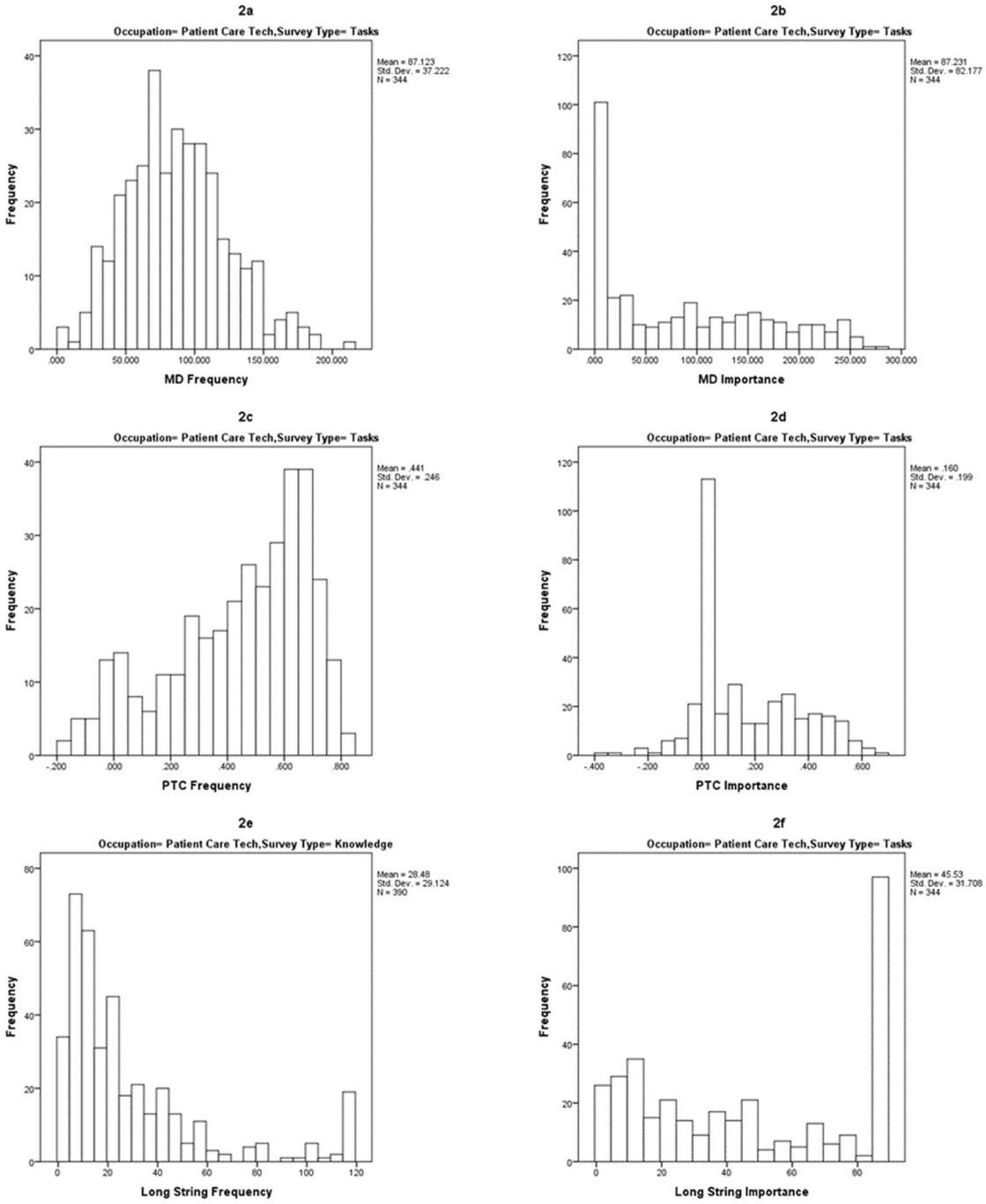


Figure B2. Histograms of task CI index values for patient care technicians. Figure B2a shows Mahalanobis distance–frequency values, B2b shows Mahalanobis distance–importance values, B2c shows person-total correlation–frequency values, B2d shows person-total correlation–importance values, B2e shows long string–frequency values, and B2f shows long string –importance values.

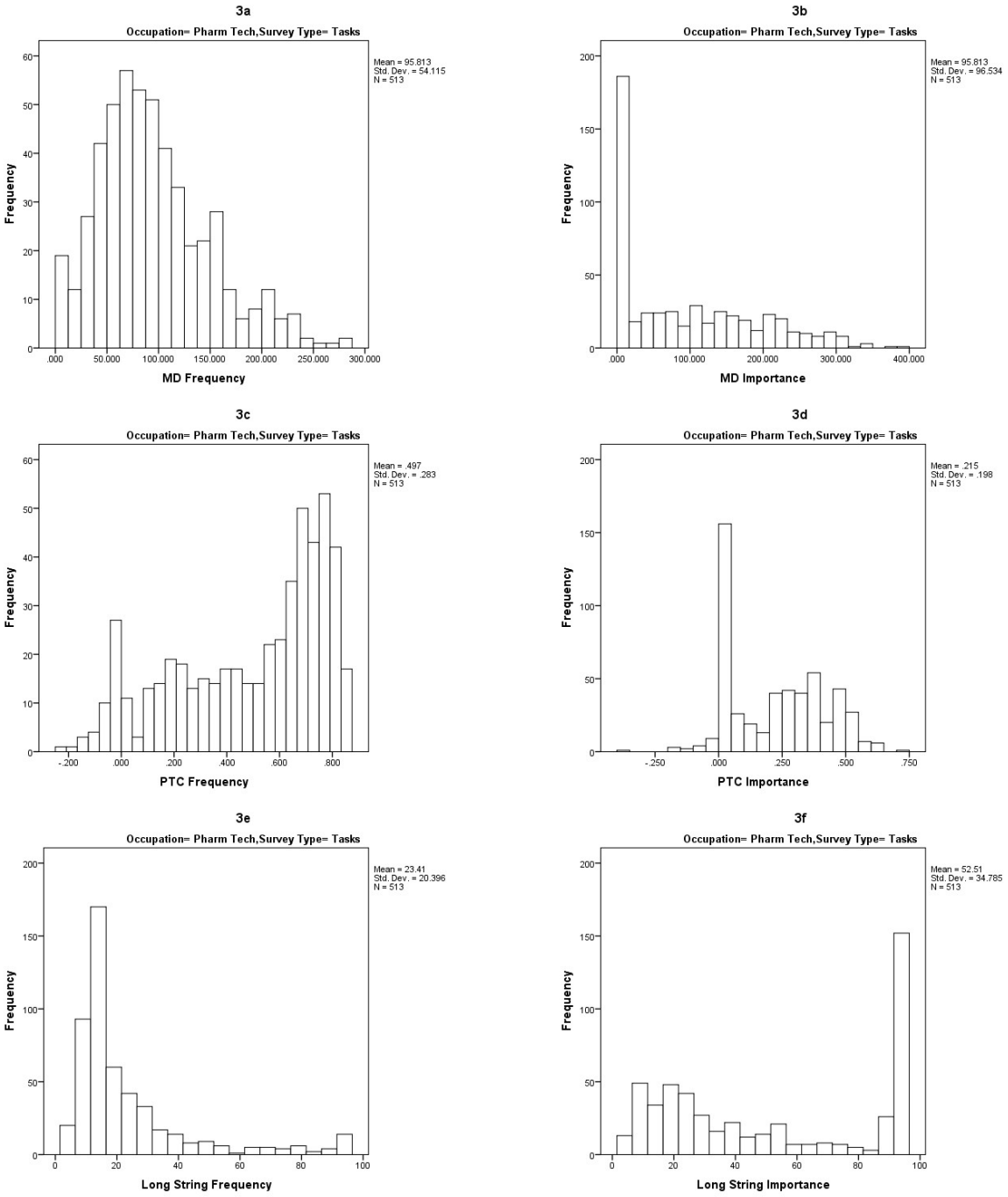


Figure B3. Histograms of task CI index values for pharmacy technicians. Figure B3a shows Mahalanobis distance–frequency values, B3b shows Mahalanobis distance–importance values, B3c shows person-total correlation–frequency values, B3d shows person-total correlation–importance values, B3e shows long string–frequency values, and B3f shows long string–importance values.

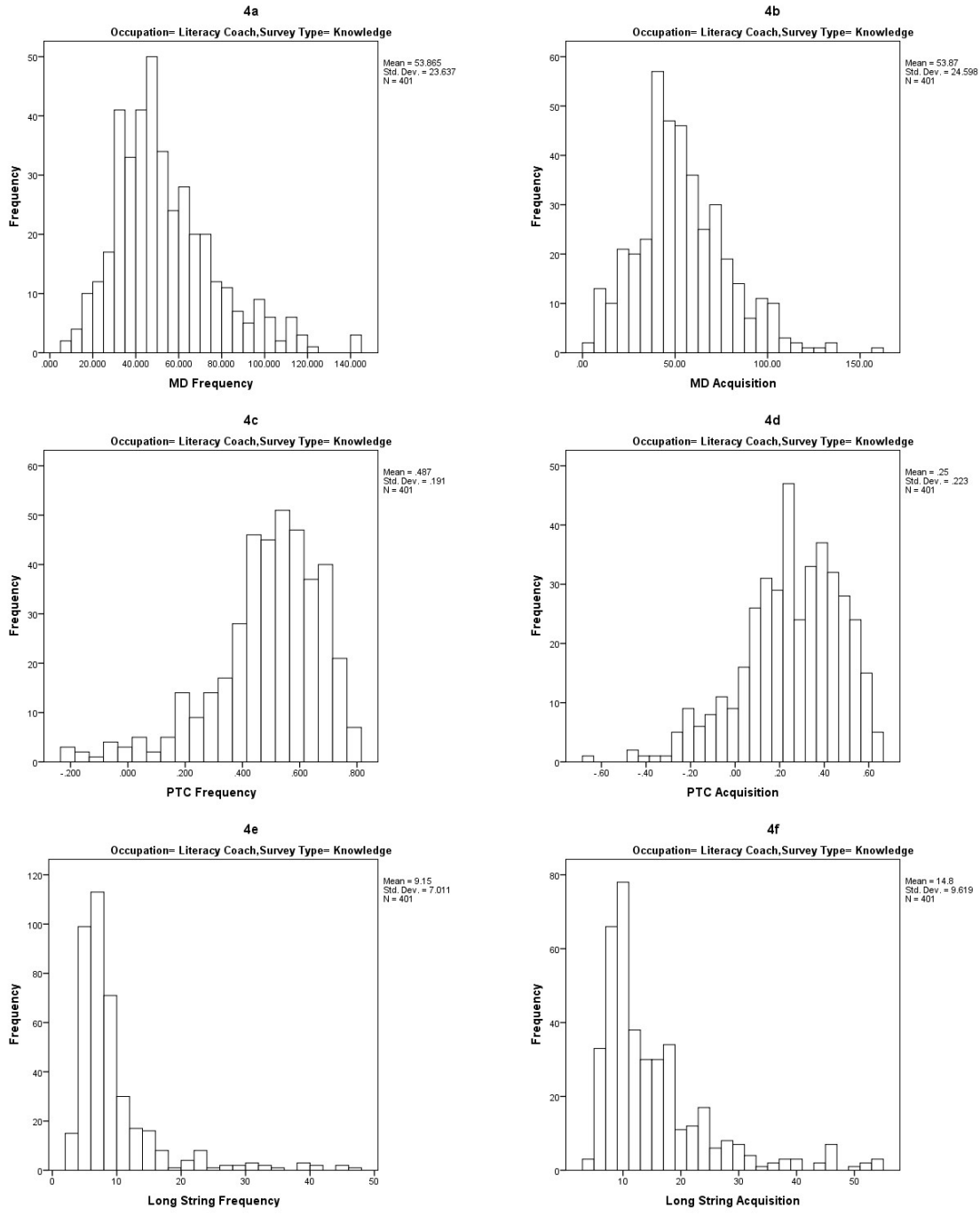


Figure B4. Histograms of knowledge CI index values for literacy coaches. Figure B4a shows Mahalanobis distance–frequency values, B4b shows Mahalanobis distance–acquisition values, B4c shows person-total correlation–frequency values, B4d shows person-total correlation–acquisition values, B4e shows long string–frequency values, and B4f shows long string–acquisition values.

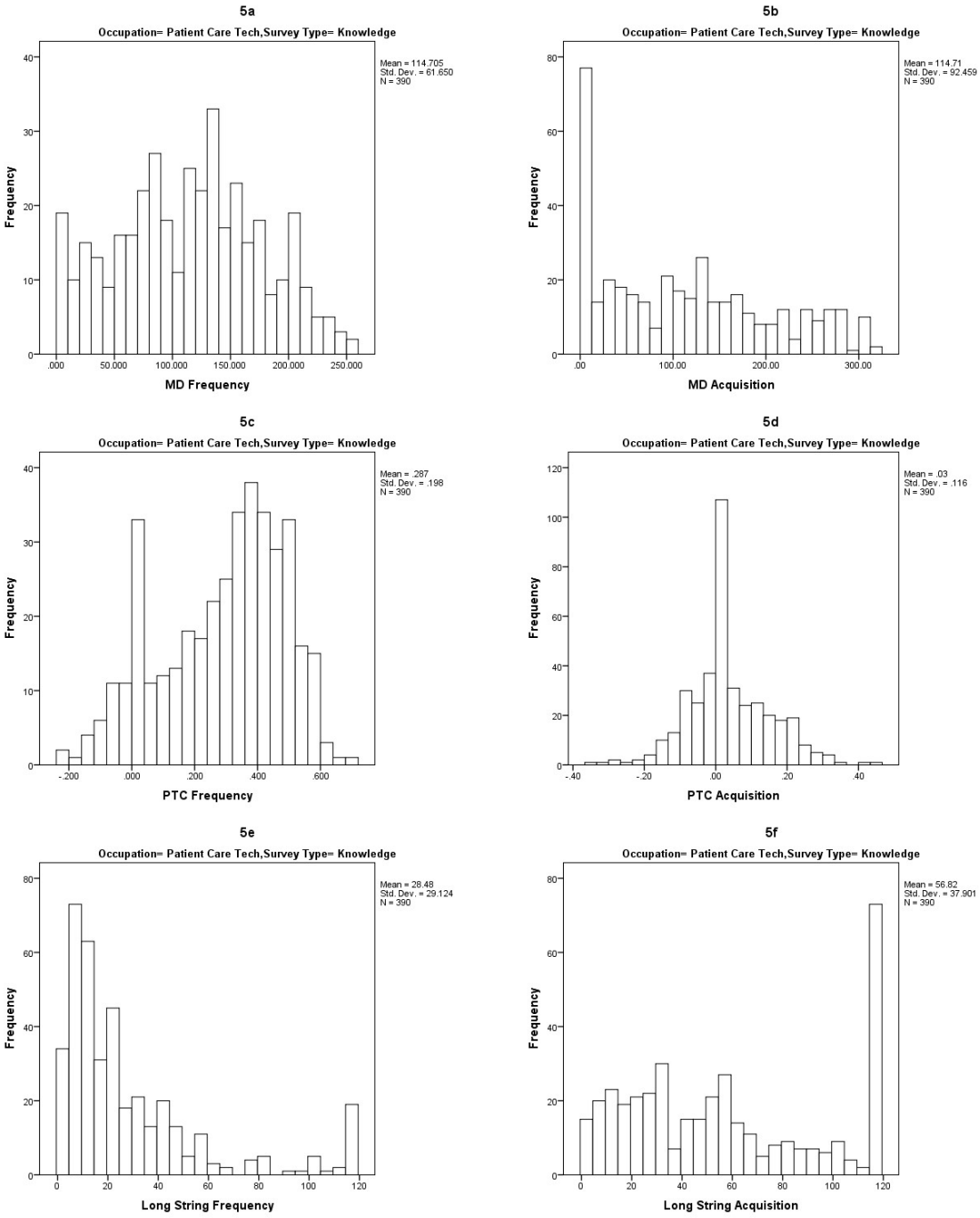


Figure B5. Histograms of knowledge CI index values for patient care technicians. Figure B5a shows Mahalanobis distance–frequency values, B5b shows Mahalanobis distance–acquisition values, B5c shows person-total correlation–frequency values, B5d shows person-total correlation–acquisition values, B5e shows long string–frequency values, and B5f shows long string–acquisition values.

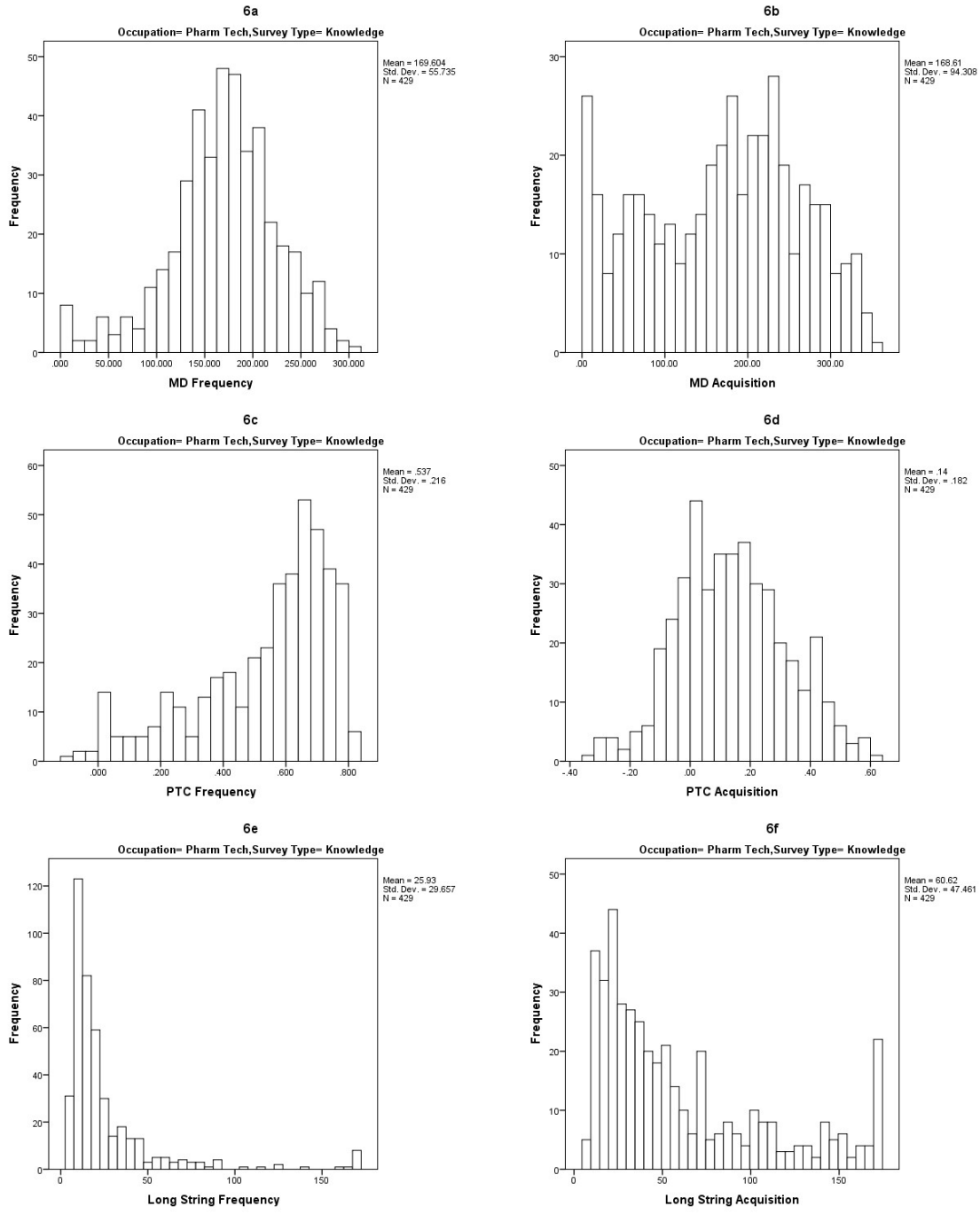


Figure B6. Histograms of knowledge CI index values for pharmacy technicians. Figure B6a shows Mahalanobis distance–frequency values, B6b shows Mahalanobis distance–acquisition values, B6c shows person-total correlation–frequency values, B6d shows person-total correlation–acquisition values, B6e shows long string–frequency values, and B6f shows long string–acquisition values.