DETERMINING THE QUALITY OF OCCUPATIONAL ANALYSIS DATA

Technical Report No. 1

Prepared by:
Patrick Rolwes
Daniel Nagle-Pinkham
Antonio Gonzalez III
Ariana Cunningham
Luke Kachelmeier
Dylan Willis CTR

Prepared for:
HQ AETC A9/SAS OA
Randolph AFB, TX 78150

November 2022

Approved for public release; Distribution unlimited

Summary

This report details the development and validation of the Occupational Analysis (OA) Flight's process to ensure the quality of data for occupational analyses. Assessing the quality of OA data will help ensure the data is appropriate and improve outcomes of the data itself. The development process included a detailed literature review on current research and a test of the proposed method for ensuring accuracy with simulated data. This effort resulted in the creation of a research-based method for assessing the quality of OA data.

Acknowledgements: Special thanks to LinQuest Team for their assistance and expertise in running simulations.

Background

"It's our people that remain our most competitive advantage over any adversary...
It's not the F-35, B-21, or any other platform. CMSgt BASS

Effective and efficient training is the key to making sure the Air Force has a competitive advantage over any adversary. Occupational analysis data supports key decisions on what to train, and more importantly from a risk perspective, what not to train. Therefore, the quality of OAs is critical to ensuring career fields across the Air Force have the data they need to make informed decisions.

To ensure the quality of OA data, the OA Flight conducted a study on rigorous techniques for filtering data, understanding the impact of the length of surveys on data quality, and metrics that should be used to demonstrate the quality of OA data. This study and technical report were conducted to address these areas and outline a plan for ensuring that OA flight members provide high quality data.

The OA flight members currently conduct occupational analyses for most enlisted Air Force career fields. OA also conducts occupational analyses for officer and civilian career fields, by request. The frequency of occupational analyses for any given career field varies dramatically from every 3 years to every 10+ years. The frequency of study is loosely based on the speed of change within the career field, but can also be influenced by how often a career field requests a study.

The purpose of occupational analysis is to collect task-level data to validate Air Force training. OA data informs career fields on which duty areas and tasks require more or less training focus than they are currently being given. Additionally, members of the Studies and Analysis Squadron Airman Advancement Flight (SAS/AA) use OA's data to inform how they write relevant and valid questions on enlisted promotion tests. AFPC/DSYX has also used OA data to inform ASVAB benchmarking.

We first identified four research questions:

1. How do behavioral scientists distinguish "good" job analysis data from "bad" job analysis data?
2. How should OA distinguish "good" survey data from "bad" survey data?
3. How should OA remove "bad" survey data from occupational analyses?
4. How should OA display data to ensure that (non-scientist) customers use it appropriately?

We begin by summarizing the most relevant findings of our literature review. Next, we describe the metrics, which we deemed most suitable for evaluating accuracy and thresholds for filtering data. Third, we describe the process of testing the identified metrics and thresholds on recently collected occupational analysis data. Experts in academic, applied, and military settings were consulted to gather feedback on our proposed process. Finally, we explain methodology related issues which were identified over the course of our research and discuss relevant literature related to these issues. Implementing the proposed accuracy standards will ensure that OA is providing high quality data and should greatly improve the confidence in resulting products.

Literature Review

Over 150 peer-reviewed articles relevant to the quality of occupational analysis data were reviewed. Special emphasis was placed on articles presenting theoretical frameworks for evaluating occupational analysis data quality and meta-analytic reviews of occupational analytic data quality. The most relevant of these articles are detailed in the following section.

Definitions

The following definitions were an important baseline for this research effort, as the members of the research team come from multiple, different disciplines. For the purposes of this study, accuracy refers to how close the measured value of some variable comes to the "true value." For example, how close is occupational analysis data to the actual job being performed by airmen. Validity refers to whether a researcher is measuring what is intended to be measured. Reliability refers to the likelihood that similar results will be observed if a study is repeated multiple times. Precision is defined as the variance within the responses, with lower variance equating to higher precision.

Throughout this report, we use the term "data quality" to refer to our degree of confidence in the data based on its' accuracy, validity, and reliability.

Job Analysis Accuracy Theory

Though evaluating accuracy is a routine practice in most areas of psychology research, it poses some unique challenges for job analysis. A "job" is not a static concept, making it difficult to define a single job consistently. A job differs from person to person, organizations, and across time. Given this conceptualization of a job, the assumption of Classical Test Theory that there is a "true score" does not hold and variance in the data cannot be assumed to be due to measurement error (Morgeson & Campion, 1997; Sanchez & Levine, 2000; Harvey & Wilson, 2000).

If there is no true score for a job, how can one measure the accuracy of job analysis data? Sanchez and Levine (2000) suggest assessing the quality of job analysis data based on the quality of the outcomes of the data. In practice, this theory would involve assessing the quality of the processes and systems developed using the data, thus establishing consequential validity. However, this is a suboptimal approach. First, if consequential validity is the only means of verifying the accuracy of job analysis data, then it is impossible to assess job analysis data that has not been used to develop personnel processes. Thus, in initial development of personnel systems, blind trust in the data would be necessary. When considering OA data, assessing consequential validity would require a great deal of resources and increased workload, which OA is not currently equipped to handle. Therefore, it is infeasible for OA to rely on consequential validity.

Second, Sanchez and Levine (2000) assume that job analysis data is completely relative to the individual performing the job, and therefore cannot be measured independently. Harvey and Wilson (2000) strongly disagree with this assertion, claiming there is an objective reality to job analysis data. Specifically, Harvey and Wilson (2000) distinguish between work

characteristics and work specifications. Work characteristics are the description of what is done on the job, whereas work specifications are the inferred requirements needed to perform those work characteristics. Harvey and Wilson (2000) contend that Subject Matter Experts (SMEs) can objectively evaluate work characteristics that are easily observable and verifiable.

Morgeson and Campion (1997) propose a solution to assessing job analysis data accuracy. Instead of using a single metric to assess accuracy, Morgeson and Campion (1997) suggest indexing the accuracy of job analysis data using a variety of accuracy metrics. More specifically, Morgeson and Campion (1997) recommend using interrater reliability, interrater agreement, discriminability between jobs, dimensionality of factor structures, mean ratings, and completeness of information to assess accuracy in a more holistic way.

Here, we adopt Morgeson and Campion's (1997) theory of assessing the quality of job analysis data. This is a holistic view which addresses concerns about assessing accuracy when there is no "true score" in the traditional sense. Next, we will shift our focus to the metrics which we will use to assess accuracy.

Interrater Reliability

Interrater reliability is defined as consistency across raters and it indexes covariation between raters (Shrout & Fleiss, 1979). Interrater reliability is the most commonly used metric to assess job analysis accuracy (e.g., Duvernet et al., 2015). Researchers typically use interrater reliability to ensure that respondents with the same job are answering similarly to one another. Research suggests that interrater reliability is typically highest for task level data versus more broad data such as skills or competencies (Duvernet et al., 2015) and higher when raters are experts (Duvernet et al., 2015; Voskuijl & Sliedregt, 2002). Additionally, frequency and importance scales tend to have higher interrater reliability (Duvernet et al., 2015).

There are two main ways to measure interrater reliability: a simple Pearson Correlation or Intraclass Correlation Coefficient (ICC). ICC is best suited for data structured as groups rather than paired observations. Because ICC uses a pooled mean and standard deviation, it is more nuanced than Pearson Correlation. For example, if Respondent A answers 1, 2, and 3 on a Likert Scale and Respondent B answers 3, 4, and 5 on the same questions, the Pearson Correlation would be a 1 because the variance is identical. In contrast, ICC detects the difference between the two patterns of responding. Thus, we will use ICC rather than Pearson Correlations because it provides a series of benefits for the type of data we are assessing.

Between Job Discriminability

Between job discriminability refers to the extent to which raters can distinguish between different jobs. The goal is to ensure that the raters can distinguish between jobs. Between job discriminability can be measured using correlations between jobs to measure between job variances. Similar jobs should have higher correlations and vice versa. By measuring between job discriminability, discriminant validity can be established. OA is not currently capable of measuring between job discriminability, as there is little to no overlap of tasks across career

fields for comparison purposes. However, between job discriminability is an important consideration going forward, as OA looks to evaluate across career fields.

Mean Ratings

Mean ratings allow a researcher to identify responses that are significantly elevated or depressed relative to other ratings. Such ratings are an indicator of rater bias.

Completeness

Completeness refers to ensuring that the data reflects the entirety of the job that is being measured. Completeness in job analysis data is typically established by asking multiple sources (e.g. incumbents, supervisors, and job experts) if the entirety of the job is covered by the measure. OA establishes this completeness by conducting subject matter expert (SME) visits in conjunction with surveys of a representative sample of incumbents.

Other Accuracy Metrics

Other metrics discussed were intrarater reliability, intrarater agreement, dimensionality of factor structure, and item discriminability. All of these metrics were worth consideration for indexing the accuracy of OA data, given their use in past research.

Methods

We hypothesize that the accuracy of job analysis data is dependent on the rigor of the study design and the accuracy of the respondents. Rigor of the study design is critical to ensure that the right questions are being asked to reach the targeted information. Accuracy of the respondents ensures that the answers to the questions are appropriate. While study design is important, the research team determined that OA's current methodology was sufficient for its intended purpose. However, there are many potential ways to improve and modernize the OA methodology. Therefore, potential suggestions for study design are included at the end of this report. The main concern for OA in determining the accuracy of OA data is the accuracy of respondents. The following outlines the proposed method for identifying and controlling for the accuracy of respondents.

Metric Determination

As discussed above, we identified the most relevant potential sources of inaccuracy in our data: conformity bias, loss of motivation, impression management, social desirability, demand effects, information overload, heuristics, categorization, carelessness, halo effect, leniency and severity effect, and method effects. Based on the relevant sources of inaccuracy, accuracy metrics were derived to identify and address these sources. The metrics used to identify inaccuracy were interrater reliability, careless responder metrics, and mean ratings. Many of these potential sources of inaccuracy were based on the respondents' ability and intent to answer items accurately and without distraction. Therefore, in evaluating the quality of OA data it was determined that identifying potential careless responders was a high priority, due to our samples likely experiencing significant survey fatigue and task overload and having no incentive to complete the surveys.

Interrater Reliability

Given the prevalence of the use of interrater reliability in the job analysis literature for evaluating the quality of job analysis data, interrater reliability was determined to be an essential metric (Duvernet et al., 2015). Historical OA documents indicate that traditionally OA had measured interrater reliability, but for unknown reasons the metric was dropped at some point. Reestablishment of interrater reliability is critical in OA's effort to reestablish the quality of OA data. ICC was the chosen metric for measuring interrater reliability, specifically ICC two-way random effects consistency with multiple raters (Koo & Li, 2016). In interpreting ICC, several factors need to be considered. ICC has been shown to be impacted by such factors as specificity of tasks, type of rating scale, respondent type, intended use, number of respondents, number of items, and breadth of the specific job (Duvernet et al., 2015). While there is variation across career fields, a specific cutoff was desired for standardization and ease of interpretation. In considering these factors and how they relate to OA data, in conjunction with research on ICC interpretations (Shrout & Fleiss, 1979) and meta-analytic results in job analysis data (Duvernet et al., 2015; Johnson et al., 1981; Voskuijl1 & Sliedregt, 2002), the research team determined an appropriate ICC for OA was 0.70 or greater. Given the large number of items in some OA surveys, the sensitivity of ICC to the number of items in a survey was a concern. In response, the research team devised a method to control for the number of items, in order to achieve a more accurate and comparable representation of ICC. For this method, 50 items are randomly selected from the survey and ICC is run on these items. This procedure is repeated 9 more times. If the ICC is greater than or equal to 0.70 for greater than 50% of the trials, then the ICC is considered to be acceptable.

Careless Responder Metrics

In order to ensure the OA data consists of accurate responses across a career field, it is essential to identify careless responders before measuring the ICC. In survey literature there are several metrics used to identify careless responders, including over and under achiever metrics, intraindividual response variance (IRV), outliers, speeders, and test questions. The purpose of the over and under achiever metrics is to identify participants who claimed to do an unreasonably small amount of task or large number of tasks. Such participants are likely not exercising appropriate attention in their responses or are not the intended recipients of the survey. IRV measures the variability of an individual's responses. Low variability is an indication of straight lining and high variability is an indication of random responding (Dunn et al., 2018). The outlier metric compares an individual's responses to the responses of other participants. If responses are significantly different from the average, it is an indicator of insufficient effort responding (Meade and Craig, 2012). However, given the fact that in OA data outliers may just be an indication of someone who performs a unique job within the career field, the outlier metric is only considered an indication of careless responding in conjunction with additional metrics, as will be explained in greater detail. The speeder metric measures how many seconds it takes participants to complete each item on the survey. Research has indicated a relationship between time spent on a survey and response effort (Wise & Kong, 2005). If the participant does not spend sufficient time to read and process the survey items, it is a strong indication of careless

responding. The test questions are directed questions in the survey which periodically ask participants to select a particular answer. For OA surveys, having a test question for every 50 tasks is our recommended number of test questions. Participants who fail to correctly answer a significant number of these items are likely not exhibiting sufficient effort to include their responses in the results.

Two-Pass System and Thresholds

Once the metrics for identifying careless responders were identified, cutoffs were determined for when to classify a respondent as a careless responder and a method was devised for removal of these responders. Particular concern was given to minimizing the possibility of identifying legitimate responders as careless responders. This concern resulted in the proposed two-pass system. In the two-pass system, respondents are evaluated on each of the careless responder metrics twice. In the first pass, careless responder cutoff thresholds are very conservative, targeting the most egregious careless responders. In the second pass, the cutoff thresholds are more liberal, but respondents must be considered a careless responder in multiple metrics in order to be removed. It is believed that this two-pass system minimizes the risk of removing good responses, while ensuring the majority of careless responders are removed.

Thresholds were determined based on a couple of criteria. First, relevant psychology research was used to obtain and support thresholds. Where research was scarce or conflicting, the research team considered what thresholds would minimize the risk of identifying legitimate responders as careless responders while still capturing the majority of careless responders. Several tests were run using simulated data to ensure the thresholds were appropriate and did not remove too much data. Table 1 displays the careless responder metrics and the proposed thresholds.

Table 1: Careless Responder Metrics and Thresholds

Area of Identification	Metric	1st Pass Threshold	2nd Pass Threshold
Metric Block	Under-Achiever	<SME Input - 5%	N/A
	Over-Achiever	>SME Input + 5%	N/A
	IRV (CCCer)	>2 Standard Deviations	>1.5 Standard Deviations
	Outlier	N/A	>99% Chi Square
Duty Area	Speeder	<1	<2
	Test Question	<50% Right	<70% Right

For the achiever metrics, what constitutes an over- or under-achiever is highly dependent on the career field. Therefore, subject matter expertise from individual career fields is necessary to determine the correct thresholds. The proposed thresholds are based on SME input adjusted by 5% in order to ensure good responders are not incorrectly identified. There is no second pass

threshold for the achiever metrics. IRV is identified by greater than 2 standard deviations from the mean for the first pass and 1.5 standard deviations from the mean for the second pass. As mentioned previously, the outlier metric is not sufficient on its own to justify identifying someone as a careless responder, as there is a possibility of the responder having a unique job within the career field. Therefore, the outlier metric is only applied in the second pass. In addition, Mahalanobis distance, the primary metric used to measure outliers in many studies, was not considered appropriate for OA data, given the large holes often found in OA data matrices. As an alternative to Mahalanobis distance, a sum of squares outlier metric was proposed. The threshold to be considered an outlier was the 99^{th} percentile in a chi-square distribution with degrees of freedom equal to the number of tasks the respondent has a non-zero answer. For the speeder metric, respondents were identified as careless responders if they took less than 1 second to respond per task on the first pass and less than 2 seconds to respond per task on the second pass. This threshold is based on the suggested threshold of Huang et al. (2012). The direct response questions are an objective metric. However, human error in responding to the question is still a possibility (Meade & Craig, 2012). Therefore, conservative thresholds of greater than 50% of test questions answered correctly for the first pass and greater than 70% of test questions answered correctly for the second pass were proposed. OA surveys present a respondent with a list of tasks which are organized by duty area, and these items are evaluated by the respondents with three different metrics (task frequency, task criticality, or task learning difficulty). The careless responders metrics are built around the organization of the OA surveys. Careless responders were identified by metric block for the achiever metrics, the IRV metric, and the outlier metric. Careless responders were identified by the duty area level for the speeder and test questions metrics. After identifying a careless responder, they should be removed at the metric block level, that is a respondent's answers could not be removed for one duty area but not another.

Testing

In order to test these metrics and thresholds, an experiment was conducted using simulated careless responder data. The simulated data consisted of various types of careless responders, based on the research and researcher experience. The following are the different types of careless responders simulated:

- Random responders have an equally likely chance of answering the potential responses for a task
- {3,4} respondents randomly answer options "3" and "4" only
- 95% 4 respondents answer "4" 95% of the time and randomly choose for the remaining 5%
- Xmas Tree refers to the traditional Christmas tree pattern from scantron exams where the respondents answer the first option in sequence to the last option then last to first in reverse

• Long String (LS) Switcher respondents provide the same response for a long stretch of tasks (randomly determined between 10 and 50) and then switch to a different response value and repeat it for a long stretch of tasks (again, random number between 10 and 50)

• Switcher respondents simulate responders who may pay attention to a survey for a random number of tasks (random number chosen between 10 and 50) and then switch to low variance responses (respond with "3" or "4" for a random stretch of 10 to 50 tasks). These careless responders then switch back to "good responses" for a random switch, alternating between these two conditions until the end of the metric block.

In addition, for each type of respondent, the speed and number of test questions answered correct were varied. Table 2 displays the various different combinations utilized for this study.

Table 2: Combinations of Simulated Careless Responder Data

Combination	Type	TQ	Speed		Combination	Type	TQ	Speed
A	Random	0.25	Fast		**J**	Xmas Tree	0.25	Fast
B	Random	0.6	Medium		**K**	Xmas Tree	0.6	Mediu m
C	Random	0.95	Slow		**L**	Xmas Tree	0.95	Slow
D	{3,4}	0.25	Medium		**M**	LS Switcher	0.25	Mediu m
E	{3,4}	0.6	Slow		**N**	LS Switcher	0.6	Slow
F	{3,4}	0.95	Fast		**O**	LS Switcher	0.95	Fast
G	95% 4	0.25	Slow		**P**	Switcher	0.25	Slow
H	95% 4	0.6	Fast		**Q**	Switcher	0.6	Fast
I	95% 4	0.95	Medium		**R**	Switcher	0.95	Mediu m

Data from two OA surveys were used in this study: 2A9X3 (Bomber/Special Electronic Warfare and Radar Surveillance Integrated Avionics) and 3E5X1 (Engineering) Air Force Specialty Codes (AFSCs). These surveys were selected because they had time metrics available for analysis. The simulated careless responder data was added to these datasets.

Results and Discussion

Tables 3 and 4 display the results of the simulated data. The proportion of real respondents identified as careless responders was also included. While there is some variation between the career fields, the interpretation of the results are largely the same. The proposed careless responder metrics identified between 4%-22% of real respondents as careless responders. This is consistent with research on careless responding, which has found between 10-20% of real respondents exhibiting careless responder behaviors (Meade & Craig, 2012; Huang et al., 2015).

Overall, the proposed thresholds were able to identify a majority of the simulated data. However, there were several careless responder types that were particularly difficult to detect. The random responders who answered test questions correctly and at a medium or slow pace

were more difficult to detect. The risk of such responders significantly impacting the data was considered low enough to not be a major concern. Simulated responders who responded in a patterned Xmas tree response who answered test questions correctly and at a medium or slow pace were also more difficult to detect. Again, the risk of such careless responders significantly impacting the data was considered low enough to not be a major concern. However, further research into how to detect such responders may be worthwhile. The other particularly difficult careless responders to detect were either type of switcher who answered test questions correctly and at a medium or slow pace. These types of respondents may warrant further discussion, as it is reasonable for certain responders who are conscientiously completing the survey, but who do not perform a large number of tasks to suddenly "switch" when tasks they do perform appear. Improved survey design can help to mitigate the risk of such a response pattern from a conscientious responder. For example, by limiting the tasks rated to only the duty areas the respondent performs, such a pattern becomes less likely.

The proposed thresholds identified a higher proportion of simulated data for the Task Frequency metric block than the Task Criticality or Task Learning Difficulty metric blocks. These results may be due to the sample size differences between the blocks, as the simulated data consisted of a larger proportion of the data in the Task Criticality and Task Learning Difficulty metric blocks. For the Task Criticality and Task Learning Difficulty metric blocks, the 95% 4 respondents who answered test questions correctly and at a medium pace were largely undetected.

Table 3: 2A9X3 Proportion of Careless Responder Simulated Data Flagged

Careless Responder/ Data Type	Prop TF Pass 1	Prop TF Pass 2	Prop Tagged	Prop TC Pass 1	Prop TC Pass 2	Prop Tagged	Prop TLD Pass 1	Prop TLD Pass 2	Prop Tagged
A Rand 0.25 Fast	82%	100%	100%	76%	94%	99%	86%	100%	100%
B Rand 0.6 Med	18%	38%	49%	15%	22%	34%	16%	40%	50%
C Rand 0.95 Slow	0%	2%	2%	0%	1%	1%	0%	2%	2%
D {3,4} 0.25 Med	100%	NA	100%	74%	83%	96%	74%	59%	90%
E {3,4} 0.6 Slow	100%	NA	100%	15%	46%	54%	16%	30%	42%
F {3,4} 0.95 Fast	100%	NA	100%	14%	97%	98%	14%	63%	69%
G {95% 4} 0.25 Slow	100%	NA	100%	77%	79%	95%	76%	66%	92%
H {95% 4} 0.6 Fast	100%	NA	100%	25%	99%	99%	29%	93%	95%
I {95% 4} 0.95 Med	100%	NA	100%	1%	1%	2%	0%	2%	2%
J Xmas Tree 0.25 Fast	78%	100%	100%	86%	73%	96%	75%	100%	100%
K Xmas Tree 0.6 Med	17%	46%	55%	18%	2%	19%	16%	42%	51%
L Xmas Tree 0.95 Slow	0%	2%	2%	0%	0%	0%	1%	2%	2%
M LS Switcher 0.25 Med	80%	71%	94%	69%	42%	82%	71%	69%	91%
N LS Switcher 0.6 Slow	48%	35%	66%	17%	22%	35%	14%	34%	43%
O LS Switcher 0.95 Fast	44%	77%	87%	17%	49%	58%	11%	82%	84%
P Switcher 0.25 Slow	72%	54%	87%	73%	0%	73%	71%	35%	81%
Q Switcher 0.6 Fast	38%	73%	83%	34%	33%	55%	30%	74%	82%
R Switcher 0.95 Med	9%	2%	11%	0%	0%	0%	0%	0%	0%
Real Respondents	16%	3%	18%	13%	11%	22%	10%	8%	18%

Green highlight indicates greater than 80% of simulated careless responders identified. Yellow highlight indicates less than 20% of simulated careless responders identified.

Table 4: 3E5X1 Proportion of Careless Responder Simulated Data Flagged

Careless Responder/ Data Type	Prop TF Pass 1	Prop TF Pass 2	Prop Tagged	Prop TC Pass 1	Prop TC Pass 2	Prop Tagged	Prop TLD Pass 1	Prop TLD Pass 2	Prop Tagged
A Rand 0.25 Fast	93%	100%	100%	94%	100%	100%	93%	100%	100%
B Rand 0.6 Med	28%	35%	53%	33%	27%	51%	30%	51%	65%
C Rand 0.95 Slow	0%	1%	1%	0%	0%	0%	0%	0%	0%
D {3,4} 0.25 Med	95%	90%	100%	92%	89%	99%	94%	88%	99%
E {3,4} 0.6 Slow	30%	40%	58%	28%	46%	62%	30%	35%	54%
F {3,4} 0.95 Fast	8%	100%	100%	11%	100%	100%	18%	100%	100%
G {95% 4} 0.25 Slow	100%	NA	100%	96%	80%	99%	99%	100%	100%
H {95% 4} 0.6 Fast	100%	NA	100%	63%	100%	100%	62%	100%	100%
I {95% 4} 0.95 Med	100%	NA	100%	36%	0%	36%	51%	0%	51%
J Xmas Tree 0.25 Fast	93%	66%	98%	93%	88%	99%	92%	100%	100%
K Xmas Tree 0.6 Med	30%	1%	30%	29%	15%	40%	22%	40%	53%
L Xmas Tree 0.95 Slow	0%	0%	0%	0%	0%	0%	0%	1%	1%
M LS Switcher 0.25 Med	94%	59%	98%	96%	20%	97%	92%	56%	97%
N LS Switcher 0.6 Slow	47%	23%	59%	30%	25%	47%	22%	33%	48%
O LS Switcher 0.95 Fast	36%	51%	68%	8%	54%	58%	11%	87%	89%
P Switcher 0.25 Slow	93%	0%	93%	96%	0%	96%	94%	0%	94%
Q Switcher 0.6 Fast	41%	50%	71%	41%	46%	69%	30%	61%	73%
R Switcher 0.95 Med	11%	0%	11%	0%	0%	0%	0%	0%	0%
Real Respondents	20%	3%	22%	10%	7%	17%	2%	3%	4%

Green highlight indicates greater than 80% of simulated careless responders identified. Yellow highlight indicates less than 20% of simulated careless responders identified.

The results from this study largely support the proposed metrics and thresholds and demonstrate their effectiveness at identifying careless responders. It appears unlikely with these metrics that any careless responders who are not identified would significantly impact the results of the data. Such a careless responder would need to answer test questions correctly and take the survey at a reasonable speed, which is likely to be a small number of careless responders.

Future Research

While this effort can serve as a starting point, additional follow-up is necessary to validate and adjust the OA data quality process. Future research should further validate the proposed data quality process, and ensure it functions as intended across career fields and in all circumstances. Such research should include a comparison of various thresholds to ensure the chosen thresholds are optimal. Examination of the quality of data from various demographics may give further insights into what factors influence data quality and what populations to focus on. OA should also test if there is a significant difference between the results before and after removal of careless responders, in order to establish the severity of the problem, and infer the level of importance of removing careless responders. Future research should also further

examine if flagged careless responders are objectively careless. Such a study could take the form of interviews of survey responders to gain insight into responder motivations while taking the survey. Additionally, further research into how survey design impacts the quality of OA data would provide new insights and offers an opportunity to further improve the quality of data (see Appendix).

Conclusion

After review of the relevant literature and initial testing, a process has been developed to evaluate the level of data quality for OA and ensure accuracy is improved by removing careless responders. In the job analysis literature, the quality of data is dependent on the quality of the methodology used and by the respondents' ability to answer accurately (Harvey & Wilson, 2000). In addition, the apparent leading theory on job analysis accuracy proposes the use of multiple accuracy metrics to index the quality of data. Therefore, the proposed data quality process for OA's job analyses uses multiple careless responder metrics, in conjunction with interrater reliability, to evaluate and ensure the quality of data. An initial test using two datasets and simulated careless responder data found that the careless responder metrics were able to capture a significant number of potential careless responder types. Future research should further validate this process with more data and test various other thresholds and metrics. Overall, the proposed data quality process was found to be sufficient for evaluating the quality of OA data.

References

Chatfield, R. E., & Royle, M. H. (1983). Methods to Improve Task Inventory Construction. NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN DIEGO CA.

Chipperfield J.O, & Steel D.G (2009). Design and estimation for split questionnaire surveys. *J Office Stat, 25*, 227-244.

Dierdorff, E. C., & Wilson, M. A. (2003). A meta-analysis of job analysis reliability. Journal of Applied Psychology, 88(4), 635.

Diehr P, Chen L, Patrick D, et al. (2005). Reliability, effect size, and responsiveness of health status measures in the design of randomized and cluster-randomized trials. *Contemp Clin Trials*, *26*, 45-58.

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. Journal of Business and Psychology, 33(1), 105-121

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.

Fleishman, E. A., & Mumford, M. D. (1991). Evaluating classifications of job behavior: A construct validation of the ability requirement scales. Personnel Psychology, 44(3), 523-575.

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*, 330-338.

Harvey, R. J., & Wilson, M. A. (2000). Yes Virginia, there is an objective reality in job analysis. Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 21(7), 829-854.

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.

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.

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.

Johnson, J. W. (2000). Factor analysis of importance ratings in job analysis: Note on the misinterpretation of Cranny and Doherty (1988). Organizational Research Methods, 3(3), 267-284.

Kerlinger, F. (1964). Foundations of behavioural research. New York: Holt

Kost, R. G., & da Rosa, J. C. (2018). Impact of survey length and compensation on validity, reliability, and sample characteristics for Ultrashort-, Short-, and Long-Research Participant Perception Surveys. *Journal of clinical and translational science*, *2*, 31-37.

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

Manson, T. M., Levine, E. L., & Brannick, M. T. (2000). The construct validity of task inventory ratings: A multitrait-multimethod analysis. Human Performance, 13(1), 1-22.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychol Methods*, *1*, 30-46.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. Psychological methods, 17(3), 437.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. Journal of applied psychology, 82(5), 627.

Morgeson, F. P., & Campion, M. A. (2000). Accuracy in job analysis: Toward an inference-based model. Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 21(7), 819-827.

Morgeson, F. P., & Dierdorff, E. C. (2011). Work analysis: From technique to theory. In APA handbook of industrial and organizational psychology, Vol 2: Selecting and developing members for the organization. (pp. 3-41). American Psychological Association.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., & Levin, K. Y. (1997). O* NET Final Technical Report.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Levin, K. Y., ... & Dye, D. M. (2001). Understanding work using the Occupational Information Network (O* NET): Implications for practice and research. Personnel psychology, 54(2), 451-492.

Raghunathan T.E, & Grizzle J.E (1995). A split questionnaire survey design. *J Am Stat Assoc. 90*, 54-63.

Sanchez, J. I., & Levine, E. L. (2000). Accuracy or consequential validity: which is the better standard for job analysis data?. Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 21(7), 809-818.

Schippmann, J (2000). Strategic job modeling.

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

Turner, M. (2019, February 13). Lost in translation: Verbally communicating reliability and validity evidence. SIOP. Retrieved July 6, 2022, from https://www.siop.org/Research-Publications/TIP/TIP-Back-Issues/2017/October/ArtMID/20295/ArticleID/1470/lost

Wiggins, J. S. (1973). Personality and prediction: Principles of personality assessment. Addison-Wesley Publishing Company.

Wise, S. L., & Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. Applied Measurement in Education, 18, 163–183.