Practical Assessment, Research, and Evaluation

Volume 23 *2018* Article 10

2018

# A Comparison of Subject Matter Experts' Perceptions and Job Analysis Surveys

Adam E. Wyse

Ben Babcock

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.



Volume 23 Number 10, September 2018 ISSN 1531-7714

# A Comparison of Subject Matter Experts' Perceptions and Job Analysis Surveys

Adam E. Wyse, *Renaissance*
Ben Babcock, *The American Registry of Radiologic Technologists*

Two common approaches for performing job analysis in credentialing programs are committee-based methods, which rely solely on subject matter experts' judgments, and task inventory surveys. This study evaluates how well subject matter experts' perceptions coincide with task inventory survey results for three credentialing programs. Results suggest that subject matter expert ratings differ in systematic ways from task inventory survey results and that task lists generated based solely on subject matter experts' intuitions generally lead to narrower task lists. Results also indicated that there can be key differences for procedures and non-procedures, with subject matter experts' judgments often tending to exhibit lower agreement levels with task inventory survey results for procedures than for non-procedures. We recommend that organizations performing job analyses think very carefully before relying solely on subject matter experts' judgments as their primary method of job analysis.

An essential component of developing and maintaining credentialing programs is performing job analysis. Both the standards created by the National Commission for Certifying Agencies (NCCA) (Institute for Credentialing Excellence, 2016; Knapp, Anderson, & Wild, 2015) and the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) stress the importance of job analysis for credentialing programs. Job analysis provides the foundation for exam content and other credentialing program requirements by linking content and requirements to what people do in the work place. There are several methods for performing job analysis, including the task inventory method (Gael, 1983), the critical incidence technique (Flanagan, 1954), the professional practice model (LaDuca, 1980), and committee-based methods that rely solely on the expertise of subject matter experts (SMEs) (Raymond, 2001). Different methods have various strengths and weaknesses, and specific professions are often better suited for the application of particular methods.

Two of the more commonly used methods are the task inventory and committee-based methods. In the task inventory method, one compiles a list of job tasks and formats them into a survey with one or more rating scales (Raymond, 2001, 2005, 2016; Wang, Schnipke, & Witt, 2005; Wang & Stahl, 2012; Wyse, Eckerly, Babcock, & Anderson, 2016). Discussions with a group of SMEs inform the list of tasks and which survey scales to use in the survey. Common task survey scales include frequency, criticality, responsibility, importance, difficulty of learning, and need at entry scales (see Raymond, 2001, 2005, 2016). The decision of what tasks should be used to define exam content and other requirements is usually based on a combination of comparing the survey results to a numerical threshold and the judgments of SMEs. For example, a task may be included if 40% of people or more reported responsibility for it with exceptions made for some tasks that are low responsibility but viewed as critical components of the job. In committee-based methods, SMEs are usually asked to come to consensus on the tasks used to define exam content and other requirements (Raymond, 2001). The committee may still be asked to estimate the percentage of people responsible for a task or if it is critical, but they usually make their decision based on their knowledge of the

field without considering survey results. Committee-based methods are common when the number of people in the field is small, it is not financially or logistically feasible to send a survey, or if obtaining survey responses is difficult.

An important question is whether committee-based and task inventory methods produce similar results. If the methods give different results, how do the results differ, and are the differences consistent across jobs? There have been four studies which have compared committee-based and survey-based methods. Ash, Levine, Higbee, and Sistrunk (1982) compared mean task ratings on time spent, difficulty, and criticality scales with those from small groups of SMEs and found high correlations between SME ratings and those from survey respondents. O'Leary, Rheinstein, and McCauley (1990) compared importance ratings for a group of five SMEs with job analysis survey responses for administrative professions and found high correlations and high correspondence between ratings above versus below a defined cut point separating important from unimportant ratings. Tannenbaum and Welsey (1993) compared committee-based and survey-based methods using an importance scale for two teacher licensure exams and found high correlations and high correspondence between ratings above versus below the midpoint of the rating scale. Maurer and Tross (2000) compared the two methods for a print advertising job using a relative time spent scale and also found similar results. While these four studies typically found high correspondence between job analysis survey responses and SME ratings, the studies focused mainly on correlations between ratings and performed comparisons when ratings were averaged. Averaging ratings is generally not a preferred analytical technique because it assumes that the rating scales are on an interval scale, when rating scales are typically on an ordinal scale (Wyse, in press; Wang & Stahl, 2012). In addition, looking at correspondence between ratings above the midpoint of the rating scale when ratings have been averaged may not be an informative comparison. It is unclear how to provide a clear interpretation to such an average, and many tasks tend to have distributions of ratings where most of the ratings are above this point. The fact that most of the ratings are above the midpoint can lead to high levels of correspondence between results when in fact there may be differences if other thresholds were used to compare ratings. Some of the other challenges with prior studies include the use of small sample sizes for the task inventory surveys and the committees of SMEs and looking at the correspondence of task inventory results with a single committee of SMEs.

This study makes an important contribution to the research literature by examining how results for the task inventory and committee-based methods compare to each other for three medical imaging credentialing programs. In our analyses, we used large random samples for the task inventory surveys and in two of the three disciplines we collected SME ratings from more than one committee of SMEs to see if patterns were consistent across different groups of SMEs. In addition, we use simple analytical techniques to make our comparisons and do not average responses for the rating scales. We also look at correspondence between results for two numerical thresholds, including using a threshold that corresponds to the point used to decide whether tasks should or should not be used to define exam content and other requirements for the three credentialing programs.

## Data and Methods

To examine the differences between the job analysis methods, we sent task inventory surveys to large random samples of people for three medical imaging professions. Survey recipients self-reported in an earlier survey as having 10 or fewer years of experience in the discipline of interest, working in the profession as their primary discipline of employment, and working full-time. We drew the random samples from a database maintained by the credentialing organization that oversees the credentialing programs. Table 1 contains a summary of the practice analyses' surveys and return results. For the first profession, we sent out 1,000 surveys and 247 usable surveys were returned (24.7% effective return rate). For the second profession, we sent out 5,400 surveys and 1,637 usable surveys were returned (30.3% effective return rate). For the third profession, we sent out 1,500 surveys and 360 usable surveys were returned (24.0% effective return rate). We tested to see if there were any significant differences between the random samples and returned usable samples for the three data sets. We found for the first and third professions that the only significant differences were that the returned sample tended to include a greater percentage of female respondents than the random samples, while for the second profession the returned sample tended to include a greater percentage of females and were somewhat

older than the random samples. The returned samples were otherwise quite representative of the random samples on key job-related characteristics, such as years of experience, educational level, place of employment, and institution size.

levels of experience serving on exam or job analysis committees for each profession. Each committee of SMEs was reviewed and approved by the board that oversees the credentialing programs. The survey for the first profession consisted of 148 job tasks, with the first

**Table 1.** Summary of Survey and Committee Participants

| Practice Analysis Survey Detail | Discipline 1 | Discipline 2 | Discipline 3 |
|---|---|---|---|
| N Surveys Sent | 1,000 | 5,400 | 1,500 |
| N Useable Returned | 247 | 1,637 | 360 |
| Usable Response Rate | 24.7% | 30.3% | 24.0% |
| Number of Tasks | 148 | 123 | 123 |
| Number Imaging Procedure Tasks | 69 | 82 | 84 |
| Number Non-Procedure Tasks | 79 | 41 | 39 |

| Committee Detail | Discipline 1, Committee 1 | Discipline 1, Committee 2 | Discipline 2 | Discipline 3, Committee 1 | Discipline 3, Committee 2 |
|---|---|---|---|---|---|
| N Panelists | 5 | 6 | 7 | 8 | 8 |
| N Technologists | 5 | 5 | 6 | 7 | 6 |
| Mean Years Certified in Discipline* | 24.4 | 32.1 | 18.2 | 10.9 | 12.3 |

*Note: We gathered mean years certified statistics post-hoc based on our technologist certification records database. These statistics were not available for the physician or physicist panelists. The number of years certified is a good proxy for years of experience for Discipline 1, as this certification represents a common entry point into the medical imaging profession. Committee members for Disciplines 2 and 3 likely had additional years of experience in those disciplines, as these certifications are not generally considered as entry into the profession.

The task inventory surveys included a frequency-responsibility scale (Raymond, 1996), where the lowest response category was "not responsible"; the rest of the scale anchors were absolute frequency descriptions of "yearly," "quarterly," "monthly," "weekly," and "daily." The credentialing organization involved has used this survey scale in many previous task inventory surveys, and results using this scale have been compared to an external data source in the past to ensure the validity of the data received (Babcock & Yoes, 2013). Different committees of eight SMEs created each task inventory survey with facilitation by a staff content specialist in the profession and a psychometrician. The committees of SMEs that developed the surveys included seven medical imaging technologists who worked in the discipline of interest and one person who was not a technologist but had extensive experience working with technologists. The non-technologist was a radiologist or medical physicist. The seven technologists had a range of years of experience, job titles, work locations, and

79 job tasks focused on non-imaging procedures and the last 69 job tasks focused on imaging procedures. The survey for the second profession consisted of 123 job tasks, with 41 non-imaging procedures and 82 imaging procedures. The survey for the third profession consisted of 123 job tasks with 39 non-imaging procedures and 84 imaging procedures. The tasks used to define exam content and other credentialing program requirements for the professions at hand are generally those for which 40% or more of people report responsibility (i.e., gave a response of "yearly" or more frequent).

In addition to the task inventory surveys, we also collected SME judgments by asking whether they thought 40% or more of people were responsible for each task before they examined the survey results. Because other organizations may use thresholds other than 40%, we also asked SMEs to make judgments using an 80% threshold. We collected judgments from two

different committees for two of the disciplines, while we were only able to collect judgments from one committee for the third discipline. Table 1 provides a summary of the number of committee members for each discipline, including how many of the committee members were technologists and statistics relating to years certified in the discipline. Like the committees that created the surveys, the SMEs who served on these committees had also been reviewed and approved by the board that oversees the credentialing programs. The non-technologists again were radiologists or medical physicists who had experience working with technologists in the discipline of interest. The committees were, on the whole, quite experienced professionals with one or two intentionally-selected committee members having somewhat less experience. The roles of the committee members (technologists, physicists, and physicians), as well as the experience levels of the panels, were typical of the panelist makeup for the certification organization involved.

We collected each of the judgments after orienting committee members to the process of job analysis and reviewing the previously-developed task inventory surveys. We collected the judgments using a computer application that listed the survey job tasks and asked SMEs to indicate whether they thought each task was above the 40% and 80% thresholds. Following the collection of data, we discussed the ratings with committee members and showed them the task inventory survey results.

We evaluated interrater agreement and between-committee agreement to assess the quality of committees' ratings. We hypothesized that the between-committee agreement, which compared the modal task ratings for each group, would be substantially higher than the interrater agreement between individuals. This would, thus, make the committee consensus ratings the best possible data to compare to the job analysis ratings. To evaluate the similarities and differences between the committee consensus ratings and the job analysis survey results, we compared the SME ratings to the task inventory survey results using percent agreement, percent of tasks with lower ratings given by SMEs, and the percent of tasks with higher ratings given by SMEs for the 40% and 80% thresholds. We used the group's mode committee rating at the task level as the consensus rating. We also examined whether the observed committee percent agreement with the data was higher than if one simply rated all the tasks as being greater than the given threshold. We additionally investigated whether there were differences in agreement for procedure and non-procedure job tasks. We hypothesized that there would be differences between committee-based and task inventory methods because prior research in other contexts suggests differences in method comparison studies. We also hypothesized that there may be some differences for procedure versus non-procedure job tasks because procedure tasks often exhibit more variation across workplace settings and experience levels. We expected that committee members based their ratings on their own settings and experiences, which may not reflect the level of responsibility across all the people that were surveyed.

## Results

### Committee Interrater Agreement and Between-Committee Agreement

Table 2 contains interrater agreement statistics for the five committees. Proportions of agreement ranged from 59% to 80% with a mean of 72%. While this is a relatively high level of agreement, it is also important to keep in mind the chance level of agreement between raters (i.e., the level of agreement if each person's set of responses was rearranged in random order). The mean level of chance agreement was 57%. This makes the aggregate kappa agreement statistic to be .35. In other words, while there was a relatively high level of agreement among committee members, this agreement on average was only 35% better than chance agreement.

Table 2 also contains the between committee agreement for the modal (consensus) responses. These statistics only applied for disciplines 1 and 3, as discipline 2 only had one committee. Agreement for these consensus numbers was substantially higher than the individual interrater agreement statistics. Agreement ranged from 77% to 87% with a mean of 83% with slightly higher agreement levels observed for the 40% threshold. The mean kappa across the four conditions was .58, indicating that the between group agreement was more than 50% better than chance. It appears that taking a group consensus rating greatly increased the agreement levels.

## Agreement Between Committee Consensus and Surveys: All Tasks

Figure 1 displays committee agreement bar plots for all tasks for the 40% and 80% thresholds. There are some clear trends observed in the plots. First, the agreement rates for the committees varied depending on whether a 40% or 80% threshold was used. Typically, agreement rates were higher for the 40% threshold than for the 80% threshold. Second, one can see that that only one out of five committees at the 80% threshold and three out of five committees at the 40% threshold had percent agreement rates that were above 80%. These results suggest that many committees on average made more than one different prediction for every five predictions that they were asked to make. These results also suggest that there can be practically significant differences in the tasks defining exam and clinical requirements if one based the task list on the task inventory survey results or the SME ratings, as the tasks lists would typically differ by 15% or more. In addition, one can see at the 40% threshold level that the agreement rates for the SMEs did not exceed what would have been obtained if one had simply classified all tasks as being greater than 40%. Most tasks were above the 40% threshold, and the committee predictions often did not do better than simply saying that all tasks should be included. Another interesting trend depicted in Figure 1 is that, when the SMEs gave different ratings than the task inventory surveys, SMEs overwhelmingly thought that fewer people were responsible for the task. These results suggest that basing the task list on the SME ratings alone would typically result in a narrower list of tasks than basing the task list on the task inventory survey results.

## Agreement Between Committee Consensus and Surveys: Procedures vs. Non-Procedures

Figure 2 provides bar plots that separate the tasks by procedures versus non-procedures. Like Figure 1, one can see several interesting trends. First, except for discipline number 3 at the 80% threshold, SMEs tended to have lower agreement statistics for procedures than

**Table 2.** Proportion agreement statistics within and between committees

| | 80% Threshold Within Committee | | | | |
|---|---|---|---|---|---|
| Statistic | Discipline 1, Committee 1 | Discipline 1, Committee 2 | Discipline 2 | Discipline 3, Committee 1 | Discipline 3, Committee 2 |
| Interrater Agreement | 71% | 75% | 80% | 69% | 59% |
| Chance Agreement | 49% | 52% | 68% | 52% | 49% |

| | 40% Threshold Within Committee | | | | |
|---|---|---|---|---|---|
| Statistic | Disp. 1, Committee 1 | Discipline 1, Committee 2 | Discipline 2 | Discipline 3, Committee 1 | Discipline 3, Committee 2 |
| Interrater Agreement | 79% | 74% | 67% | 77% | 69% |
| Chance Agreement | 60% | 59% | 54% | 67% | 59% |

Between Committee Modal Response Proportion Agreement

| Statistic | Discipline 1 | Discipline 2 | Discipline 3 |
|---|---|---|---|
| Agreement | 83% | NA | 77% |
| Chance Agreement | 51% | NA | 52% |

| Statistic | Discipline 1 | Discipline 2 | Discipline 3 |
|---|---|---|---|
| Agreement | 87% | NA | 86% |
| Chance Agreement | 63% | NA | 71% |

they did for non-procedures. These results make some sense, because there are many non-procedure tasks that are done by almost everyone. If workers did not perform these tasks, then facilities could lose accreditation or patients would be severely harmed. Second, one can again see several situations where the SME agreement rates were lower than if all tasks from that section had been included in the task list. The most notable differences again occurred for the 40% threshold, where four out of five sets of committee ratings for procedures and two out of five sets of committee ratings for non-procedures were not better than simply saying everything was greater than 40%. Results were better for the 80% threshold, where only two out of five committees were not better than saying everything should be on task list for non-procedures and all committees did better than saying everything should be on the task list for procedures. Similar to the results for all tasks, we again observed for both the procedures and non-procedures that committees overwhelmingly thought fewer people were responsible for the task than was observed with the task survey results. This finding suggests that the biggest differences in task lists would typically occur in the procedure sections, and that generally the task list created by the SMEs would contain many fewer procedures than the task list generated from the task inventory surveys. In many cases, the procedures section had less than 80% overall agreement, and the task list would contain at least 20% fewer procedures. For example, for discipline number 2, the procedure list would include a little over 40% fewer procedures at the 80% threshold and a little over 10% fewer non-procedures. These differences in tasks are practically significant and suggest that exam and other credentialing program requirements may look much different if results were based simply on SME ratings versus task inventory survey results.

## Discussion

This research examined the veracity of SME judgements in the context of job analyses. Specifically, we wanted to know whether SMEs' perceptions of the percentages reporting responsibility for different tasks were consistent with the percentages responsible found when using task inventory surveys. This is an important issue to examine, as some job analyses use SME judgement as the only guide for what tasks are and are not a part of a certification program. In addition, some credentialing programs tend to give more weight to SME judgments and allow SMEs a large amount of leeway to override task inventory survey results when deciding what tasks should be included in the task list.

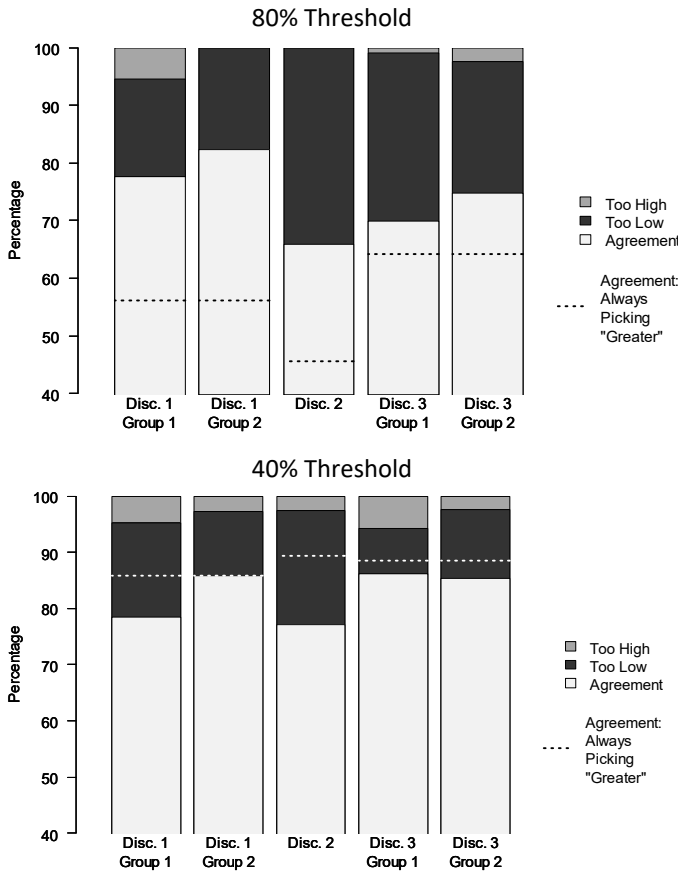


**Figure 1.** Committee Agreement Bar Plots Between the Consensus Committee Rating and the Job Analysis Survey for All Tasks

The results showed several discouraging trends and some key differences between SMEs and task inventory survey results. First, levels of committee agreement with the job analysis survey data were not as high as one would like to see. When determining whether a large percentage (80% responsible or more) of workers in the field conducted a task, the committees generally differed from the task inventory survey results in their task classifications over 15% of the time. The classification consistency was somewhat better for a lower threshold (40% responsible or more), but the percentages of agreement with the job analysis survey data were still lower than simply saying that all tasks were above the 40% threshold. It is disappointing that the SME committee agreement levels with task inventory survey results did not beat the clearly suboptimal strategy of "everything is in" for the 40% threshold. If one defined quality committee perceptions as both having greater

than 80% agreement and doing better than saying all tasks are in, only one out of the ten trials had quality perceptions (Discipline 1, Group 2 for the 80% threshold) based on this definition.

Second, committee predictions generally showed lower agreement for the procedure-based tasks compared to the more general practice tasks. This was a disappointing but not surprising trend, as the general task category contained numerous "no-brainer" tasks that obviously nearly everyone conducted in the field. The procedure-type tasks are often interpreted as the "meat and potatoes" of the task lists for the disciplines involved. If committee judgement was worse for this very important category of procedures, then having good data to help guide committee judgements becomes all the more important.

Finally, and possibly most importantly, there were systematic differences in SME judgments compared to task inventory surveys. Committee members overwhelmingly rated tasks as having lower percentages responsible than was observed when surveying a random sample of people working in the field. Thus, based on these findings, a completely committee-driven task list may be unjustifiably short compared to a task list created with a data-guided process.

Taking these results together, it appears that SMEs may not be able to make high-quality judgements about the percentage of people responsible for conducting tasks, at least in terms of being consistent with the percentages obtained when surveying large random samples of people working in the field. This seriously calls into question committee-only job analysis methods that do not rely on any other data source. Job analyses should not use committee judgement as the sole basis for including or excluding tasks from a final task list. Expert panels require some sort of external, representative data to help guide their judgements; it appears to be quite difficult to guess at what everyone

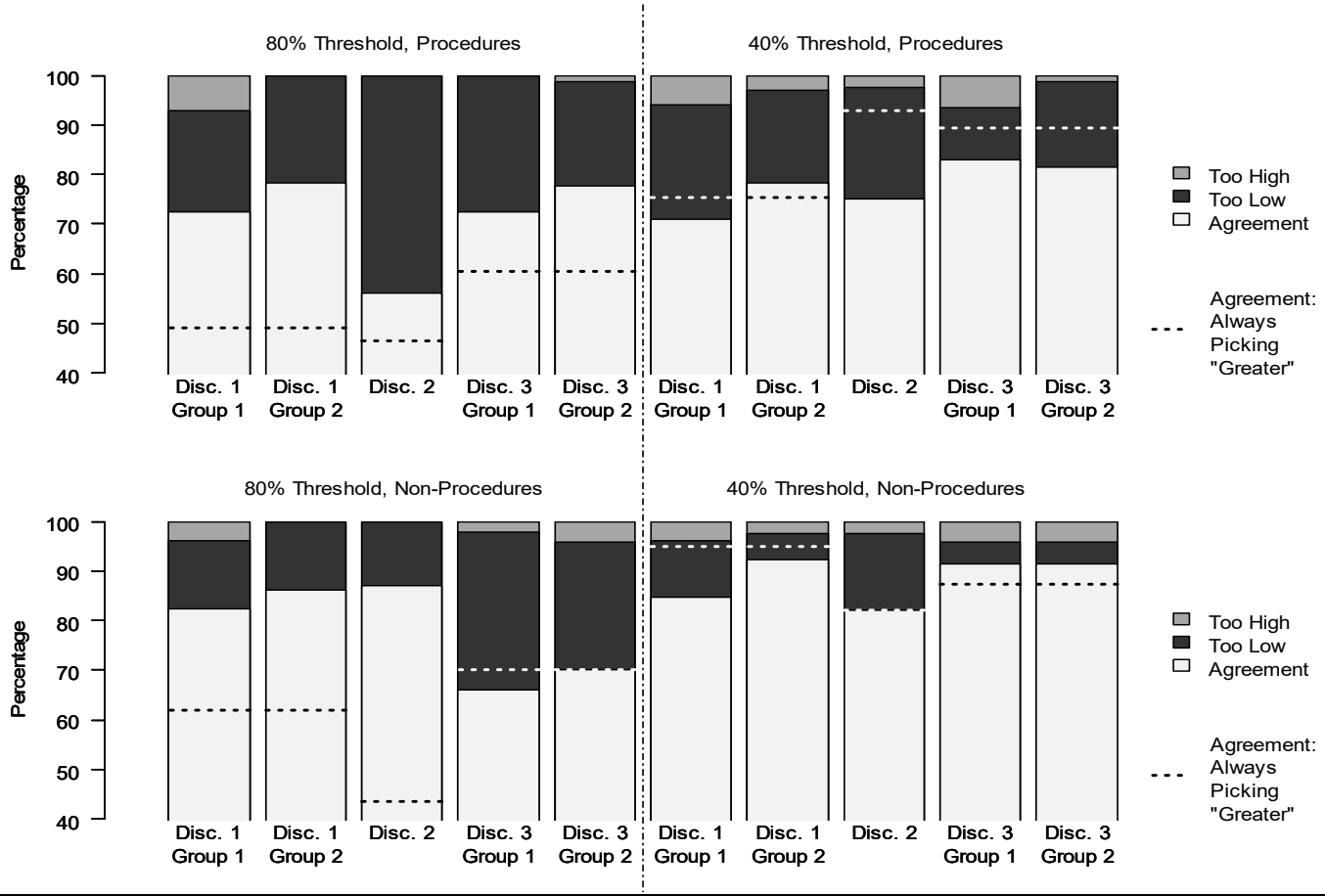


**Figure 2.** Committee Agreement Bar Plots for the Consensus Committee Rating and the Job Analysis Survey, Separated by Threshold and Procedures vs. Non-Procedures

else in a profession does on a day-to-day basis based only on one's own limited experiences.

It is also important to point out that we did find relatively high levels of interrater agreement within and between committees. These results suggest SMEs seemed to have a somewhat shared understanding of the tasks that they thought people in the profession were responsible for and that the differences between survey results and committee perceptions seemed to be held by SMEs within and across committees. It did not appear that the differences we observed between the task inventory results and committees was a function of there being large amounts of disagreement between SMEs.

One limitation of this study is that the task inventory survey results are certainly not perfect, insomuch as the results are subject to measurement error, sampling error, and individual workplace judgement error contained in any job analysis survey (see Morgeson & Campion, 1997). One cannot assume that the survey results are the perfect absolute truth. They are instead an estimation of the frequency with which people conduct each task in the workplace using a well-established and researched methodology. We do not believe that there was a high amount of inaccuracy in the task inventory survey results because we did not find significant differences between the returned and random samples on many of the key demographic variables. In addition, we correlated the task inventory survey results where applicable with Centers for Medicare and Medicaid Services data that reports the frequency with which different medical imaging procedures are billed to Medicare and Medicaid. We found correlations in the low 80s, which suggest that the task inventory survey data exhibited a high level of correspondence with these data. These findings give us a good deal of confidence that the people filling out the surveys could take the next step and give an accurate representation of the actual task frequency in practice.

The committees, however, are substantially more prone to two of the above types of error; measurement error and sampling error. The measurement error will generally be higher for the committees simply for the fact that there are less than ten committee members. The sampling error will also tend to be greater for the committees, but for two reasons instead of just one. The first reason is the same as the measurement error reason; drawing responses from fewer people will lead to greater opportunities for sampling error. The second reason has to do with the makeup of the committees. For the organization that conducted this study, the SMEs are in one or more of the following categories more often than the general population: working in a cutting-edge medical center, influential in local or national societies in the profession, and extremely knowledgeable in the profession as measured by achieving a high score on their initial certification exam. Due to the nature of the appointment process, it is likely that the committee composition will be less representative of the population of workers than the people returning the job analysis surveys. It should be noted that the recruitment and selection of SMEs by this organization in many ways parallels strategies used by other organizations, as many organizations justifiably try to include people with advanced knowledge and who are recognized industry experts on committees.

There is even an additional source of error for the SMEs that the survey respondents do not have: external workplace judgement error. The committee members will, in theory, have the same amount of error from individual workplace judgements as survey respondents. However, committee members must also think about what tasks people are doing in other workplaces (not their own workplaces). This added error for SMEs is potentially the largest source of error of all. People filling out the surveys must only think about their individual job role, not the job roles of others. So, in addition to having a higher probability for increased measurement error and sampling error, SMEs have an additional source of error not present at all in task inventory survey results. One would be hard-pressed to argue from a measurement perspective that the survey results from a large and fairly representative sample using a refined and well-researched methodology are a worse approximation of the true results in the population of interest than those of a small and quite unrepresentative, though well-intentioned, committee.

While we believe our results showing key differences between SME judgments and task inventory survey results are compelling, future research should conduct similar exercises to see if the results generalize to other contexts. It is also possible that different job contexts have key task category distinctions other than procedures versus non-procedures that could affect the quality of committee predictions. Different contexts could also have varying levels of agreement (or lack agreement) due to the mean percentage of people responsible for tasks in the survey. A majority of the tasks in this study's job analyses had high responsibility

levels as reported from the survey respondents. A job analysis where most tasks had low percentages of reported responsibility could influence committee agreement, though such a scenario would call into question either the sampling of survey recipients or whether there was truly a unified core job role in the first place. Another area for future research would be to explore how results may differ when other rating scales are used in the task inventory survey. For example, would SME ratings more closely match survey data if criticality or importance scales were used? It is possible that results may differ depending on the scale(s) used to perform the job analysis. In fact, some prior research suggests that results may be more similar when other scales are used (see Ash et al., 1982; Mauer & Tross, 2000; O'Leary et al., 1990; Tannenbaum & Wesley, 1993). Future research could also look at how results are impacted by using different thresholds from the ones used in this study, as other organizations may use other threshold guidelines in their job analyses.

This paper makes an important contribution by identifying key differences that may exist when employing two different job analysis methods. Identifying potential differences between job analysis methods is critical because job analysis provides crucial validity evidence in support of credentialing programs. The findings from this research also contribute to the growing body of literature indicating that SMEs, even with proper training, often struggle with activities asking them to quantify constructs that are important to healthy exams (e.g., providing standard-setting ratings, estimating item difficulty). Continuing research in this area will help sketch the contexts where SME judgement is most and least valuable.

## References

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

Ash, R. A., Levine, E. L., Higbee, R. H., & Sistrunk, F. (1982, March). *Comparison of task ratings from subject matter experts versus job incumbents.* Paper presented at the meeting of the Southeastern Psychological Association, New Orleans, LA.

Babcock, B., & Yoes, M. (2013). Enhancing job analysis in medical specialties with CMS data. *Evaluation & the Health Professions*, *36*, 240-246.

Flanagan, J. C. (1954). The Critical Incident Technique. *Psychological Bulletin*, *51*,327-358.

Gael, S. (1983). *Job analysis: A guide to assessing work activities.* San Francisco, CA: Jossey-Boss. Institute for Credentialing Excellence. (2016). *NCCA accreditation.* Retrieved from http://www.credentialingexcellence.org/ncca .

LaDuca, A. (1980). The structure of competence in the health professions. *Evaluation and the Health Professions*, *3*, 253-288.

Knapp, J., Anderson, L., & Wild, C. (2015). *Certification: The ICE handbook*, 2nd Edition. Institute for Credentialing Excellence: Washington, DC.

Mauer, T.J. & Tross, S. A. (2000). SME committee vs. field job analysis ratings: Convergence, cautions, and a call. *Journal of Business and Psychology*, *14*, 489-499.

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

Raymond, M. R. (2016). Job analysis, practice analysis and the content of credentialing examinations. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd edition, pp. 144-164). New York, NY: Routledge.

Raymond, M. R. (2005). An NCME instructional module on developing and administering practice analysis questionnaires. *Educational Measurement: Issues and Practice*, *24(2)*, 29-42.

Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education*, 14, 369-415.

Raymond, M. R. (1996). Establishing weights for test plans for licensure and certification examinations. *Applied Measurement in Education*, *9,* 237 - 256.

Tannenbaum, R. J., & Wesley, S. (1993). Agreement between committee-based and field-based job analyses: A student in context of licensure testing. *Journal of Applied Psychology*, *78*, 975-980.

Wang, N., Schnipke, D., & Witt, E. A. (2005). Use of knowledge, skill and ability statements in developing licensure and certification examinations. *Educational Measurement: Issues and Practice*, *24(1)*, 15-22.

Wang, N., & Stahl, J. (2012). Obtaining content weights for test specifications from job analysis

task surveys: An application of the Many-Facets Rasch model. *International Journal of Testing*, *12*, 299-320.

Wyse, A. E. (in press). Analyzing job analysis data with mixture Rasch models. *International Journal of Testing.*

Wyse, A. E., Eckerly, C., Babcock, B., & Anderson, D. (2016). A study of potential methods to increase response rates in task inventory surveys. *CLEAR Exam Review*, *16*, 15-22.

## Citation:

Wyse, Adam E, & Babcock, Ben (2018). A Comparison of Subject Matter Experts' Perceptions and Job Analysis Surveys. *Practical Assessment, Research & Evaluation*, 23(10). Available online: http://pareonline.net/getvn.asp?v=23&n=10

## Authors' Notes:

The authors would like to thank Lauren Wood for comments she provided on an earlier version of this manuscript. This article was written when the first author worked at the American Registry of Radiologic Technologists. Any opinions, findings, conclusions, or recommendations expressed in this manuscript are those of the authors and are not necessarily the official position of The American Registry of Radiologic Technologists or Renaissance.

A version of this paper was presented at the 2018 annual meeting of the National Council on Measurement in Education.

## Corresponding Author:

Adam E. Wyse
Renaissance
1813 Chatham Ave
Arden Hills, MN 5511

email: adam.wyse [at] renaissance.com