Methodological Issues Related to the Occupational Requirements Survey

Michael J. Handel
February 26, 2015

The views expressed in this document are those of the author and not necessarily those of BLS or SSA.

EXECUTIVE SUMMARY

The Social Security Administration (SSA) has contracted with the Bureau of Labor Statistics (BLS) to produce occupational data for use as the main source of information about job demands in determining eligibility for Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI) disability benefits. SSA uses five steps of sequential evaluation to determine whether claimants are eligible for benefits under these programs. The first three steps are decided primarily based on medical factors, but at steps four and five, eligibility depends on both a claimant's limitations based on his or her impairments, and the physical and mental demands of work that jobs require. Claimants found not disabled have the right to appeal in administrative law hearings and the Federal court system. The proposed Occupational Requirements Survey (ORS), currently in development, is intended to be an important component in a new Occupational Information System (OIS) that will be the primary source of information on job demands used in SSA's disability programs and considered in judicial appeals. The DOT's suitability for this purpose, given its growing age, has been a source of increasing concern, and SSA has explored alternatives since the late 1990s. ORS is an interagency program in which BLS is supplying technical expertise and institutional capacity to assist SSA by collecting important information about occupational requirements for inclusion in its OIS.

SSA's disability programs are large public benefits programs and an important part of the U.S. social safety net. The programs have been the subject of significant attention from political actors inside and outside the Federal government and from claimants' legal advocates. The long search for a DOT replacement has also brought different technical expert groups into the discussion. Given SSA's disability programs' importance to beneficiaries and the general public, the political attention it has received, and the professional interest the proposed data collection program has attracted, it is clear that BLS is entering somewhat new territory. The ORS' primary purpose is to assist directly in the administration of a large public program, rather than to provide standard statistical information that policy-makers use for general guidance as needed. The large responsibility and scrutiny associated with this data program argue for careful efforts to ensure the data's quality.

After completing several exploratory research phases, BLS requested a methodological report that would provide advice on achieving high levels of reliability, validity, and accuracy given the ORS' parameters. Specific concerns included the relative merits of different sources of information and modes of data collection. Review of program documents for this report also raised questions regarding the high levels of precision desired for variables dealing with frequency and duration of certain task requirements. The report was also asked to address concerns raised by the public in a recent official comment period when earlier program documents were released for public reactions. The comments relating to BLS' role on ORS clustered around concerns that in-person observations of job performance by trained field

analysts, which were a prominent part of the DOT, would play a much smaller role compared to interviews, and that the occupational classification system would be more aggregated than the DOT. In addition to addressing these concerns, this report also provides substantial background on general issues of reliability and validity as they relate to the ORS project.

The central recommendation is that the ORS conduct its own systematic tests and validation studies in order to ensure that the final survey design aims for the most reliable, valid, and accurate measures that can be expected given the various possible alternatives. While there are many existing studies of the reliability and validity of job analysis methods and a selection of the most relevant are reviewed, a systematic review of the literature would not be very productive and would leave many issues unresolved given the differences between the research design of most academic studies and the nature of the ORS program. The ORS is much larger than most other studies, which imposes certain constraints, and the measures are quite distinctive, reflecting SSA's specific needs. Given the significance of the proposed data, a dedicated effort tailored to the specific circumstances of the ORS program is recommended.

To accomplish this task the ORS needs to develop a coherent master plan for conducting methodological research. A strategy document should give a clear, compact summary and explanation of the ORS' specific data needs and the alternative data sources and collection procedures under consideration, including procedures, such as job observations, that are only feasible on a limited basis for the final phase of data collection but that can serve as a benchmark or gold standard for assessing validity more widely in smaller-scale methodological studies. The master plan should also identify occupations, measures, and specific occupational requirements with particular importance due to the frequency with which they present themselves in Social Security disability claims so that program resources, such as job observations, finer occupational detail, and enhanced measurement, can be targeted to particularly critical areas. This planning document is necessary both as a roadmap for the work that a validation phase must accomplish and because detailed study of reliability, validity, and accuracy requires prior specification of intended procedures and purposes, and required levels of precision. Where gold standards are difficult to find or utilize, alternative methods of data collection can be compared to assess levels of agreement, their ability to discriminate among occupations that are known to differ based on prior knowledge, and their levels of inter-rater reliability.

Systematic tests should be designed by an experienced industrial psychologist that seeks to identify the most significant sources of variance among raters using the same methods across jobs. Sources of variance to be considered can include

a. difficulty of the item (e.g., duration)
b. characteristics of the respondent
c. length of interview
d. characteristics of the job and establishment rated (e.g., skill level, task homogeneity, industry, urban/rural location, organizational size)
e. field economist (e.g., experience, training, judgment, personal style)

f. field office (e.g., variation in training, informal practices across offices)

As no set of tests can investigate all of these sources of variance, prior project reports should be mined and field economists and supervisors debriefed to identify those that are most likely to be the largest contributors. Clearly, rater-related variance is the traditional focus of reliability studies and should be a principal concern of ORS reliability studies. In addition, given the known difficulty in capturing the duration for certain data elements and the social distance separating many firm officials in large organizations from physically demanding front-line jobs, special attention should be given to these items and situations.

In short, the ORS needs to design and conduct a well-planned series of tests and validation exercises to evaluate the efficacy of different approaches to data collection and optimize the final design of the program. This involves

1. identifying all gold standards that might serve as validation criteria for accuracy
2. defining a reasonable range of other methods to assess convergent validity when gold standards are unavailable
3. identifying significant sources of measurement error for assessing reliability, including duration items and respondents relatively distant from the front-line job
4. considering methods for distinguishing error variance and true heterogeneity within critical occupations, and measuring their absolute and relative sizes
5. relating standard measures of validity and reliability to rates of classification disagreement across key boundaries (e.g., sedentary vs. non-sedentary) to assess the practical implications of ratings disagreements that are observed

Finally, the ORS should consider mining existing large-sample databases (e.g., O*NET, National Compensation Survey, DOT) for the insights they can provide regarding the properties of measures and procedures similar to the ORS, and the likely magnitude and practical significance of within-occupation variation in job ratings where the information is available to further address concerns about the occupational classification's level of detail.

INTRODUCTION

Beginning in the late 1960s the Social Security Administration (SSA) used the Department of Labor's Dictionary of Occupational Titles (DOT) as its primary source for occupational information in the disability determination process (Miller et al. 1980). Although the DOT is well-known for its level of informational detail on an extraordinary 12,000 occupations, most information in the DOT is now forty or even fifty years old, and even the most recent, partial update completed in 1991 is approaching twenty-five years old. For various reasons, SSA finds that the DOT's replacement, the Occupational Information Network (O*NET), fails to meet its needs and has asked the Bureau of Labor Statistics (BLS) to help design and implement a custom data collection program, the Occupational Requirements Survey (ORS), focused specifically on providing information that helps SSA accomplish its mission.

The ORS program draws on BLS' accumulated expertise and infrastructure, particularly with respect to its longstanding National Compensation Survey (NCS) program. This alleviates the need for SSA to develop its own internal capacity for large-scale, occupational data collection and avoids duplication by finding complementarities between Federal agencies. The ultimate goal is to produce a database covering the entire range of jobs in the U.S. economy at a relatively fine level of occupational detail in keeping with SSA's need to make accurate disability determinations in light of the availability of jobs at different levels of skill and other demands. BLS's role is primarily is to work as an internal contractor for SSA while keeping within the scope of its own mission. The substantive domains of interest are physical requirements, specific vocational preparation (SVP), cognitive job requirements, and workplace environmental conditions that may affect an SSA claimant's capacity to perform a job (e.g., dust levels that are excessive for asthmatics).

After completing several exploratory research phases, BLS seeks a review of methodological issues and previous research that can help ensure collection of high-quality data and meet the other requirements of various stakeholders. Key questions include:

- Do ORS measures of physical demands, SVP, environmental conditions, and cognitive demands meet accepted standards of reliability, validity, and accuracy?

- What are the relative merits of different source(s) of information? (e.g., managers, HR and other company officials, supervisors, job incumbents, trained project staff, other experts, job descriptions, other written materials)?

- What is the best mode of collecting information from these sources? (e.g., structured interviews, conversational interviews, group interviews, telephone vs. in-person interviews, standardized self-complete surveys, job observation by trained field staff)?

Although not mentioned explicitly, issues relating to the proper level of occupational aggregation are implicated in these questions and are a frequent concern among stakeholders. The concern is that the usefulness and validity of the data will be compromised by foregoing the DOT's detailed occupation codes and relying on the more aggregate Standard Occupational Classification (SOC) system. This is a very complex issue and because it relates to both questions of

reliability/validity and user acceptance of the final database, it is worth considering in some detail in this report.

Some indication of the nature of the issues can be conveyed by noting that individuals hold *jobs*, some of which have formal titles that have broad currency, others are not recognized across employers or apply to a very small share of the workforce, and others lack formal titles entirely. It is convenient for official statistics to describe job-holding patterns for the workforce as a whole in terms of *occupations*, which may correspond to some job titles that have broad popular currency but are generally higher-level rubrics that encompass many particular jobs and job titles. SSA also uses occupational information to determine the incidence of various *job characteristics*. The histories of occupational classification systems within the United States and internationally, and the significant revision of these systems that commonly occurs every 10-20 years, are testimony to the fact that most occupations are not some kind of unambiguous, naturally-occurring entity that merely needs accurate and precise recording, like a respondent's age, which can be resolved to the level of an exact day, month, and year. Current standards for classifying jobs into occupations are the result of great effort over several decades, and some aspects remain uncertain, debated, and imperfect, and are likely to remain open indefinitely, even apart from the issue of how classify new jobs (e.g., social media coordinator). The spirit of many outside comments suggests a common belief that "true" occupations have an objective reality and can be identified easily, and that individual jobs can be assigned to these occupational categories without error. None of these propositions has a strong foundation, as explained further in this report, but they drive common concerns regarding the level of occupational detail in the ORS and require serious consideration.

This raises the broader point that basic methodological questions do not exist in isolation from program goals and the broader institutional context. This report will not venture into distant lands that are properly the province of other agencies and the policy-making and political processes. However, there will be a handoff between BLS' outputs for ORS and those actors, processes, and institutions. Whatever simplicity is gained by considering methodological issues in the narrowest sense possible may be more than offset by the problems that arise after a data collection program is launched, when the drawbacks of treating an interface with stakeholders as if it were walls of a silo sealed off from them become apparent. It is always better to design a study as well as possible from the outset rather than forging ahead and trying to fix problems afterwards. The situation is complicated by the fact that the ORS must navigate a complex set of cross-pressures. These competing claims need to be recognized if the final design of the ORS is to balance them in reasonable fashion, and stakeholders are to understand why various concerns are or are not reflected in program outputs.

The most significant institutional consideration that must be recognized at the outset is the requirement that all government agencies work within the budgets authorized by Congress and Executive agencies, such as OMB. While it should be obvious enough not to need formal statement, examination of numerous documents from other stakeholders in the course of preparing this report makes clear the need to recognize explicitly a basic principle: *many things are possible when resources are infinite, fewer are possible when resources are finite*. Although not a particularly controversial concept, much of the debate among agencies and stakeholders regarding valid measures, measurement strategies, and levels of detail in occupational coding,

seems to take place absent awareness of this basic point. As will be argued below, there is good reason to believe that resource constraints are a major reason neither the DOT has been maintained nor has its replacement, O*NET, adopted its data collection procedures. Indeed, consideration of the issues in this report will show the question of feasibility arises repeatedly, as one can always improve a data product or most anything else by devoting additional resources to it *ad infinitum* provided the extra effort is at least moderately well-directed.

Taking a broader institutional perspective, the ORS, as an inter-agency cooperative effort meant to support a large public benefits program, is in the position of needing to satisfy diverse stakeholders and their interests to varying extents, including:

(1) SSA's needs for timely, relevant, usable, and credible information that is effective in helping it meet statutory mandates for its high-volume caseload while complying with standards established in administrative legal rulings and Federal courts, as well as satisfying Congressional oversight and demands
(2) OMB's need to control costs and ensure resources are used efficiently as a watchdog for taxpayer interests, which includes encouraging inter-agency cooperation to avoid duplication and upholding Federal agency standards,
(3) BLS's requirements that inter-agency cooperative efforts not stray too far from its core mission or compromise its standards and reputation as a source of high-quality, objective data
(4) ORS respondents' desires that the burden of survey collection in terms of time, effort, and intrusiveness remain reasonable, which is also an OMB requirement and a practical requirement for any data program's success
(5) The public's right to have a fair and accurate determination process for eligibility for disability benefits as defined by current law, keeping both false negatives and false positives within reasonable and feasible limits
(6) The need for SSA claims processors and staff, claimant advocates, diverse experts, and adjudicators involved in the disability determination process to view the occupational database replacing the DOT as trustworthy, relevant, interpretable, transparent, and impartial[1]

These concerns emerged from a review of documents from the Occupational Information Development Advisory Panel (OIDAP) phase, public comments to OMB, and the general literature review conducted for this report. It is important to enumerate them because the validity of any occupational information system is determined by how well many of them are addressed even as the differences among them make tradeoffs and shortfalls inevitable. To take just one example, stakeholder interests in points (5) and (6) suggest data elements should be numerous and detailed, but respondents' interests in point (4) suggest the opposite, even aside from budget constraints affecting the government actors. Likewise, most external stakeholders may be unaware, but should be advised, that OMB clearance of surveys such as ORS requires descriptions of respondent burden hours and cost with an eye towards keeping them under control, and requires use of standard classification systems, which likely does not include the DOT's unique occupational classification system (Somers and Salmon, slide 23).

[1] For convenience this report uses the term "adjudicator" to refer to both administrative law judges and Federal court judges. This may differ from SSA's use of the same term.

The most obvious conflict the ORS must negotiate is the understandable desire of many stakeholders for a maximalist database that is likely beyond the capacity of both participating agency budgets to support. This is properly a concern that non-government stakeholders should address to the political authorities in Congress and the Executive who make policies and budgets, rather than the operating agencies that must adhere to them. This report takes the current policy context as given regarding such issues as current statutory provisions, eligibility criteria and determination procedures, and agency budgets, under the assumption that the basic parameters within which the ORS is envisioned to operate will remain constant in the short run. The goal of the report is to shed light on questions of social science methodology and data quality within current limits of feasibility, rather than in terms of standards that assume few or no constraints. Yet even if sufficient resources were available there is no guarantee that those selected as ORS respondents would agree to the intensity of participation desired by ORS users, which would limit the depth of the data collected in any case. Any discussion of ORS data quality, indeed the entire ORS, exists within a context that includes a complex and somewhat tangled set of actors and interests. Nevertheless, it is legitimate for all stakeholders to expect that the ORS strives as much as possible to meet their needs, which is a key motivation for this report.

As the first entry in the numbered list suggested, the ORS is unlike most BLS data collection efforts in the sense that it will be a key input into a large-scale benefits program that has a long and close relationship with an extensive body of administrative case law and procedure. As one review of SSA's dissatisfaction with O*NET noted, "The process of disability determination can be quite litigious, and those in charge of making the determination prefer to minimize the risk of legal challenges…" (Hilton and Tippins 2010, p.167). The legal process adds an additional layer of complexity to the usual concerns over reliability, validity, and accuracy but raises issues that are sufficiently specialized to be beyond the scope of this report. The key point is that the burden the ORS is likely to bear in legal proceedings argues for particular care in program design.

The major common issues that emerge from a review of stakeholder concerns are uneasiness over the level of occupational detail in the ORS and primary reliance on data collection methods other than direct observation by trained field analysts. These issues are addressed in this report.

Some stakeholders have more specific and varying concerns with individual data elements (e.g., "standing and walking" should be measured separately, "bending" is different from "stooping," etc.). Although BLS may have had an advisory role with respect to these measures, this report takes the position that BLS plays a support role in helping SSA meet its goals. Evaluating the validity of items with respect to the disability construct is within SSA's jurisdiction and expertise as primary sponsor of the ORS, and outside BLS' core competence and mission. Therefore, concerns over construct validity at this level are directed more properly to SSA, though it can be noted in passing that insofar as such concerns call for *more* detailed information on job characteristics than previously available in the DOT they move the goalposts from meeting to exceeding established standards. This report will address narrower issues of construct and criterion validity, such as the relationship between ORS items and possible gold standards, as this more accurately reflects the scope and nature of BLS' role in the ORS program.[2] BLS is

[2] Concretely, whether or not "standing and walking" may be measured together or must be measured separately is a decision for SSA, and the extent to which either can be measured well using which methods is a concern for BLS.

supplying technical assistance to support SSA's mission. To put it simply, defining and operationalizing relevant job requirements is SSA's task and measuring them as well as feasible is BLS' task.

The key conclusions of the report can be summarized briefly:

- A number of basic considerations regarding reliability, validity, and accuracy have not been explicitly recognized or appreciated in most discussions, including those regarding occupational classification systems, although they are largely common knowledge among researchers, and it is important to clarify them in order to avoid unnecessary confusion
- The DOT is commonly viewed as a gold standard against which any replacement should be benchmarked but many aspects of the DOT's data collection process and the reliability and validity of its measures have been misunderstood and the actual nature of the DOT data is more complex and ambiguous
- Neither the methodological literature on the DOT nor other key sources, such as O*NET and industrial/occupational (IO) psychology studies, provide information on reliability and validity in sufficient detail or relating to sufficiently similar items and collection procedures as those envisioned by the ORS to address the questions raised in the beginning of this report. At best, they provide a general orientation to the issues and illustrative examples of previous experiences regarding measurement reliability, validity, and accuracy. Specific characteristics limiting their utility include:
    - Validation studies of dissimilar measures (e.g., highly abstract constructs) cannot be assumed to generalize to the behaviorally specific measures the ORS intends to collect.
    - Large-scale, national data collection programs differ in critical respects from small-scale, job-specific and sector-specific studies, meaning certain procedures that might be feasible on a small scale are precluded on a large scale due to cost considerations
    - The standards and other considerations involved in occupational data collected for lower-stakes purposes, such as hiring, training, compensation, vocational guidance or job referral, do not necessarily apply to measures intended to be used in the administration of a large-scale Federal benefits program that has substantial connections with the administrative legal system and formal case law
- For these reasons, the key action point in this report is that the ORS needs to conduct its own experiments and validation studies among a feasible set of alternatives in order to
    - assure itself that the final items and procedures it adopts are the most reliable, valid, and accurate that can be reasonably expected, and
    - provide reasoned assurances to stakeholders that its conclusions are well-grounded in evidence.
    - It is advisable to have one or more IO psychologists and vocational rehabilitation experts participate actively in the design of these experiments and validation studies in order to ensure sound design and stakeholder acceptance
- Two key issues that need to be resolved before deciding what experiments and validation exercises to run
    - SSA needs to clarify its needs and priorities because these are the criteria against which the validation exercises must be judged

    - BLS needs to clarify the particular data collection options that are feasible so that it knows which exercises are experiments to find the best techniques to use and which are validation exercises to benchmark those techniques against gold standards that are not feasible for full-scale use
- Some basic points can be made with respect to the quality of the ORS instrument and mode of collection on the basis of well-known psychometric, psychological, and survey research principles
    - Reliability and validity are enhanced by targeting more detailed information collection at points on a trait continuum that are a focus of particular interest. For example, computer-adaptive education tests ensure that each test-taker receives additional questions proximate to the ability level demonstrated on previous items to achieve more reliable discrimination of their precise trait level. By contrast, a one-size-fits-all set of items containing many items across the full range of difficulty can be expected to include large numbers that one could have anticipated would almost all be answered either correctly or incorrectly (or positively or negatively), providing much redundant information and losing opportunities for more relevant items. Insofar as SSA prioritizes in-depth knowledge on a tractable subset of occupations it makes sense for ORS data collection efforts to prioritize them as well in terms of finer occupational coding detail, additional data elements, and focused validation exercises. Given scarce resources, if the criticality of information is spread unevenly across the occupational spectrum then the allocation of resources should be targeted to reflect that fact.
    - Items that are concrete, as are most ORS items, are better than those where the referent is abstract and indefinite, which requires personal interpretation and introduces another source of likely error variance. However, items that are extremely specific or require very precise physical measurement or time accounting risk pressing up against or exceeding the cognitive capacities of respondents regardless of their work roles (incumbents, supervisors, etc.). If certain facts are not salient for respondents, such as exact weights, frequencies, or time spent in certain mundane activities, there may be no reason to expect people to be able to supply this information regardless of the mode of elicitation. Of course, if measurement errors are random they will cancel out when averaged, so this may not be a problem but the ability of sources to supply the information at the level of precision desired is a concern that runs through this report.

After providing background information, this report will proceed generally in the order indicated above, discussing general issues of reliability and validity as they relate to the ORS, current ambiguities in SSA goals and BLS data collection options, and key findings from the research literature on the DOT and other relevant sources. Specific stakeholder concerns, including those expressed during the OMB comment period, will be addressed at relevant points. A final section, subject to revision for the final draft in light of BLS comments, will suggest specific directions for experiments and validation studies.

I. BACKGROUND

A. HISTORICAL BACKGROUND

The DOT provides a catalogue of occupations in the U.S. economy as well as descriptions of the work performed in each occupation. Data were collected by analysts working in field centers who collected the bulk of the data on which the DOT is based by visiting business establishments, observing workers in jobs, and recording and scaling the information observed. Data collection procedures included coverage of particular industries; these industries were sometimes very broadly specified (e.g., retail trade) and sometimes very narrowly specified (e.g., button manufacturing).

Establishments within each industry were chosen for analysis, with some effort being made to choose "typical" establishments, but selection depended on the establishment's willingness to participate. Once a site was selected, analysts prepared a report summarizing the structure of the workplace and the organization of jobs within it. For each analyzed job the analyst prepared a job analysis schedule, recording the tasks entailed in the job, the machines, tools, or work aids used, the working conditions, and a variety of other information. On this basis a description of the job was prepared, and the job was rated with respect to 46 characteristics (worker functions and worker traits).

In the past several years, problems with the DOT have been raised. Some of these issues include the fact that no standards were defined for describing occupations and were thus very uneven across jobs. Classification of different types of jobs was also difficult to determine. The DOT contained an overrepresentation of manufacturing fields and not enough in service and trade, and too few observations made for particular jobs, resulting in fewer jobs identified in such areas as clerical work. Approximately 75% of the jobs identified have incomplete data. Furthermore, excluding a partial update in 1991, the DOT has not been fully updated since 1977. As occupations and their duties and responsibilities have changed over the last 20 plus years, the relevance of DOT-based information in the SSA disability adjudication process has declined as well.

DOT sampling methods have also been criticized. No standard sampling methodology was implemented, but rather locations and professions to sample were selected haphazardly and determined by common sense (e.g., geographic locations that have a high rate of a certain profession). Establishments were selected to obtain a variety of sizes (small, medium, or large businesses) across different vocations. Finally, worker functions, traits, and job complexity measures were somewhat problematic in that they were vague and ambiguously defined. Thus, researchers recommended that a comprehensive survey be conducted to update these missing data and jobs from the DOT, one of the biggest problems being that many of the occupations included in the DOT are now outdated such as "envelope addressor."[3]

[3] Ferguson, G. R. (2013) *Testing the Collection of Occupational Requirements Data.* Joint Statistical Meeting – Government Statistics Sections.

In an effort to update the information contained in the DOT, the Occupational Information Network (O*NET) was established in 1998 by the DOL to summarize information and characteristics of different jobs (Hilton and Tippins 2010). Data were collected by contacting establishments to participate and then offering job incumbents multiple ways to respond (e.g., web and mail versions of the survey). Job incumbents answered questions about worker characteristics (e.g., abilities, occupational interests, work values, work styles), requirements (e.g., skills, knowledge, education), experience (training, entry level requirements, licensing), occupational requirements (e.g., generalized work activities, detailed work activities, organizational context, work context), workforce characteristics (e.g., labor market information, occupational outlook), and occupation-specific information (e.g., tasks, tools and technology). However, O*NET was not without its criticisms as well. For example, approximately 75% of the occupations were identified through probability-based sampling, and respondents for 25% of the occupations were identified by other less scientifically rigorous methodologies. To gather information for the skills and abilities domains, respondents for all occupations were identified by methodologies other than probability-based sampling. Others argued that O*NET used vague constructs in its questions, leading to measurement error and low reliability. Importantly, and most relevant for the current research, O*NET data do not meet the statutory requirements of the SSA Disability Determination process.

Because of this, SSA approached the Bureau of Labor Statistics (BLS), specifically the National Compensation Survey (NCS). NCS is a national survey of business establishments conducted by the Bureau of Labor Statistics (BLS).[4] Initial data from each sampled establishment are collected during a one-year sample initiation period. Many collected data elements are then updated each quarter while other data elements are updated annually for at least three years. The data from the NCS are used to produce the Employer Cost Index (ECI), Employer Costs for Employee Compensation (ECEC), and various estimates about employer provided benefits. Additionally, data from the NCS are combined with data from the OES to produce statistics that are used to help the President's Pay Agent and the Federal Salary Council recommend changes in GS pay under the Federal Employee's Pay Comparability Act. In order to produce these measures, the NCS collects information about the sampled business or governmental operation and about the occupations that are selected for detailed study. Each sample unit is classified using the North American Industry Classification System (NAICS). Each job selected for study is classified using the Standard Occupational Classification system (SOC). In addition, each job is classified by work level – from entry level to expert, nonsupervisory employee to manager, etc. These distinctions are made by collecting information on the knowledge required to do the job, the job controls provided, the complexity of the tasks, the contacts made by the workers, and the physical environment where the work is performed. Many of these data elements are very similar to the types of data needed by SSA for the disability determination process. All NCS data collection is performed by professional economists or statisticians, generically called field economists. ORS data collection has been designed to meet the statutory requirements of the SSA disability adjudication process. SSA will use this occupational data on Specific Vocational Preparation, Physical Demands, Environmental Conditions, and Cognitive Demands to

[4] Ferguson, G. (2013) "Testing the Collection of Occupational Requirements Data." Proceedings of the 2013 Joint Statistical Meetings, Montreal, Canada.

determine if updated employment requirements data for occupations can and will help in the determination that SSA uses in administering the Social Security Disability Insurance (SSDI) program.

B. POLICY CONTEXT

SSA's Social Security Disability Insurance (SSDI) program is intended as a safety net for workers and their families who suffer from severe physical or mental conditions that prevent them from continuing to work at a level that exceeds a certain minimal threshold. The governing statute defines "disability" as the inability to engage in substantial gainful activity by reason of a medically determinable physical or mental impairment expected to last at least a year.

SSA determines eligibility for benefits in a five-step process. By the end of the third step the claimant who has met current earnings and medical hurdles has their residual functional capacity to perform work-related activities (RFC) classified according to the five exertional levels of work defined in the Dictionary of Occupational Titles (sedentary, light, medium, heavy, very heavy). The final two steps require occupational information, such as provided previously by the DOT, to compare their own functional capacities to those required by available jobs:

- Step 4. *Previous work test.* Can the applicant do the work he or she had done in the past? If the individual's residual functional capacity equals the previous work performed, the claim is denied on the basis that the individual can return to their former work. If the claimant's residual functional capacity is less than the demands of his or her previous work, the application moves to Step 5.

- Step 5. *Any work test.* Does the applicant's condition prevent him or her from performing "*any other kind of substantial gainful work which exists in the national economy?*," meaning work that "*exists in significant numbers*" either in the region of residence or in several regions of the country.[5] If yes, the application is accepted and benefits are awarded. If not, the application is denied. In this step the RFC is applied against a vocational grid that considers the individual's age, education and the transferability of previously learned and exercised skills to other jobs. The vocational grid directs an allowance or denial of benefits.

Over 50% of all initial disability claims require vocational review at Step 4 or both Steps 4 and 5 (Karman 2009). According to some legal analyses, the claimant bears the burden of proof in Steps 1-4, while SSA bears the burden in Step 5, for which the DOT or some similar authoritative source of job information is quite important (Hubley 2008).

Federally-funded, state-operated Disability Determination Services conduct an assessment of all the claimant's functions to determine residual functional capacity related to their relevant past work or any other work that exists in the national economy, taking into account the claimant's age, education, and past work. This is the root of SSA's interest in information on SVP, physical demands, environmental conditions, and cognitive demands. In 1966, SSA and DOL entered an interagency agreement to publish data in a form that met SSA needs in a companion volume to the DOT, Selected Characteristics of Occupations (SCO). Physical residual functional capacity

[5] Quotations are from the Social Security Act Section 223(d)(2).

is based on DOT categories for physical job demands. SSA's Medical-Vocational Guidelines (grid rules), published 1978, are also based on DOT definitions.

Claimants can appeal decisions and may appoint a representative to act on their behalf, who may be an attorney or lay advocate. Appeals include three levels beyond SSA. Hearings before administrative law judges (ALJ) involve oral and written arguments, scheduled witnesses, questioning of witnesses by the claimant's side, and questions from the ALJ. In addition to medical experts, witnesses can include vocational experts, who have long relied on the DOT as an authoritative reference volume. If the claimant is dissatisfied with the decision at this level, he or she can request a review before SSA's Appeals Council, and, finally, file suit in U.S. district court. A very small number of cases can reach the U.S. Court of Appeals and U.S. Supreme Court (Morton 2014, p.13). For example, a recent ruling by the Seventh Circuit of U.S. Court of Appeals written by Judge Richard Posner, *Browning v. Colvin* (9/7/14), criticized various aspects of the DOT's use in disability determination specifically (for excerpt, see http://www.ssaconnect.com/index.php/privateconnect).

In June 2014, about 9 million disabled workers and 2 million children and spouses of disabled workers received benefits, averaging $1,146 monthly for workers and slightly over $300 monthly for the others. It appears the vast majority of cases workers receive disability benefits until they reach full retirement age or die, while a small number leave due to medical improvement or earnings above the threshold for benefits eligibility (Morton 2014, p.4). One source indicated 2.6 million disability claims were filed in 2011, which gives some indication of the agency's substantial workload.[6]

SSDI is primarily funded through a portion of the Social Security payroll tax that is directed to the Federal Disability Insurance (DI) Trust Fund. In FY 2013, the DI trust fund disbursed more than $139 billion in benefits. Social Security trustees project that the fund will be exhausted in the fourth quarter of 2016 under provisions of current law, after which continuing tax revenues will be sufficient to pay around 81% of scheduled benefits (Morton 2014, p.17). A similar situation in 1994 prompted Congress to change the DI's funding formula, extending its solvency to what was then projected to be 2016, with which current projections coincide exactly. Nevertheless, some believe the growth in benefits, which is depleting the fund, is due to more than predictable demographic changes and propose various ways to limit future growth (Morton 2014, p.18).

There is significant academic and political contention regarding the long-term growth in the number of SSDI recipients and total program spending, with some saying the program is too lax (Autor and Duggan 2006; Wolf and Engel 2013; Lieberman 2013)[7] and others saying eligible claims are rejected in order to ration an inelastic budget and cope with high caseloads at the expense of those in legitimate need (Hubley 2008; cf. Lipsky 1980). Some research indicates the program's most recent growth is related to disproportionate job loss experienced by those with

[6] Occupational Information Development Advisory Panel (OIDAP), *Annual and Final Report to the Agency*, 2011-12 (2012), p.10.

[7] See also *Social Security Administration Oversight: Examining the Integrity of the Disability Determination Appeals Process*, hearing before the Committee on Oversight and Government Reform, U.S. House of Representatives, June 10, 2014.

disabilities during the recession (Kaye 2010) and that SSDI is not used as a substitute for UI benefits once the latter are exhausted (Hyde 2014). "Despite a widespread belief that Government disability benefits provide an adequate safety net for working-age adults with disabilities—perhaps even an enticement to leave the labor force when times are tough—the very high rate of unemployment following job loss indicates that a large proportion of those losing jobs either need to remain in the labor force or choose to do so" (Kaye 2010, p.29). GAO was asked to investigate the issue of program abuse and recently issued its report.[8] Further discussion of these issues and the policy context can be found in Vallas and Fremstad (2014). This report will not address the relative merits of these positions except to say that they demonstrate clearly that in addition to serving a very large number of beneficiaries and being the focus of an even larger group of applicants, SSDI is also part of broader debates over the size and structure of government benefit programs that has played a prominent role in American policy and political life for several decades.

In addition to policy makers, analysts, and commentators, there are several organized professional groups of experts and advocates that are active or involved in this area. Industrial/organizational (IO) psychologists are experts regarding measurement of job requirements, vocational and rehabilitation experts frequently testify in administrative law hearings, and lawyers and other advocates specializing in disability cases represent claimants, many belonging to the National Organization of Social Security Claimants' Representatives (http://nosscr.org/). Given the centrality of SSA's occupational information to their own concerns, all of these actors have been involved in or observed SSA's search for a replacement for the DOT since the late 1990s, expressed concern over its course, and will subject the ORS to considerable scrutiny after its release. As noted, SSDI decisions are subject to legal challenges from individual claimants believing they have been denied unfairly and there seems little doubt that the ORS will be subject to legal challenge. Indeed, it is the job of disability lawyers to try to discredit any unfavorable evidence or argument as part of their obligation to represent their clients. As part of this process, IO psychologists and vocational and rehabilitation experts will be asked their thoughts about the reliability/validity of the ORS. Their responses and other statements will be scrutinized for any indication that the ORS is tilted against claimants. [9] Consequently, stakeholders' views of the data's quality matter for its ultimate usability.

This situation is not unique. Other official statistics have been subject to debate or controversy among professionals, politicians, the media, and the broader public. Measures of GDP are not free from conceptual and measurement problems. Possible mismeasurement of productivity has been a source of great concern following the slowdown observed beginning in the early 1970s. Household and payroll surveys provide differing estimates of total employment, only partly amenable to reconciliation despite great efforts, which has fueled divergent views of short-term economic trends (Bowler and Morisi 2006). Critics claim official occupational projections underestimate rising workplace educational requirements (Bishop 1992, 1997; Carnevale, Smith, and Strohl 2013). Federal occupational injury statistics are known to have problems with undercounting (Ruser 2008; Nestoriak and Pierce 2009). The official poverty line has been the

[8] "Disability Insurance: Work Activity Indicates Certain Social Security Disability Insurance Payments Were Potentially Improper," GAO-13-635, 2013.

[9] For examples of such arguments with respect to the DOT, see http://jamespublishing.com/shop/social-security-disability-advocates-handbook/.

focus of both academic and political debate on various grounds (Johnson and Smeeding 2012). The Consumer Price Index has been subject to numerous criticisms that it overestimates inflation (Boskin Commission) or that it underestimates inflation (inflation hawks), sparking fears of different sorts of political manipulation (Greenlees and McClelland 2008).[10] Recent unemployment figures have been accused of partisanship.[11] The Decennial Census is no stranger to controversy (Alonso and Starr 1989; Anderson and Fienberg 2001; Prewitt 2003). The veracity of Federal and local crime statistics has been questioned. Nationally-reported rates of high school graduation, state and local reports of test scores, and educational testing programs themselves have come under fire regarding their accuracy and validity. Academic debates and popular controversies across all major government agencies are the principal focus of an entire book on the disputed quality of official statistics and other respected data, currently in its fourth edition (Maier and Imazeki 2013). These concerns vary widely in their cogency and it is unrealistic to expect that major data collection programs like the ORS will avoid controversy entirely given the inevitability of data imperfections and the divergent professional viewpoints, partisan loyalties, and material interests that can be found in the broader policy environment. Nevertheless, the preceding underscores the importance of keeping official statistics above this kind of doubt and questioning wherever possible (cf. National Research Council 2013). The ORS will succeed if it is an acknowledged source of credible and objective occupational information. This is only possible with thoughtful program planning and design, and credible empirical evidence regarding data quality.

In key respects the ORS represents a departure from previous BLS data collection programs. Although this report did not study formally the Employment Service's use of the DOT and O*NET in determining eligibility for UI benefits, informal impressions suggest a significant difference in the use of those databases and the intended use of the ORS. Although the DOT and O*NET data programs were not unmindful of SSA's needs, the latter do not appear to have been central to their design or purpose. The DOT and O*NET appear to have been designed mainly, though not exclusively, as an aid to relatively low stakes decision processes, such as vocational counseling, career guidance, job referral, and job placement; SSA was one constituency among many.[12] By contrast, the *primary* purpose of the ORS is to support a high-stakes benefits determination process that involves many well-established, often organized, actors and is subject to a well-developed tradition of legal review. ORS' central purpose from the outset is to assist

[10] For an overview, see, "Why won't inflation conspiracy theories just die already?" by John Aziz, The Week (August 4, 2014), http://theweek.com/article/index/265764/why-wont-inflation-conspiracy-theories-just-die-already, accessed 12/22/14.

[11] See "Jack Welch questions jobs numbers" by Chris Isidore, CNN Money (October 5, 2012) http://money.cnn.com/2012/10/05/news/economy/welch-unemployment-rate/, accessed 12/22/14, and "Unsubstantiated Allegations that the Philadelphia Regional Office Manipulated the Unemployment Survey Leading up to the 2012 Presidential Election to Cause a Decrease in the National Unemployment Rate," Office of Investigations, Office of Inspector General, U.S. Department of Commerce, Report Number 14-0073 (May 1, 2014).

[12] "The DOT is first of all an instrument to unify job descriptions and categorize jobs so that employment service offices can classify and code applicants' occupational qualifications and employers' job openings in a commonly understood language…The DOT was developed primarily for counselling, job placement, and training…" from "The Butcher, the Baker, and the Missilemaker," (no author), Monthly Labor Review (May 1966, p.481f.) "The fourth edition DOT reflects the continued primacy of job-worker matching as the reason for its existence" (Miller et al. 1980, p.196). O*NET is intended as a "resource for businesses, educators, job seekers, HR professionals, and publicly funded government programs" (Lewis, Rivkin, and Frugoli 2011, slides 11 and 87ff.).

SSA's disability determination process. Less extensive or documented methods of data validation that may have presented few issues for previous data collection programs should not be assumed sufficient in the present context. The ORS will be subject to greater scrutiny than the DOT or O*NET. The challenge will be to satisfy the widespread desire for an authoritative source of information and guidance while meeting the kinds of concerns raised by stakeholders regarding the technical adequacy and neutrality of various plans considered during SSA's lengthy search for a replacement to the DOT. The history of external and internal comments regarding SSA's efforts suggests a demanding standard that would be a high hurdle for any data collection program to clear.

For these reasons it is imperative that BLS conduct its own reliability and validity studies so that its efforts and evidence can provide some assurance to experts, claimants, advocates, examiners, and adjudicators that the ORS is a technical resource compiled with the greatest care practicable and aiming to provide an accurate and objective portrait of job requirements in the U.S. economy. Absent such validation studies it will be harder to maintain the necessary confidence of key stakeholders. However, it is also true that reliability and validity are technical questions that are not well understood outside professional circles and do not necessarily correspond to popular impressions. In the interests of clarity, these issues are considered in the next section.

II. RELIABILITY AND VALIDITY ISSUES

A. GENERAL ISSUES

Social science approaches to issues of measurement quality are influenced strongly by a long discussion and debate within psychology over the meaning of the concepts *reliability* and *validity*. Briefly, reliability refers to the reproducibility or repeatability of measurement values under similar conditions, while validity concerns the extent to which the measurement methods or quantities measured are appropriate or acceptable for the particular purpose(s) for which they will be used. Much of this discussion developed in the context of efforts to measure complex, abstract concepts important to psychology, such as general cognitive ability, achievement in academic subjects, job fitness, and mental states and personality traits (e.g., burnout, extroversion). Such constructs are not directly observable and can only be measured using various kinds of indirect indicators, such as educational and cognitive tests, and attitude scales. These tests and survey batteries are commonly called "instruments" because their purpose is measurement of some quality, albeit intangible. The GATB and vocational interest inventories used by the DOT and O*NET belong to these traditions. In this conception what is observed is always only a measurement of the target concept, while the true magnitude or value of the trait under study always remains unknown. This makes the most sense when the variables are theoretical or conceptual (constructs, latent variables), and can be measured only indirectly.

This view differs from common conceptions in which a knowable true value is taken for granted and measurement adequacy is indexed by how well a method approximates that value. This view is more sensible for physical quantities that can be measured directly and for which high-precision methods of measurement exist. Such "gold standard" methods of measurement usually involve physical instrument readings or chemical tests (e.g., assays), as in some, though not all,

medical and health research. ORS' measures of physical requirements and environmental conditions, which are highly concrete, may approximate this situation, though the cognitive and mental demand constructs and the disability construct itself are more similar to examples in the previous paragraph; SVP is likely intermediate. Finally, discussions of reliability and validity also draw on terminology and illustrations from prototypical rating activities, such as judging Olympic sports or clinical diagnoses, in addition to psychological tests, attitude scales, physical measures, and chemical tests, as will be evident at points in the discussion below.

As implied by the preceding, reliability assumes that observed values always contain measurement error in addition to any signal of the unobserved true score. By contrast, validity refers to the need to ensure that what is measured is appropriate given the intended uses of the scores. Reliability is usually considered a prerequisite for validity because excessive measurement variability casts doubts on the ability to estimate true values with great confidence. However, high reliability is insufficient for concluding a measure has high validity, as it is possible to measure an inappropriate concept or operationalization, or an off-target (miscalibrated) value of a suitable measure quite consistently. Indeed, a measure with lower reliability and higher validity may be considered superior to one with higher reliability and lower validity depending on the specific magnitudes. However, while validity is usually considered more critical, its precise meaning and measurement are subject to much greater conceptual debate and practical difficulties than reliability (e.g., absence of gold standards or other criteria). Consequently, reliability indices are often the default for assessing data quality even though this is recognized as sub-optimal.

A point worth stressing is that while it is tempting to consider reliability and validity as *general* qualities of a measure or method (i.e., *Is it reliable and valid?*), reliability and validity have meaning only in the context of *specific* procedures, methods, and goals. For example, individual ratings may have disappointing reliability on their own but very high reliability when averaged and used as occupation-level scores. Indeed, quality differences between individual and occupational-level scores based on the same underlying data can be dramatic. This is directly relevant for the ORS project for which the question of the unit of analysis (individual job vs. occupational average) remains murky. Nevertheless, whether a measure meets a standard for reliability often depends on how it is used, so this issue will have to be clarified at some point by SSA. Are only averages and other summary statistics needed or is the individual-level data in itself to be used in some fashion? SSA knows its own procedures better than BLS and will need to explain them before one can determine which reliability indices are appropriate in the present context.

Likewise, in recent years most psychologists are emphatic that validity is not intrinsic to a measure but is a characterization of the intended use to which it is put. A standard driving test may be invalid for certifying fitness to operate a semi-trailer truck but may easily meet validity standards for certifying competence to operate ordinary vehicles. In other words, an instrument is valid for a particular purpose but not necessarily for others, even those that might appear close in kind. Ambiguity regarding precise purpose complicates a validity study because validity describes the quality of the relationship between the measure and the purpose for which it will be used, not the measure considered by itself. Even with respect to the one type of driving, the exam may be valid to use on a pass-fail basis if capacity is considered dichotomous

(presence/absence), but invalid for making more precise distinctions. This leads insurance companies, which *do* require finer discrimination, to obtain more detailed information on driving safety level, including sometimes direct monitoring of driving behavior with on-board electronic measurement and recording devices. This information is much more precise than scores on a driving exam, which do not have adequate validity for *this* purpose, but the greater precision of measurement can be gained only at greater cost. The corresponding situation for ORS is whether or to what extent it uses objective measuring instruments, such as weight scales, thermometers, and decibel meters, to measure object weight and extreme temperature and noise in the workplace. Such instruments enable greater precision but the additional cost requires persuasive argument that the gain in criterion validity is both necessary for present purposes and worth the expenditure. In short, asking whether a measure is valid only makes sense if one specifies the purpose or procedure for which the measure will be used, including the required level of precision. *There is no such thing as reliability or validity in general or apart from particular problem contexts.*

The need to specify precisely the intended uses of an instrument also arises because measuring whether jobs cross a single threshold of exertional difficulty or not is generally easier than measuring multiple levels, simply because the former is a subset of the latter. If classifying cases on either side of a single threshold involves some difficulty and error, then classifying the same cases more finely with multiple thresholds will multiply the opportunities for difficulty and error, as well. In other words, how finely SSA needs to classify jobs will affect the measured levels of reliability and validity of ORS items and scales. For example, if SSA is interested only in whether a job is "sedentary" or not then only classification errors with respect to that threshold matter and all other potential classification errors (light vs. moderate, moderate vs. heavy, heavy vs. very heavy) are irrelevant and uncounted in calculating an index of reliability. Imprecise measurement across an irrelevant part of the trait continuum is ignorable and not error, just as there is no need to devote great efforts assigning detailed letter grades for a pass-fail test. Likewise, there is no problem if some respondents report a job involves frequent lifting of 25 lbs. and others with a different job title in the same SOC report lifting 35 lbs. with the same frequency if the ORS classifies exertional levels using the interval 21-50 lbs. Pointing to any difference as if it were consequential per se fails to recognize that *all within-category variation is ignorable*. This point should be understood clearly by stakeholders, as well as SSA. It means that if there is interest in distinguishing four different levels of non-sedentary jobs, for example, then classification errors proximate to all thresholds matter and the potential number of classification errors rises by a factor of four, all else equal. In either case an index of data quality will clearly underestimate reliability if it treats a variable as continuous and counts all disagreement among measurements as error when the actual intent is to treat the variable categorically and treat all within-category distinctions as irrelevant. Clearly, *a measure's reliability, validity, and potential error rate are only meaningful given the intended level of measurement exactitude (coarseness/fineness)*, which typically affects the cost of collecting the data, as well.

Finally, even if the purposes and procedures are well-specified there remains the issue of judging whether standards of acceptable reliability and validity have been met. Reliability and validity are measured on continuous scales, generally ranging from 0 to 1.0; they are not binary quantities (presence/absence). One can speak of a measure's estimated **level** of reliability or

validity but not whether the measure *is* or *is not* reliable/valid in an absolute sense except in the most extreme cases. Because all measurements contain error and measurement quality is a continuum there are few clear-cut standards for acceptable levels of reliability and validity. In general, there are no obvious or universal cutoffs or break points separating unacceptable and acceptable levels of reliability or validity. Any decision regarding acceptable levels of reliability and validity involves judgment, though benchmarking relative to other, generally similar cases provides some basis for those decisions. However, the question cannot be whether ORS measures exhibit less than perfect reliability and validity because this is true of all measures. Recognizing this, the U.S. Office of Personnel Management's Assessment Decision Guide (n.d.), designed to help government agencies develop defensible selection tests for hiring and promotion, noted

> *In practice, [test] scores always contain some amount of error and their reliabilities are less than 1.00. For most assessment applications, reliabilities above .70 are likely to be regarded as acceptable.…validity coefficients for a single assessment rarely exceed .50. A validity coefficient of .30 or higher is generally considered useful for most circumstances* (pp.6f.).[13]

These considerations apply to job measurement, as well as pre-employment testing. IO psychologists have long been aware of "the disconcerting fact that a high level of disagreement often exists between incumbents *within the same job title* [i.e., sub-occupation level] regarding the tasks they perform," due to some unknown combination of random error and true within-group variance (Harvey 1991, p.109, emph. orig.). (For ease of discussion, this report will generally refer to true within-group variance as **heterogeneity**.)

The preceding relates to the concern that the nature of individual jobs or narrow occupations may be mischaracterized by SOC-level occupation values. Leaving aside for the moment questions regarding *how much* measures at different levels of occupational aggregation might differ, discussed next, the notion that statistical measures correspond perfectly to individuals' true values or are perfectly predictive at the individual level misconstrues the nature of statistical conclusions. Even the best possible actuarial model will not *predict exactly* when a *specific individual* will die, and no one argues seriously that grades and college entrance exams are definitive measures of individual-level academic performance or potential even though they carry great weight in college admissions decisions. Measurement and prediction always perform better for groups than for specific cases. Cherry-picking individual instances in which an SOC-level measurement differs (by some unspecified amount) from the value at an individual job or finer-occupation level is not a recognized method for measuring overall validity because it is the *absence* of such anomalies that is unusual. As Lee Cronbach, a leading test and measurement psychologist and creator of the coefficient α bearing his name, observed, "perfect prediction is a false ideal" (Cronbach 1984, p.140).

> *Long ago, William James warned psychologists against trying to "write biographies in advance." Whether a validity coefficient is large enough to warrant prediction from the test depends on the benefit obtained by making predictions, the cost of testing, and the*

[13] Available at http://www.opm.gov/policy-data-oversight/assessment-and-selection/reference-materials/assessmentdecisionguide.pdf (accessed 12/7/14).

cost and validity of alternative selection methods. To the question, "What is a good validity coefficient?", the only sensible answer is, "The best you can get" (Cronbach 1984, p.140).

Tufts University biostatistician Gerard Dallal has a useful discussion of the issues surrounding how much estimates differ from true values in a discussion of sampling error (pollster's "margin of error") that applies equally well to the uncertainty resulting from imperfect reliability and validity:

> *In a survey, there's usually no hypothesis being tested. The sample size determines the precision with which population values can be estimated. The usual rules apply—to cut the uncertainty (for example, the length of a confidence interval) in half, quadruple the sample size, and so on. The sample size for a survey, then, is determined by asking the question, "How accurately do you need to know something?" Darned if I know!*
>
> *Sometimes imprecise estimates are good enough. Suppose in some underdeveloped country a 95% confidence interval for the proportion of children with compromised nutritional status was (20%, 40%). Even though the confidence interval is quite wide, every value in that interval points to a problem that needs to be addressed. Even 20% is too high. Would it help (would it change public policy) to know the true figure more precisely?*

Dallal illustrates the point further with an example from a leading statistician, William Cochran, who described an anthropologist who wanted to know the incidence of a blood type on the island he was studying to within 5% in order to make a classification decision regarding their likely genetic ancestry. Dallal quotes an extended passage from one of Cochran's widely-used books and adds his own commentary (in brackets):

> *An error limit of 5% in the estimate seemed to him small enough to permit classification into one of these types. He would, however, have no violent objection to 4 or 6% limits of error.*
>
> *Thus the choice of a 5% limit of error by the anthropologist was to some extent arbitrary. In this respect the example is typical of the way in which a limit of error is often decided on. In fact, the anthropologist was more certain of what he wanted than many other scientists and administrators will be found to be. When the question of desired degree of precision is first raised, such persons may confess that they have never thought about it and have no idea of the answer. My experience has been, however, that after discussion they can frequently indicate at least roughly the size of a limit of error that appears reasonable to them. [Cochran had a lot of experience with sample surveys. I don't. I have yet to have the experience where investigators can "indicate at least roughly the size of a limit of error that appears reasonable to them" with any degree of confidence or enthusiasm. I find the estimate is given more with resignation.]*

> *Further than this we may not be able to go in many practical situations. Part of the difficulty is that not enough is known about the consequences of errors of different sizes as they affect the wisdom of practical decisions that are made from survey results. Even when these consequences are known, however, the results of many important surveys are used by different people for different purposes, and some of the purposes are not foreseen at the time when the survey is planned.*[14]

Finally, it should be noted explicitly that the presence of classical measurement error, which is by definition random, does not tilt the disability determination process in favor of either the acceptance or denial of claims on its face. Systematic bias in ratings is clearly undesirable, but the consequences of random error are not obvious *à priori*. Insofar as errors cancel out within occupations or do not push occupations across category thresholds, such as sedentary/non-sedentary, they have no practical import at all. Insofar as they generate random misclassification of some occupations across thresholds then it may increase rates of both false positives and false negatives in disability determination at Step 4 but the magnitudes are unknown and possibly difficult to determine because related occupations suitable for the claimant might be subject to misclassification in the opposite direction. Similar issues apply to imprecise measurement of economy-wide job demands used at Step 5. Dallal's question is apposite, *How much would it help the overall accuracy rate of job classification and disability determination to know the true figure more precisely?* (i.e., how much difference does it make in practice?). The existence of random error always generates unease—*everyone would prefer perfect measurement*—but the mere presence of random error is insufficient to demonstrate (or exclude) any kind of specific harm or benefit to claimants or claim-deniers because such error generally cancels out when averaged within (and across) occupations. Of course, if some *true* within-occupation variation is mistakenly assumed to be error and averaged away then information is lost but there is no strong reason to expect that within-occupation heterogeneity occurs systematically on one or the other side of occupational means, though that does not exclude the possibility that it happens to do so. Again, systematic bias is an entirely separate issue and discussed in other sections.

Several points emerge from this consideration of general issues. Assessing reliability and validity appropriately requires prior statement of the exact purposes and procedures to which measures will be put (e.g., individual scores vs. averages; dichotomous vs. polytomous classification vs. continuous scores). Essentially no measure approaches perfect reliability or validity but random error per se (imperfect reliability) has no intrinsic practical consequences assuming reasonable sample sizes and the use of averages, though the possibility of such consequences cannot be ruled out entirely à priori. Indexes of reliability and validity should be as large as possible but are continuous measures, with few natural cutoffs or breakpoints to help decide how large is large enough, and users are rarely able to supply such standards on a reasoned basis. These conditions are not unique to ORS or any individual survey or study—they are normal, as the quoted passages from the Federal employment hiring guide, a leading test and measurement psychologist, and a widely-consulted text on health research indicate. Indeed, the dilemma of deciding at what point to discretize a continuous variable is common to any Federal agency or unit when setting standards involving money, such as the poverty line, minimum

[14] http://www.jerrydallal.com/lhsp/size2.htm accessed 12/16/14. Also available as The Little Handbook of Statistical Practice, by Gerard Dallal at http://www.amazon.com/dp/B00847SM6A. Again, text within brackets in the penultimate paragraph is from this source, so "I" refers to Gerard Dallal in that passage.

wage, and SSA's cutoff for substantial gainful activity; observing natural breaks is rare and standards must be chosen on a pragmatic and inevitably somewhat arbitrary basis, whether the continuous metric is money or a reliability/validity index.

Insofar as one wants additional foundation for standards one can consider the track record of existing similar measures and the number and importance of any decisions affected substantively that would result from using different standards. For the ORS, one might ask a study to determine, for example, how many or the total share of jobs that would be misclassified as sedentary or not as a result of observed limitations in reliability and validity using the best feasible method of data collection compared to a clearly superior but more costly method that provides a suitable benchmark ("gold standard"). If the share of jobs misclassified and the consequences of that misclassification are small by any reasonable definition then the less costly method has acceptable reliability and validity.

Finally, it is to be expected that cases closer to cut-points will be misclassified more often than cases falling clearly on one or another side of the threshold, and there will be greater disagreement among raters and rating methods because the task requires finer discrimination among trait levels, which is more difficult (Rindskopf 2002, p.13026). Police officers monitoring traffic in a 30 mph zone have no trouble using a coarse technique like simple visual inspection to classify cars traveling 15 mph and 45 mph as under and over the legal limit, respectively. However, consulting a well-calibrated radar speed detector may be necessary to achieve anything close to similar accuracy for cars traveling at speeds like 28 mph and 33 mph because those tasks require much finer discriminations. The bottom line is that the level of instrument precision required is contingent, not absolute.[15]

For the ORS, this implies more resources (survey items, occupational detail, intensive data collection) should be used to gain more precision for the most critical situations, such as jobs or occupations near the border of a trait continuum that separate sedentary from non-sedentary jobs. If a case "misses by a mile" a certain standard of interest then seeking high precision or exact values simply wastes resources collecting more information than will be used. Directing resources toward cases that are more difficult to classify will increase reliability and precision of measurement where it counts. Of course, this assumes one knows or has a set of standards to guide such targeting.

B. CURRENT AMBIGUITIES AFFECTING ORS RELIABILITY AND VALIDITY ASSESSMENT

One of the main difficulties in providing precise guidance regarding reliability and validity studies relevant to the ORS is that its data collection procedures and the exact purposes or uses to which the information will be put remain unclear in several key respects. While some of this uncertainty motivated the request for this report and some reflects the immense body of formal and tacit knowledge that attaches to SSA's operations, which are not easily communicated in compact form, the ambiguity regarding SSA's exact needs and goals and BLS' feasible set of data collection procedures makes the character of this report necessarily tentative at points.

[15] Kane (1996, pp.366f.) discusses the issue of measurement precision in the context of classification decisions using criterion-referenced tests.

Given their importance for understanding issues of reliability and validity, issues relating to both SSA and BLS are discussed below, as well as some of their implications for investigating reliability and validity of the ORS.

1. SSA ambiguities

The greatest strength of SSA's approach to occupational information is its emphasis on measuring concrete facts and behaviors, which are generally easier to measure than more abstract constructs. The main caveat to this generalization is that SSA seeks to measure duration and frequency of certain behaviors much more precisely than is commonly the case, indeed, perhaps more precisely than the DOT. This may present cognitive difficulties for respondents and practical difficulties for field analysts. There are other areas in which SSA's needs are quite unclear and therefore it is difficult to say how well they could be met by different data collection procedures that might be considered. A measure is valid insofar as it measures what it is intended to measure. The problem is it is not clear at the time of this writing exactly what it is SSA intends. In the absence of fuller information it is not easy to assess SSA's actual needs and, consequently, how they might best be met.

At the most basic level, ORS must produce a complete variable list containing short, clear definitions, level of measurement (continuous, polytomous, dichotomous) and labels or definition of units, and whether the variable will be used individually or as an intermediate in the construction of a final index. Definitions of variables and units for all indices should be included in the list. Specific recommendations regarding methodological issues are constrained by the lack of clarity on these points because the meaning and measurement of reliability and validity is situation-specific. As explained in a later section, the statistical methods for continuous and categorical differ greatly, so no concrete recommendations are possible until this basic information is clarified. In addition, as noted, the finer the measurement units (continuous, polytomous, dichotomous) the more resources are required for measurement and the greater the potential for error. Subsequent sections will show the appropriate indices of reliability and validity often differ by level of measurement, as well. Likewise, one of the biggest questions is whether individuals' scores will be used in a substantive fashion beyond calculation of occupational averages because averaging continuous variables typically increases measured reliability substantially, though it also purges any true within-occupation variation, as well. The treatment of measurement error for categorical variables is a little more complicated because averaging is inappropriate, but the assumption of symmetrically distributed measurement errors is potentially problematic if there strong floor and ceiling effects for polytomies.

Likewise, SSA needs to provide greater detail on how the ORS data elements will be used in disability determinations so that individual issues can be addressed and broader questions regarding the targeting of resources can be considered. For example, if there is a variable, such as excessive workplace noise, that is to be scored dichotomously and used on its own rather than as part of a composite index, how is within-occupation variation to be treated? An informal field test found that the noise from electric jackhammers used for road repair that are currently under consideration for adoption in New York City measured 15 decibels lower than the noise from

conventional pneumatic drills used currently, which register around 100 decibels.[16] Because decibels are measured in common logs, small differences in scale values often represent great differences in noise levels. Although OSHA and NIOSH use different standards, both recognize that the difference between 100 db and 85 db is large enough to alter a job's categorization,[17] and it is conceivable that the same is true for SSA's classification of noise levels. If the new drill were to diffuse across *jobs* within *occupations*, even those defined finely enough to satisfy all stakeholders (e.g. DOT occupations), it is not clear how this heterogeneity is to be addressed, i.e., how are *occupations* to be scored on a dichotomous variable if the true (error-free) proportions for some variable are 0.55 vs. 0.45 or some other difficult split? This situation becomes only more complex in the case of polytomies and in the presence of measurement error, which clouds determination of the level of heterogeneity itself quite apart from the question of how to address it. If SSA wants to use the frequency distributions of categorical variables and the quartiles of continuous variables to characterize occupations, this would be a reasonable choice but the appropriate reliability/validity indices would have to be selected accordingly because IO psychology usually uses job averages on the assumption that all within-title variability is random error (Harvey 1991, p.110ff.). How SSA would like to treat various forms of within-occupation variation must be made clear at some point if the appropriate methods for validation studies are to be determined.

Given the ambiguities of distinguishing measurement error from heterogeneity within occupations, one could perform sensitivity analyses that assume, alternately, that all within-occupation variation is random error and that all variation represents genuine heterogeneity to see how much difference these polar assumptions make to the frequency distributions of different job characteristics over the full sample. If the differences are not large enough to matter for SSA's purposes then occupation-level measures of central tendency are sufficient. If there are consequential differences between the two approaches then there is the possibility that *occupation-level* measures do not provide a completely accurate guide to the distribution of *job* characteristics in the national economy. A consequential difference could be one that alters a characteristic's prevalence based on the criterion of its presence or absence in "significant numbers" of jobs, which an element of Step 5 decision-making. Unfortunately, no operational definition of "significant numbers" could be found in the materials examined for this report, so this is another issue that may require clarification. In fact, it will be very difficult to interpret reliability/validity studies if there is no accepted cutoff in the SSDI determination process against which one can evaluate the reliability levels of different methods of data collection and processing.

The level of informational detail desired is also ambiguous. Step 4 decisions and the use of skill transferability analyses in the determination process are possibly the motivations for SSA's desire for detailed profiles of job requirements across the entire occupational spectrum. This would seem to require a relatively comprehensive occupational information database and clearer decisions regarding the treatment of within-occupation variation.

---

[16] Sam Roberts, "With Electric Jackhammers, Plans to Quiet an Earsplitting City Sound," New York Times, 12/17/2014, http://www.nytimes.com/2014/12/18/nyregion/electric-jackhammers-and-revised-noise-rules-may-help-quiet-new-york-city.html (accessed 1/28/2015)

[17] See "Occupational Noise Exposure," https://www.osha.gov/SLTC/noisehearingconservation/ (accessed 1/28/2015)

By contrast, Step 5 appears to set a lower threshold for denying claims because it is based on whether the claimant can perform any "kind of substantial gainful work" found in significant numbers given the applicant's age, education, past work SVP level, and RFC. This appears to motivate SSA's particular interest in the prevalence of *low-skill, sedentary jobs*, which is a much smaller subset of the information needed for Step 4. These may be high priority cases warranting additional ORS resources for study.

If the goal is to make relatively few categorical classifications of job requirements, as implied by Step 5, then the task is greatly simplified because one would focus on relatively fewer borderline occupations rather than devoting extensive efforts to make equally reliable and fine discriminations of job requirements at all levels. If profiles are required for all occupations, even relatively rare ones, at a uniform level of detail and reliability/validity then the ORS' task becomes much greater.

If there is some intermediate position such that more detail is desired for jobs meeting Step 5 criteria then it would be helpful to have some guidance on the relative degree of precision desired for these two components of the database given that it is easier to meet a more demanding level of precision for a subset of jobs as opposed to the entire set. A sound estimate of how many jobs require low skill and physical effort is considerably easier to achieve than estimates of more detailed levels of skill and physical demands across the spectrum of jobs. As the preceding section tried to make clear, the uses to which the data will be put need to be considered carefully and to be well-specified because they influence strongly the information that is actually needed, its cost, and the standards and methods of reliability and validity used to investigate its quality. The question is, Are all occupations to be profiled in equal depth or is it reasonable to focus more attention on the occupations that result in most applicants for benefits and their near-substitutes? If this is an acceptable balance between level of detail and feasibility, then the question becomes the degree to which SSA claimants are concentrated in a small number of jobs/occupations and skill/exertion levels.

Illustrating the possibilities for targeting resources are findings on the occupations, skills, and other job requirements most frequently cited by SSA disability claimant population. Ferguson (2010) drew a random sample of 4,000 administrative records out of a database of more 1 million jobs with 21,000 unique job titles. The database contains 21,000 unique job titles but they include highly detailed occupations because the data is collected in free text form rather than in standardized codes (e.g., "Cashier for a restaurant," "Burger King Cashier," "Cashier at Burger King," "Grocery Store Cashier," "Cashier at Superfresh"). Analyses showed that 20 titles accounted for 38% of claimants (see table). Clearly, there is substantial concentration of claimants in terms of their previous occupation, which gives a considerably different impression from claims that the ORS needs to be at the same level of detail as the DOT.

THE MOST FREQUENT *CATEGORIZED* OCCUPATIONS OF 2008, HELD BY DISABILITY CLAIMANTS*

Categories of Occupation*	National Population	Percent of Total Population	Sample Size**
CASHIER	51,256	4.9%	513
LABORER/GENERAL	51,213	4.9%	512
DRIVER	49,117	4.7%	492
CONSTRUCTION	44,220	4.3%	442
SERVICE	35,789	3.5%	358
MANAGER	22,638	2.2%	226
SALES	21,634	2.1%	216
ADMINISTRATIVE	17,693	1.7%	177
CLERK	14,805	1.4%	148
HOUSEKEEPING	13,455	1.3%	135
MAINTENANCE	12,371	1.2%	124
WAITRESS/WAITER	11,814	1.1%	118
SECURITY/POLICE/GUARD	11,083	1.1%	111
JANITOR	10,763	1.0%	108
TEACHER	10,118	1.0%	101
MACHINE	10,075	1.0%	101
ASSEMBLY	8,485	0.8%	85
Total Occupations 1,035,641*	396,529	38.3%	3967

A similar study of 5,000 randomly chosen records from 2009 by Trapani and Harkin (2011) found "substantial limitations in the quality of occupational information that SSA obtains from claimants and in the applicability of the DOT taxonomy to our current caseload." About 16% contained insufficient information to identify the occupation, 1.4% matched no DOT title, and 4% matched a combination of jobs. It seems that the reliability of past determinations has been affected by accuracy and reliability issues pertaining to the claims processing activity itself, as Trapani and Harkin cite "Limitations in occupational information obtained from claimants," as well as "Limitations using the DOT" as challenges for both their study and SSA occupational assessments. In response to an OIDAP member's question at the public presentation of findings, Harkin elaborated that the problem of insufficient occupational information on claim forms "is not caused by poor training but by the overwhelming caseloads that adjudicators [SSA examiners] carry and the complexity of the forms that are sent to claimants."[18]

The most common titles Trapani and Harkin identified in their claimant sample were similar but not identical to those found in Ferguson's study, with 11 titles accounting for 27.7% of all cases (see below). The most common 50 titles account for 45% of all prior jobs in the records sampled. Of the 5,274 instances of past relevant jobs associated with the study claimants, the authors could identify 1,171 distinct DOT titles, which comprise about 9 percent of the total number of titles listed in the DOT, as the authors observe. Insofar as the current SOC contains

[18] Occupational Information Development Advisory Panel Quarterly Meeting Minutes, May 4, 2011, p.12. A survey of vocational experts found that 30% of cases contained insufficient information given by the claimant to identify the jobs (*ibid.*, p.20).

insufficient detail it seems likely that a study of these titles could provide a reasonable number of additional codes that would cover the most common contingencies. Indeed, the authors concluded, "Relatively small number of job titles account for relatively large proportion of work performed by claimants, suggesting that targeted OIS data collection can produce information broadly applicable to SSA claims." This observation may hold the key to reconciling stakeholder desires for occupational detail and the practical requirements limiting such detail.

1. Cashier-Checker (4.2%)
2. Nurse Assistant (3.8%)
3. Fast-Foods Worker (2.7%)
4. Home Attendant (2.6%)
5. Cashier II (2.6%)
6. Laborer, Stores (2.4%)
7. Material Handler (2.1%)
8. Packager, Hand (1.9%)
9. Stock Clerk (1.8%)
10. Cleaner, Housekeeping (1.8%)
11. Janitor (1.8%)

The distribution of strength levels for jobs with sufficient information is below. In the absence of knowledge regarding the distribution of jobs across strength levels in the workforce overall it is difficult to draw firm conclusions, but it seems likely that "sedentary" jobs are significantly under-represented in this claimant sample. This also provides some perspective on where data collection efforts might be targeted.

1. Sedentary – 11.6%
2. Light – 35.3%
3. Medium – 39.6%
4. Heavy – 11.4%
5. Very Heavy – 2.1%

Trapani and Harkin found the distribution of SVP given below and concluded, "a substantial majority of the jobs held by our claimants have been unskilled (22.4%) and semi-skilled (40.4%) jobs that required a relatively short time, from < 1 to 6 months, to learn."

1. SVP 1 – 0.6% (of all SVP citations for Past Relevant Work)
2. SVP 2 – 21.8%
3. SVP 3 – 23.7%
4. SVP 4 – 16.7%
5. SVP 5 – 7.9%
6. SVP 6 – 10.0%
7. SVP 7 – 15.6%
8. SVP 8 – 3.8%
9. SVP 9 – 0.0%

The five most frequently identified SVP-Strength combinations comprised nearly half of all jobs with sufficient information:

1. SVP 3-Light (10.6% of all SVP-Strength citations for past relevant work)
2. SVP 2-Medium (9.5%)
3. SVP 3-Medium (9.2%)
4. SVP 2-Light (8.3%)
5. SVP 4-Medium (7.8%)

The twenty most commonly cited functional limitations in the Trapani-Harkin sample are listed below. The first ten limitations account for nearly 56% of all limitations cited in the sample, and the twenty comprise about 83% of all limitations cited. Trapani and Harkin comment, "Exertional and Postural limitations represent the most prevalent categories of functional limitations cited in our case files, but various categories of mental limitations are also cited relatively frequently." Those that appear to be mental limitations are listed in bold below. Again, this information provides some indication of where ORS items on job demands may be best targeted.

1. Lift/carry occasionally (76% of all cases)
2. Lift/carry frequently (76%)
3. Stand/walk (76%)
4. Sit (75%)
5. Climbing ladder/rope (54%)
6. Climbing ramp/stairs (40%)
7. Crawling (39%)
8. Crouching (39%)
9. Stooping (37%)
10. Kneeling (35%)
11. **Maintain attention (30%)**
12. **Carry out detailed instructions (29%)**
13. Balancing (29%)
14. **Understand detailed instructions (28%)**
15. **Avoid hazards (28%)**
16. **Complete workday (28%)**
17. **Respond appropriately to changes (24%)**
18. **Interact with public (23%)**
19. **Accept instructions from supervisors (19%)**
20. **Perform within schedule (15%)**

Trapani and Harkin found that Step 5 decisions are most frequently made based on grid rules and that the five most frequently encountered situations were (in descending order of prevalence):

Vocational Rule	Residual Functional Capacity	Age	Education	Past Work	Decision
204.00	Heavy	All	All	All	Not Disabled
201.06	Sedentary	Advanced age (Age 55 and over)	High School Graduate or More	Skilled or semiskilled	Disabled
202.21	Light	Younger individual (Under age 50)	High School Graduate or More	Skilled or semiskilled	Not Disabled
202.06	Light	Advanced age	High School Graduate or More	Skilled or semiskilled	Disabled
202.14	Light	Closely approaching advanced age (Age 50-54)	High School Graduate or More	Skilled or semiskilled	Not Disabled

2. BLS ambiguities

The ORS instrument reviewed for this report has many valuable qualities. It is thoughtful, precise, and concrete. But there are also a number of ambiguities in the data collection methods and post-collection processes that need to be clarified before studies of reliability and validity can be designed. BLS seeks to ascertain optimal data collection methods, meaning maximum reliability and validity within the constraints of feasibility. There appear to be a number of relevant decision variables that could be the subject of experimentation.

a. Who or what are the sources of information

Job incumbents are a major source of information in almost all methods of job analysis. It is a truism that job incumbents know their work more intimately than anyone else. Information from incumbents is often supplemented and triangulated with information from supervisors, managers, HR officials, and official job descriptions. The latter methods cost less, are often more convenient, and may be the only possible sources given employer preferences regarding access to their employees and hesitation regarding worker downtime. External observers also may avoid the problem of a *limited frame of reference* that might affect incumbent responses when questions involve some elements of subjectivity, such as ratings, or comparison to conditions in other jobs, whether implicit or explicit. The current design of the ORS seems to involve little use of incumbents as sources of information, though their use in validation exercises has been mentioned. This raises a number of issues.

The existing literature is not so consistent that one can speak confidently about either (1) the *existence* of great differences between responses of incumbents and non-incumbents in general

or (2) the *relative quality* of measures that rely on one or the other, or (3) on some combination of their responses. Such studies are not very numerous and their design and measures are both heterogeneous with respect to one another and different from the kinds of questions in the ORS, which serves as a caution regarding their use as guidance for ORS practices. Some studies are discussed in the literature review section.

Nevertheless, given this uncertainty, it is strongly advisable to conduct experiments on the effect of respondent identity on measurement quality. These experiments should determine the magnitude of measurement variation across sources (*consensus*), as well as within them (*stability*), and the relationship between each source and a common gold standard measurement, such as physical instruments or job observations by a field economist who has not participated in the interviews to avoid contamination (*accuracy/bias/validity*).

The first paragraph in this section highlighted *cognitive* reasons for differences in accuracy by source, but it is also important to note *motivational* reasons. Incumbents may inflate their autonomy or decision making power due to conscious or unconscious self-presentation or self-enhancing motives (Sanchez and Levine 2012, p.411). Possible motives for incumbents' misreporting physical demands are more complex as there is no direct incentive for them to do so. By contrast, there are clear incentives for company officials to understate physical demands and under-report the severity of adverse working conditions given their potential implications for worker's compensation costs and drawing unwanted attention from Federal and state government enforcement agencies such as OSHA and EPA. For this reason it is **essential** that ORS have some systematic channel for incorporating incumbent perspectives, particularly with respect to occupations that prior knowledge leads one to expect have significant physical demands and negative environmental conditions.

Finally, there are substantive reasons for an incumbent survey. While it is possible to derive measures of central tendency at the occupational level from an employer survey and use other sources, such as the OES, CPS, and Census for weighting purposes, it may be more difficult to employ such a strategy if one wants to use measures of within-occupation variation in the ORS database, such as percentiles or category frequencies. While it may be possible to derive estimates, it is certainly more straightforward to survey individuals and derive means, percentiles, and category frequencies for both occupations and the overall economy directly from the job-level microdata.

Likewise, while the general thinking has been to organize the ORS database around the concept of occupations, it is useful to note that a probability sample of workers would also be a probability sample of *jobs*. The stakeholder objection that occupations should be coded at what is effectively an infinitely fine level of detail would be met by a database that could summarize job trait prevalence with job-level information. Information on traits that jobs in different occupations share with one another but not with other jobs in their respective occupations would be preserved in this database, rather than averaged away. One could run frequency procedures on one or a flexible combination of characteristics and determine the number of *jobs* in the economy that satisfy the query. As with every design, this suggestion is not without limitations. The identity of the jobs would be based on incumbent-supplied titles and descriptions, not SOC codes, so standard Census/CPS procedures would have to be tightened to make sure they are

informative. The survey questions on job characteristics would have to be interpretable for respondents regardless of education level and language. Finally, the issue of measurement error would need to be considered more deeply. Refraining from using the standard solution of averaging to eliminate random error will preserve heterogeneity—but it will also preserve the error that was the object of averaging itself. It is difficult to avoid this tradeoff. Nevertheless, OIDAP did hear presentations from the ACS top staff in considering the possibility of a supplement for job-holders on the nature of their work. ACS annual samples are about 200,000 persons per year (including non-workers), so there is clearly potential for obtaining numerous ratings per job.[19] This report did not uncover information on OIDAP's final evaluation of this option.

b. Where and how to collect information

BLS collects some job information through personal visits to workplaces, but most data is gathered remotely by telephone and, to a lesser extent, email. BLS seems to have an ongoing dialogue internally regarding the relative merits of standardized interviewing as opposed to guided or conversational interviews. The latter give field economists greater discretion to vary procedures according to the situation if they believe it is conducive to greater cooperation and highest quality information. This issue is discussed further in reviewing internal reviews of ORS practices (Section III D).

Site visits may involve interviews only or a combination of interviews and job observations. ORS experience reviewed in Section III D suggest in-person interviews are likely to be higher quality than telephone interviews because interviewees tend to be more responsive and engaged in face-to-face interaction. Given cost considerations, telephone interviews are likely to be more common, so this source of variation should be studied in experiments and validation exercises.

However, the most prominent issue is the possibility of supplementing various kinds of reported information with direct observations of jobs by BLS field economists in personal visits. Various stakeholders expressed the strong wish that the ORS follow the DOT in using direct observation by trained raters as the primary method of data collection. Some IO psychologists, who conduct research on a much smaller scale than the ORS, have supported this position. There is significant concern over whether the quality of non-observational methods will be similar to the quality of DOT. However, BLS has indicated that cost consideration preclude primary reliance on personal visits and job observations, as does the policies of some employers regarding access to their workplaces and employees. Indeed, the DOT's cost is one reason it has not been updated. IO psychologists themselves have long known that "collecting task-based data remains a very costly undertaking" and the general scarcity of such measures "is probably due in large part to the costs involved," though there have been suggestions that automating the collection of incumbent responses might reduce expenses sufficiently (Harvey 1991, p.117). Web-based data collection using effective branching based on prior responses may be one way to realize this suggestion, and some believe automated natural language processing has reached the point at

[19] Occupational Information Development Advisory Panel Quarterly Meeting Minutes, May 4, 2011, p.16f.

which detailed task information can be scraped from employer and job positing web sites, but this would require separate study.[20]

In fact, it is likely that job observations are not critical for most office jobs because observation of most non-manual work is unlikely to be much more informative than interviews. Therefore, that the real choice is not between using personal observation across the board or not at all, but rather how much job observation by field economists can be used for the occupations and industries in which it is likely to provide significant informational benefits relative to interviews alone. This prompts three recommendations:

> **Recommendation:** BLS should produce a rough costing for on-site job observations so both ORS and stakeholders have some sense of the costs involved and how much use of this method is feasible. Variables whose costs may exceed greatly their incremental value in terms of capturing variation (e.g., rural location) should be a focus for economizing resources in this costing exercise.[21]
>
> **Recommendation:** ORS should construct a list of target occupations and industries for which job observation and other intensive methods would be most beneficial due to (1) potentially high ambiguity regarding their classification into exertional levels or other relevant categories; or (2) their prevalence in SSA caseloads. The general principle is that when a resource is scarce it can be used most efficiently by targeting it to the most critical areas.
>
> **Recommendation:** Independent of the two prior recommendations, on-site interviews and job observations should be used as validity criteria and gold standards in the methodological experiments and validation exercises recommended in this report wherever possible.

Given the intrinsic arguments in favor of in-person interviews and job observations, and the intensity of stakeholder views, BLS should make efforts to incorporate them into the ORS to the extent feasible. This will narrow the gap between the program and stakeholders' wishes and bolster the quality of measures directly and through their use in validation.

c. How to process information

BLS's final quality assurance practices are not clear from documents provided. There may be large differences in reliability depending upon the specific procedures used (or not used) to reconcile inter-rater differences. ORS project documents suggest substantial efforts were made to reconcile differences rather than simply accept them as given and this has potentially large

[20] See, e.g., Marc Anderberg (2012), "Job Content and Skill Requirements in an Era of Accelerated Diffusion of Innovation: Modeling Skills at the Detailed Work Activity Level for Operational Decision Support," Employment and Training Administration, U.S. Department of Labor, pp.56ff.

[21] Although representativeness is desirable, results of the Rural Manufacturing Survey conducted by the Economic Research Service, USDA found relatively small differences generally on a wide range of variables, though specific implications for ORS variables would require investigation. See "Rural Competitiveness: Results of the 1996 Rural Manufacturing Survey," by H. Frederick Gale, David A. McGranahan, Ruy Teixeira, and Elizabeth Greenberg. Agricultural Economic Report No. 776 (1999). Economic Research Service, USD A.

effects on reliability, validity, and accuracy. However, it is not clear how much of this activity reflects project start-up, regular practice, or something wholly or partly inapplicable to a switch to primarily incumbent self-complete questionnaires. While it is desirable to have high-quality information at the initial rating stage, any experiments or validation exercises will under-estimate the final reliability and validity of measures insofar as they omit subsequent quality control checks that are anticipated to be part of the program's standard operating procedures. It is good practice for experiments or validation exercises to isolate the different sources of significant variation but they should also mirror final practice as much as possible. ORS will have to identify the scenarios that are sensible to test according to its own determinations.

C. A CLOSER LOOK: MEASURING RELIABILITY

Given the preceding considerations one can be more precise about the meaning and measurement of reliability. Reliability refers the extent to which a measurement method gives reasonably consistent and stable results across replications, i.e., highly similar values when the same or equivalent instruments are used repeatedly to measure the same or equivalent targets. A single scale for measuring the weight of physical objects is unreliable if it produces widely varying readings when the same object is placed on it repeatedly—a single reliable instrument gives relatively stable values across repeated administrations. Similarly, if two scales assumed to be interchangeable indicate widely varying weights for the same or equivalent objects then neither scale can be considered reliable (or valid) in the absence of additional information. If an instrument or method is reliable then multiple versions would show relative consensus when evaluating the same object. If a human rater takes the place of the weight scale, the first situation is a form of *intra-rater* reliability, often called *test-retest reliability* (or rate-rerate reliability), while comparisons of multiple raters judging the same target approximates the second situation and is a form of *inter-rater* reliability. The common thread across different methods is that reliability is concerned with the replicability of measurement under comparable circumstances.

For the ORS, this means that if respondents in a validation exercise are surveyed initially and then re-surveyed a few weeks later they should give roughly comparable answers to the same questions on both occasions. If the activities performed recently on the job vary in normal fashion across the two occasions and evoke different recollections and responses, reliability will be lowered. If the sample of cases or occasions is large enough to reasonably represent the range of situations that the job faces, then such differences may average out at the occupational level. As with any effort to measure test-retest reliability there is always the issue of balancing two risks: (1) if the time interval between administrations is short then reliability estimates can be biased upward by memories of first-round responses or the similarity of job activities within the time interval, but (2) if the interval between administrations is long then reliability estimates can be biased downwards because stable job characteristics are more likely to have changed genuinely in the interim.

Inter-rater reliability can be measured in several ways. An experiment could have one field economist interview a respondent and compare how similarly two other field economists observing the interview independently code the responses. This inter-rater reliability exercise is comparable to the situation of comparing the weight of a single object as recorded by two

different scales, and gives some indication of the reliability of individual interviews as would be conducted normally by field economists. Video-taped interviews and taped interviews plus job observations could be used to scale up this experiment and measure inter-rater reliability across field offices.

A less labor-intensive method of measuring inter-rater reliability would compare the responses of different incumbents in the same job who answer a self-completed survey. However, unless the assumption the jobs are identical has a very strong basis one would expect such reliabilities to be lower than in the previous cases because even the "same" jobs within establishments are rarely identical and therefore responses will reflect some unknown mix of true variation in task requirements and error variance across incumbent-raters.

If one measured the inter-rater reliability of field economists' ratings for different incumbents of the same job then the measure of reliability may pick up similar inter-rater error variance among the field economists along with the variation described for the previous method (incumbent error variance + true variance across the "same" jobs that is difficult to distinguish from measurement error). However, field analyst ratings of different incumbents need not depress actual reliability relative to incumbents' self-completing surveys if the field economists' active participation reduces the impact of error variance among incumbents through probes, clarifications, and exercise of more accurate judgment in field coding responses. Clearly, this also raises the issue of the relative merits of standardized vs. conversational interviewing methods, which is an active point of discussion in ORS project documents. **It should be noted that the large element of discretion granted field economists in interpreting respondents' answers and collaborating with them in articulating responses implies essentially two sources of "rater" variation in the ORS, in contrast to job observations and closed-coded survey responses where *either* the field analyst or respondent alone performs the rating task.**

It is likely that rate-rerate reliabilities will exceed inter-rater reliabilities by considerable margins due to (1) the latter's inclusion of unobserved within-occupation heterogeneity, (2) the varying styles (leniency/severity) of different raters judging identical jobs, and (3) the built-in floor for intra-rater reliabilities resulting from single-raters' recollections and stable styles across occasions. Sole reliance on rate-rerate reliability metrics may give an optimistic picture of data quality given that ORS plans to use multiple raters.

Nevertheless, an important property of test-retest reliability worth noting is that it controls for within-occupation heterogeneity, treating the respondent as the sole rater for the moment. Assuming a job does not change between rating occasions, differences in responses across occasions reflect intra-rater unreliability only, not true differences in the jobs rated or different rater styles (Sanchez and Levine 2012, pp.402f.). By contrast, inter-rater reliability, which is arguably more critical, is not a "clean" measure because it may be influenced by all three sources of variation (target job, rater identity, and occasion or "pure noise"). **It would be very valuable for ORS to consult with a specialist to see if it is possible to devise experiment(s) that can distinguish these three components and estimate their absolute magnitudes for various SOC occupations that represent critical cases due to their frequency in SSA caseloads and variation across the range of key trait(s), such as exertion level.** Cronbach's generalizability theory may

be a useful framework to consider.[22] The desired outcome of such a study could be finding that true-score variation represents a low percentage of total variation within occupations, which would support the averaging scores within occupations for the ORS database. (This is contrary to the aim of most such studies, like those examining educational test scores, which seek maximum true-score variation and minimum noise because the individual-level value will be the basis of some decision.) Such a study could shed much light on the question of the magnitude of within-occupation heterogeneity and its implications for ratings at the 8-digit SOC level of aggregation.

While the literature review section will discuss specific reliability estimates, some general points can be made at the outset. It is both common sense and a truism of survey methodology that items that pose cognitive difficulties for respondents will have lower reliability, as well as lower response rates, all else equal. Questions that require exceptional attention or memory for characteristics that respondents treat as low-salient or unmonitored background detail may be subject to widely varying guesses among different respondents performing the same tasks. Likewise, questions tend to have higher reliability if they are relatively straightforward and relate to objective behaviors, facts, and events. Survey questions that do not require a great deal of personal interpretation, and are concrete and factual are generally more reliable than questions that are more abstract.

Many of these considerations are potential problems with many O*NET items and also imply that much of the research literature that investigates the reliability of abstract constructs, such as work autonomy, does not have great relevance for the ORS, most of which consists of highly concrete questions. However, ORS project documents suggest that if respondents are asked to complete a survey on their own asking precise frequency or duration for various physical tasks and subject to many caveats and conditions ("stooping isn't bending," etc.), there will be more measurement error than at present. In other words, there is a built-in conflict between the exactitude of the questions and the reliability of respondents' answers because of the guesswork involved in mapping unmonitored aspects of the flow of daily life into categories defined by such precise distinctions. In this context it is useful to recall that human memory and perception differ from scientific instruments, such as weight scales or decibel meters.

There are common methods that can reduce the "noisiness" of measures significantly. Reliability increases with the length of the assessment assuming the items help discriminate the trait level of the target of measurement. This argues for following the example of computer-adaptive testing and constructing the survey so that filters and branching direct respondents to a

[22] For an introduction, see, e.g., Amy M. Briesch, Hariharan Swaminathan, Megan Welsh, Sandra M. Chafouleas (2014), "Generalizability Theory: A Practical Guide to Study Design, Implementation, and Interpretation." Journal of School Psychology, 52:13-35. For an application assessing ratings of physical task demands using videos of manufacturing production workers, see Angela Dartt, John Rosecrance, Fred Gerr, Peter Chen, Dan Anton, Linda Merlino, "Reliability of Assessing Upper Limb Postures Among Workers Performing Manufacturing Tasks," Applied Ergonomics (2009) 40:371–378. For an application to job analysis using DOT GED scores, see, "Generalizability of General Education Development Ratings of Jobs in the United States," by Noreen M. Webb, Richard J. Shavelson, Jerry Shea, and Enric Morello, Journal of Applied Psychology (1981) 66:186-192. For another perspective of decomposing rating variance in job analysis, see, Chad H. Van Iddekinge, Dan J. Putka, Patrick H. Raymark, and Carl E. Eidson Jr., "Modeling Error Variance in Job Specification Ratings: The Influence of Rater, Job, and Organization-Level Factors." Journal of Applied Psychology (2005) 90: 323–334.

more intensive line of questioning more finely targeted to their rough level of job demands. By contrast, a one-size-fits-all set of questions across the full range of difficulty can be expected to include large numbers that one could foresee would be answered either entirely correctly or incorrectly. This redundant information represents lost opportunities for more relevant items. Insofar as SSA prioritizes in-depth knowledge on a tractable subset of occupations it makes sense for ORS data collection efforts to prioritize them as well in terms of finer occupational coding detail, additional data elements, and focused validation exercises. Given scarce resources, if the criticality of information is spread unevenly across the occupational spectrum then the allocation of resources should be targeted to reflect that fact.

Finally, item reliability is almost always much higher if one is interested in occupational means rather than individual-level responses. If all that is desired are occupational means then much of the error variance is averaged out, assuming reasonable occupational sample sizes. As long as items do not suffer from significant bias, discussed in the next section, then many situations in which individual-level reliability is modest resolve themselves once they are replaced by means because random error among individuals averages out.

The different methods of measuring inter-rater reliability and the ways in which reliability varies by question type and level of aggregation are examples of the ways in which reliability must be considered in relation to a project's specific procedures, methods, and goals. If the ORS considers it essential to collect highly detailed information that respondents are unlikely to be able to recall on their own but cost considerations preclude personal visits, it will be considerably more challenging to collect reliable data than if one or more of those constraints were relaxed. If the ORS is trying to decide whether to use personal interviews or self-completed questionnaires then the relative levels of inter-rater reliability can be determined through a validation study. In terms of the present report, if the ORS has no interest in abstract questions or individual-level measurement, then there would be little purpose in reviewing reliability studies for such measures.

The classical definition of reliability is the ratio of true score variance to total score variance, or $\frac{var_{true}}{var_{true} + var_{error}}$, which can be seen as an analysis of variance concept corresponding to the proportion of total observed variance explained by true score variance, similar to an R^2 statistic. As one source says simply, "A reliability is a squared correlation" (Dorans 1999, p.3). Under standard assumptions this correlation measures the relationship between the observed values and the (unobserved) "true scores." This concept of reliability has been described as the ability of measurement to differentiate between the targets or classes of targets in the sample (Kottner et al. 2011, p.96). Expressing the idea in more general terms, Tinsley and Weiss wrote, "It is always necessary to establish the reliability of a set of ratings to demonstrate that the variance in the ratings is due to a systematic rather than a random ordering of the objects" (2000, p.96).

It is possible to measure similarity between two repeated measures using Pearson correlations, but this raises a problem worth understanding clearly given the ORS' goals. Standard bivariate correlations transform both data series by subtracting the means and dividing by their standard deviations of the respective variables, rendering all scores unit-free and measuring association as similarity across variables in ranking cases. This is desirable when variable units are intrinsically incomparable, such as education and income, or when different units lack any

absolute meaning or deterministic conversion factors, such as SAT and ACT scores. In these cases, the insensitivity of Pearson correlations to linear transformations is an advantage.

Standardizing is not desirable generally when the variables use common units and one wants to know the extent of absolute agreement ($y=x$) between measures, not simply the extent to which they are consistent with one another to within a linear transformation ($y=a + bx$). There may be occasions, such as grading essay exams, rating job interview performance, or other complex judging activities, where one might want to purge the relative leniency or severity of judges from the ratings and calculate a consistency-based inter-rater reliability index even when raters use a common scale (with different styles). The similarity of relative orderings or rankings may be all that matters in these situations, so that inter-rater differences in level and spread are nuisance factors. However, this is clearly not the case with the ORS, in which raters use the same objective metric to record factual information and leniency/severity biases are non-ignorable. Most ORS measures are criterion-referenced, i.e., the concern is whether cases (jobs or occupations) meet some predetermined standard or criterion. This calls for a measure of reliability that "summarizes agreement (is y equal to x?), not linearity (is y equal to some $a+bx$?)" (Cox 2004, p.336). SSA and BLS correctly, though not necessarily explicitly, have in mind the more demanding concept of reliability as absolute agreement, which Pearson correlations do not distinguish from the weaker standard of consistency: "…correlation measures relationship, not agreement. If one measurement is always twice as big as the other, they are highly correlated, but they do not agree" (Bland and Altman 2012). Indeed, there is evidence than job incumbents assign higher ratings to their jobs than job analysts, at least for abstract constructs that are more open to interpretation and more difficult to verify through observation; however, three studies found correlations between incumbent and analyst ratings between 0.67 and 0.93 (Tsacoumis and Van Iddekinge 2006, pp.2f.). Needless to say, the design of the ORS and experiments and validation studies should aim to minimize rater variance so scores reflect only differences in the targets rated and not the varying standards of the raters themselves to the extent practicable.

This report assumes some kind of absolute agreement rather than consistency measure of reliability will be used. It is useful to be clear regarding the limitations of Pearson correlations to understand the generally accepted measures of reliability that avoid them and the fact that most *validity coefficients*, which are Pearson correlations comparing different-unit variables, can only measure concepts similar to consistency rather than absolute agreement.

The best-known measure of reliability is the intra-class correlation coefficient (ICC), which has numerous variants depending on one's purpose. Variants that measure consistency rather than absolute agreement can be ignored for this report. Published ICCs using a consistency standard may be very misleading if misinterpreted as measures of absolute agreement and there are significant inter-rater differences in means or variances. When single raters (e.g., respondents) are not re-used to rate other targets then the ICC is similar to a one-way ANOVA; when a panel of raters is used for the same cases the ICC can control for rater effects similar to a two-way ANOVA.

However, selecting the appropriate ICC still requires deciding whether scores for the target of measurement are to be used as individual ratings (e.g., job-level scores), or averaged across

raters (e.g., occupation-level scores). Variants of ICC can indicate the average reliability of individual scores or the average reliability of mean scores. "If conclusions typically are drawn from the ratings of a single judge, the average reliability of the individual judge is appropriate…Conversely, when decisions are based on a composite formed by combining information from a group of observers, it is necessary to know the reliability of the composite" (Tinsley and Weiss 2000, p.108). If SSA wants an unconditional estimate of the incidence of job demands it is conceivable that the former calculation is appropriate; for occupational profiles, it is clear that the latter is applicable. For a given set of scores, the magnitude of group-level ICCs will be greater than individual-level ICCs, usually by a considerable margin. For example, if an ICC(1,1) were 0.20, then ICC(1,*k*) for *k*=22 ratings and *k*=35 ratings would be close to 0.85 and 0.90, respectively (LeBreton and Senter 2008, p.833). As can be inferred from this example and from similarities to issues of sample size and statistical power calculations, the gain in reliability is a nonlinear function of the number of raters with declining marginal returns, so the good news is that great improvements in reliability are feasible with relatively few ratings per occupation.

If panels of raters, such as field economists, are re-used across occupations, then the raters as well as the targets (occupation) can be modeled as either a random or fixed factor in computing intra-class correlations. The various choices are represented in the diagram below, where the first digit of the ICC code corresponds to the type of rater (unique rater per target or rater panel) and the second digit corresponds to the number ratings that is the unit of analysis (individual-level ratings or means from *k* ratings). Further consideration of which is most appropriate for ORS depends on the clarification of the program's intended data collection and use of the information collected.

UNIT OF ANALYSIS	**RATERS**		
		Panel	
Is the interest:	**Unique**	**Random**	**Fixed**
Individual job	ICC(1,1)	ICC(2,1)	ICC(3,1)
Occupation mean	ICC(1,*k*)	ICC(2,*k*)	ICC(3,*k*)

Note: *k* generally refers to the number of raters over which a measure is averaged.

A problem with ICCs is that they are sensitive to the sample's range of values on the measured trait. ICCs are similar to R^2 in the sense that the proportion of variance explained depends on the amount of true variance among the targets rated, as well as the average amount of error involved in rating each target. Sampling cases with a wider range of trait values will increase measured reliability for a given level of within-target error under this approach. This means, for example, reliability will always appear greater for systolic than diastolic blood pressures even when both are measured at the same time on the same person with the same device. This is not a result of diastolic being more difficult to measure but simply the way the math works out given their different ranges; it is a statistical artifact (Bland and Altman 1990, p.338). In other words, the measured reliability of an instrument will reflect both its own quality in some ideal sense and the nature of the sample used to derive a reference value to benchmark its readings. If true variability between cases is low then a high proportion of the observed variation in values will be error variance. However, if the group is very diverse, meaning true variance is high, then even undesirably high levels of within-group disagreement may be masked by high ICC values. The

potential for inflated measures of reliability, as well as other practical considerations, have led some to use an absolute measure of agreement or a relative measure using different reference variances (Tinsley and Weiss 2000, p.101; Rindskopf 2002, p.13024). All are motivated by the fact that agreement within a group is not conditional on between-groups differences, which are part of all ICCs.

One measure of reliability that eliminates the problems with "proportion of variance explained" metrics simply converts the raw variance into the standard deviation for interpretability. The standard deviation has the advantage of indexing (dis)agreement in the same units as the original items and does not require comparison to variances across classes of targets (e.g., occupations), i.e., the measure is not benchmarked relative to anything. However, standard deviations for a given trait are usually larger for cases at higher levels of the trait, often because floor effects at the lower end are no longer present. The coefficient of variation (sd/mean) (CV) is sometimes used to adjust for this kind of nuisance variation. Because the CV is normalized by the mean it is unit-free and permits reliability comparisons not only across the range of a single scale but also across variables that differ qualitatively and whose units have very different numerical ranges. A problem with the coefficient of variation is that it is very sensitive to small variations of the mean as the mean approaches zero, producing its own kind of nuisance variation.

An alternative, the r_{wg} index, also measures rater agreement for a single target (e.g., one job or occupation) by avoiding comparisons to between-target variance. The formula involves calculating the within-target variance divided by the variance of some theoretical null distribution and subtracting that quantity from 1 (i.e., $1 - \frac{S^2_{observed}}{\sigma^2_{expected}}$). The null ($\sigma^2_{expected}$) is usually the variance assuming all judges made completely random ratings. Thus, r_{wg} can be interpreted as the proportional reduction in error variance relative to expectation, indexing agreement on the familiar 0.0-1.0 range. Some use a rule of thumb that $r_{wg} \geq 0.70$ indicates high, rather than low, inter-rater agreement, but argue that the standard should be higher if the costs of a wrong decision are high and possibly more relaxed if not (LeBreton and Senter 2008, p.835f., 839).

Variants of r_{wg} test for bimodality to cover the possibility of systematic disagreement, which may be relevant for debates regarding the proper level of occupational aggregation that have arisen with respect to the SOC codes. As with ICC(, *k*) measures, the size of r_{wg} can be attenuated if the number of raters is low (e.g., $k<10$) (LeBreton and Senter 2008, p.826ff.).

An objection to r_{wg} is that assuming ratings are completely random is unrealistic and an overly lenient benchmark against which to evaluate observed error variance. Instead of using only a rectangular distribution as the baseline some suggest using additional benchmark variances for left- or right-skewed or thin-tailed distributions to account for possible response biases toward the extremes and the center of the rating scales, even if ratings were otherwise random.

All of this points to the need to have some standard of comparison against which to benchmark within-group (error) variance, but also continuing dissatisfactions with the current options. There are benchmarks that can make ICCs or r_{wg} appear large on grounds that are extrinsic to the

goal of gauging measurement quality in some circumstances, i.e., inflated, but no decisive technique for resolving the problem with some alternative benchmark.

Most ORS measures are examples of criterion-referenced measures in which the concern is whether individuals or occupations meet some predetermined standard or criterion. In this case, reliability is a question of the rate of classification agreement across raters, methods, and occasions (Rindskopf 2002, p.13026), rather than continuous-variable methods such as ICC. If the ORS mainly collects categorical rather than continuous data, one might be tempted to measure agreement by the proportion of ratings that are identical across raters. Unfortunately, for characteristics that are rare or very common a high percentage of agreement may result from chance. As Harvey noted, inter-rater reliabilities can be inflated if a rated attribute does not apply to a large number of jobs: "the numerous DNA [does not apply] ratings provide a built-in level of 'reliability' that can mask serious disagreements on the items that *are* part of a job" (Harvey 1991, p.112, emph, orig.).

Though not addressing the issue Harvey raised, Cohen's kappa corrects for the possibility that some level of agreement is expected due to chance. Kappa ranges from 0.0 to 1.0 depending on the fraction of possible non-chance agreement that is observed. Kappa can be used with both dichotomous and polytomous nominal variables, and variants exist for multiple raters. A version for ordinal variables weights disagreements differentially according to distance, i.e., gives partial credit for close differences, but some argue weighted kappa is equivalent to an ICC measuring consistency rather than absolute agreement (Tinsley and Weiss 2000, p.114). In addition, when distributions are extreme there is not much possible agreement above chance agreement so even ordinary kappa will not yield large values, making the reliability index sensitive to the trait's prevalence in addition to the quality of the measurement method, analogous to the restriction of range problem with ICC (see, e.g., Uebersax (a) nd). One solution is to present both rates of raw agreement and kappa statistics. Other well-respected but less-used methods for assessing agreement among multiple raters are Kendall's coefficient of concordance (Kendall's W) and Krippendorff's alpha. More sophisticated kinds of categorical data analysis, such as log-linear models for contingency tables and latent class analysis, have also been proposed but are also not widely used (Agresti 2012; Uebersax (b)).

This section has provided a brief outline of a very large literature. There are several methods for measuring reliability that have various strengths and weaknesses, are subject to considerable debate, and whose applicability depends partly on the problem. None provide a mechanical cutoff for determining whether a measure meets some universal standard of acceptability. External expectations that there are standard methods and cutoffs for reliability to which the ORS can be held misconceive the nature of the field. Likewise, the ORS needs to clarify its goals and procedures before any recommendations can be made regarding which methods to consider selecting from among the menu of options.

D. A CLOSER LOOK: THE MEANING AND MEASUREMENT OF VALIDITY

Validity is commonly represented as indicating whether one is measuring what is intended to be measured. Beneath the consensus regarding this general statement is a great variety of interpretations and operationalizations. A common typology of "kinds" or aspects of validity, or

kinds of validity evidence, is the division into face, content, construct, and criterion validity. Construct validity is divided into convergent and divergent validity, and criterion validity distinguished according to whether it is predictive or concurrent. Many of these are recognized in the Federal government's *Uniform Guidelines on Employee Selection Procedures* (http://www.uniformguidelines.com/). Other categories are internal, external, ecological, and consequential validity, some of which do not apply at all to ORS. There have been various efforts to unify all of these concepts as facets of construct validity, conceived broadly, among many other suggestions. This report will focus on areas of core relevance for the ORS, rather than delving into all of these tangled debates (see Sackett, Putka, and McCloy 2012). One point that is a firm part of the current consensus among specialists but not well appreciated by others is that validity is not a property of tests or other measuring instruments, but rather the inferences drawn from them.[23] Again, one must ask not whether scores are valid, rather their level of validity for a particular purpose or use.

This has direct relevance to understanding the concept of face validity, which refers to whether measures appear to be appropriate for their intended purpose on their face. Face validity is often treated lightly because it involves rather subjective judgments usually without clear guidelines or numerical measurement, and the outward appearance of validity is no guarantee that validity measured through quantitative means is high. However, face validity cannot be neglected. It is clearly a problem with many O*NET items, and the poor face validity of O*NET's physical demand items, in particular, motivated SSA to seek alternatives such as the ORS. The ORS items have much greater face validity due to their concreteness, but the program needs to recognize that the method of data collection (i.e., remote interviews vs. personal visits) may affect users' perceptions of ORS' validity. The purpose of the ORS is to support the disability determination process. If stakeholders consider the data lacks validity on their face then it will be difficult for the ORS to achieve its purpose. Ultimately, the success of the program depends on user acceptance, including the legal system, which relates to face validity and argues for methodological studies that can provide some assurance of the validity of ORS' mode of data collection to outside stakeholders.

Content validity refers to the extent to which an instrument represents fully all facets of a construct. For example, as noted previously, there are many suggestions from stakeholders that measures should be added to account for various contingencies. The completeness of the instrument with respect to the disability construct is properly SSA's concern and will not be discussed here. Insofar as the question of the appropriate level of occupational detail arises, this may also be considered a question of content validity and will be addressed in a separate section. Like face validity, assessment of content validity usually, though not always, involves qualitative judgments rather than numerical indices.

Construct validity, which some see as the unifying or most general conception, has traditionally been interpreted in a narrower sense to refer to whether the measures represent a coherent, usually unidimensional, concept. Quantitative measures of internal coherence construct validity are used most commonly with multi-item scales intended to measure a latent construct, a

[23] "Validity always refers to the degree to which that evidence supports the inferences that are made from the scores…. validity depends on the accuracy of score-based inferences and is not a property of a test itself" (Popham 1997).

hypothetical conceptual entity not easily defined in narrow operational terms, such as cognitive ability, self-esteem, or physical fitness. Latent constructs typically cannot be observed directly, so indicators are used to measure or calculate them. Scale coherence can be measured with Cronbach's α index of internal consistency or more formally with factor analysis or principal components analysis to show all scale items load highly on a single latent variable and there is no substantial cross-loading of indicators from other constructs that might suggest absence of unidimensionality or construct contamination. These and other methods of construct validation test for the plausibility of a construct that can be labeled and distinguished from others, and the adequacy of its measurement. The abstract quality of the construct makes the meaning of the indicators and scales a question requiring investigation, rather than something obvious on its face (Sackett, Putka, and McCloy 2012, p.93).

Many O*NET variables, especially in the Abilities and Generalized Work Activities sections, are attempts to use individual items to measure higher-level constructs derived from factor analyses in previous research using much longer item batteries. This research extracted factors from batteries of items and labeled them using some generality that seemed appropriate, and O*NET sought to economize by using these labels for individual items in place of the long list of raw items themselves. Not surprisingly, SSA found these holistic, "factor-level items" far too vague for disability determination and sought much more concrete and less ambiguous indicators that it determined have a definitional (deterministic) relationship to constructs such as sedentary/non-sedentary and skilled/non-skilled. This is most clearly visible in the physical demand items, which identify discrete, observable behaviors that can be measured directly and whose meanings present few interpretive difficulties on their face. Factor analytic methods and similar tests of internal coherence are not appropriate in this context because the higher-level construct is pre-defined in terms of specific facts and behaviors, which are the direct objects of measurement and whose meaning is not at issue. However, other characteristics, such as cognitive demands, are somewhat closer to latent constructs. Those indicators might be analyzed with latent trait models derived from item response theory because their focus on estimating item difficulty is more congruent with SSA's needs than traditional factor analysis.

There are more general and simpler approaches to construct validity that may be quite applicable to assessing the validity of ORS measures. If a measure correlates strongly with others designed to tap the same or similar constructs it exhibits convergent validity, and if it does not correlate as strongly with measures of different constructs it exhibits divergent validity. For example, the sum of SAT verbal and math correlates 0.92 with the ACT composite, indicating very strong convergent validity (Dorans 1999, p.2). Such a strong correlation would be unusual for many other kinds of measures, including measures of job demands. However, one would hope occupation-level ORS physical demand variables would correlate strongly with similar physical demand variables from other databases, such as the NCS, O*NET, and the DOT (convergent validity), and less strongly than with cognitive demand variables from the ORS itself (divergent validity). ORS physical demand measures might also correlate with the incidence of occupational musculoskeletal injuries from occupational health databases.

Whereas reliability coefficients usually measure replicability by gauging similarity between repeated measures that use the same method, validity coefficients generally gauge similarity between measures derived from different methods. Because the units usually differ across

variables, construct and many criterion validity coefficients typically index consistency rather than levels of absolute agreement, though this is not always the case (e.g., comparisons of supervisor and incumbent ratings on the same scales to assess convergent validity).

Another useful method of assessing construct validity is contrasting known groups, which compares trait means or proportions for groups that are known to differ, often quite widely, on the construct of interest to see how well the measure can distinguish them. For the ORS this could involve comparing the proportion of different occupations that are classified into sedentary and non-sedentary categories to determine whether the patterns and magnitudes conform to general expectations and results from other databases, such as the DOT, NCS, and O*NET. To the extent that one measure of a construct distinguishes contrasting groups more sharply than alternatives, it has greater discriminating power, which is an indication of its relative strength.

Taking content and construct validity together, one can say that measures have high validity insofar as they cover all facets of the concept of interest and do not pick up variation reflecting other constructs. If a measure contains a great deal of *construct-irrelevant variance* (e.g., socially desirable responding), it may cause the measure to be biased. Unlike random measurement error, which is the main focus of reliability analyses, bias does not average out generally if individual ratings are aggregated into occupational means. Because it implies consistent, systematic divergence between observed and true values, bias is not captured by most measures of reliability but it will compromise a measure's validity. As can be seen, content and construct validity are concerned with both the *conceptual adequacy* of measures, largely beyond the scope of this report, as well as the effectiveness of a construct's operationalization and the quality of the numerical values that result in practice, which are more consistent with popular understandings of the term "validity" and also this report's concerns.

Evidence regarding criterion validity focuses more directly on common-sense concepts of validity and involves the degree to which a measure is consistent with information obtained from other, independent instruments that are closer to the true object of interest. Common examples of criterion validity coefficients are correlations between college entrance exams and college grade-point averages, and correlations between scores on pre-employment screens (e.g., tests, interviewer ratings) and performance measures of those selected for employment (e.g., supervisory ratings, output records). For example, standardized test scores, such as SAT and ACT correlate between 0.45 and 0.56 with cumulative college grade-point average (Bridgeman, Pollack, Burton 2008, p.5; Schmitt et al. 2009). Like the ORS, criterion-referenced measures are often used to make binary decisions (accept/reject) or classifications (yes/no), as opposed to some evaluation on a continuous scale, raising issues of where to set level thresholds or cutoff scores.

However, raw correlations related to selection processes are often biased downward due to restriction of range. There are no criterion measures for those failing to pass the screening process and those who pass usually have disproportionately high scores on the predictor variable because its purpose is to serve as a selection device. Predictive validity studies can use the distribution of test scores for the entire applicant pool to correct for range restriction in the predictor, though not for the criterion variable. However, if both predictor and criterion values are available only for those selected, e.g., a cognitive test administered to current employees,

then a concurrent validity study must use norms or plausible estimates to correct the restriction of range for both variables. Some studies also correct for downward bias in validity correlations due to random measurement error (unreliability) in the criterion variable. Restriction of range is not directly relevant for the ORS, which is concerned with the population of jobs, not persons, and will measure the full range of traits using representative sampling procedures. However, it should be noted that validation studies of job analyses may use estimates of population variances and criterion reliability that increase the sizes of reported coefficients significantly. A recent meta-analysis found graduate admissions tests correlated about 0.40 with graduate school GPA *after* correcting for restriction of range and criterion unreliability (Kuncel and Hezlett 2007). The raw correlations between measured cognitive ability and job performance are usually in the 0.20s, but are usually 0.40 or above after corrections (Schmitt 2014, p.46). The differences can be substantial. If assumptions in any particular study are mistaken they may produce upward bias, turning underestimates into overestimates. Without entering debates on these practices, it is clear that the utility of such criterion validity coefficients for ORS planning or benchmarking is limited given the very different methodologies.

A form of criterion validity assessment that is applicable directly to the ORS and is closer to popular conceptions of validity involves comparing measures that are preferred for their cost and convenience to costlier methods that are considered superior on substantive grounds, which are the criteria against which the former are evaluated. For example, the public's preference for ratings derived from job observations by field analysts rather than interviews is based on the belief that the former is likely to be closer to the truth for various reasons, such as analysts' training and the procedure's greater objectivity. In this case, ratings from job observations are the criteria against which more scalable methods are benchmarked.

It should be noted that specialists in measurement would be quick to point out that while there may be good reasons to believe ratings derived from analyst observations are *closer* to true values, there are also good reasons for not assuming they are equal to the true values. For example, any job observation represents a sample from a broader population of all incumbents and occasions one could witness, not a full-year, continuous-time recording of all incumbents' every action. Different analysts will produce different ratings of the same situation, reflecting possible rater predispositions (fixed effects) as well as random variation. SVP and mental job requirements are much less observable than physical demands so primary reliance on verbal reports is unavoidable for measuring these kinds of traits and little is likely gained from behavioral observations. Finally, skill ratings in the third edition of DOT have been shown to grossly under-estimate job requirements for a number of female-dominated occupations both absolutely and relative to certain male-dominated occupations (Cain and Treiman 1981, pp.269ff.). Although the problem was corrected in the fourth edition that appeared in 1977 the example shows that even professional observers can be biased by halo effects and social preconceptions, supporting the premise of test score theory that true scores are essentially unobservable.

Medical research provides one of the few contrasting perspectives that uses a concept of accuracy, defined as the extent to which results agree with a known standard, usually an instrument measuring something physical or chemical. The criterion is a more objective measure of the specific focus of interest, rather than the unobserved constructs and indirect measures that

are common in the psychological literature. If a measure can be assumed to be nearly error-free then the validity or bias of the simpler or less costly alternatives can be measured as the mean absolute deviation from that gold standard and the standard deviation of the deviations in the case of continuous variables. Small values for both indicate strong agreement. Bland-Altman plots graph the difference between two methods on the y-axis against their mean values on the x-axis, drawing horizontal lines at zero, the mean difference, and the boundaries of the "95% limits of agreement" (±1.96 times the standard deviation of the difference). A non-standard hypothesis test may be possible for the mean difference but traditional t-tests are problematic because they set a high hurdle for rejecting the null that an observed value differs from zero, which becomes a very lenient standard in situations where the desired outcome is the absence of a significant difference from zero, i.e., a failure to reject the null is the favored hypothesis. In practice the magnitude of the mean absolute difference between a measure and the gold standard may be the best indication of bias. The limits of agreement indicate the expected differences between the two methods for approximately 95% of cases assuming the differences are mostly random and therefore normally distributed. If the bias is small and the expected range of random differences (limits of agreement) is narrow, then the two methods are essentially equivalent and the methods can be used interchangeably. "How small the limits of agreement should be for us to conclude that the methods agree sufficiently is a clinical, not statistical decision. This decision should be made in advance of the analysis" (Bland and Altman 2003). The graphs can also be used to identify outliers, group differences, and non-constant bias and variability across the range of trait values, which render the estimated limits of agreement potentially invalid and potentially require a log transformation or use of percentage difference instead of raw difference. Bland-Altman plots can be used to compare two or more methods whether or not one is treated as a gold standard. Conceivably, they could be used in studies of the ORS to compare job-level ratings with occupational means to address questions regarding the magnitude and significance of within-occupation variation. There is a large literature on different methods based on Bland-Altman, and arguments for other approaches that should be consulted if they are judged relevant to ORS validation studies.[24] Unlike most other validity indices, these methods involve comparisons involving common units, so they measure absolute agreement for continuous variables rather than consistency. As Bland and Altman (2003) note, correlations, including validity coefficients, are "blind to the possibility of bias" and other forms of systematic disagreement when the weaker consistency standard is met.

For a binary classification (e.g., sedentary/non-sedentary), common indices include sensitivity, the percentage of positives that the measure identifies as positive, and specificity, the percentage of negatives that the measure identifies as negative (i.e., does not misidentify as positive). Highly sensitive measures produce few false negatives and highly specific measures produce few false positives. Related indices calculate conditional probabilities from the other perspective, showing the percentage of cases identified as positive by the measure that is actually positive (*positive predictive value*), and the percentage of cases identified as negative by the measure that is actually negative (*negative predictive value*). An overall index of the accuracy of a binary classification method is the likelihood ratio, which is the *percentage of true positives testing*

[24] For an example of some of the issues with Bland-Altman plots, particularly in the presence of gold standards, see, Will G. Hopkins, "Bias in Bland-Altman but not Regression Validity Analyses." Sportscience. 8:42-46 (2004) and Alan M Batterham, "Comment on *Bias in Bland-Altman but not Regression Validity Analyses*." Sportscience. 8:47-49 (2004), both available at http://www.sportsci.org/.

positive divided by the *percentage of true negatives testing positive* or $\frac{(sensitivity)}{(1-specificity)}$. The likelihood ratio can be interpreted as indicating how much more likely a positive classification will be made when the case is actually positive than when the case is actually negative.

Sometimes a continuous measure is used to make binary classifications (e.g., level of an antibody→medical diagnosis, college entrance exam score→admissions decision). Varying the stringency or leniency of the cut-off used for a positive classification will involve a tradeoff between sensitivity and specificity (e.g., higher thresholds that reduce the number of false positives usually increase the number of false negatives). The tradeoffs involved in using a more or less stringent cutoff for a continuous measure can be graphed in a receiver operating characteristic (ROC) curve to find the optimal value.

Although the case of medical tests seems supportive of the popular notion of accuracy and true values, Bland and Altman argue that even "gold standards" for scientific measurement contain measurement error in practice:

> *Comparing methods of measurement should be a matter of estimating how closely two methods agree, not whether they agree or not. Most biologic measurements are made with error, and exact agreement will not happen, even when using the same method twice. We can decide how much disagreement might be acceptable before the study and that might well vary for different purposes* (Bland and Altman 2012).

The final sentence raises the issue of the need to interpret the size of errors in terms of their practical importance discussed earlier. Tinsley and Weiss (2000, pp.114f.) expand on this point in warning against the assumption that error magnitudes have context-free meaning:

> *Some very small errors can be quite serious (e.g., a mistake of one gram in preparing a prescription or a mistake of one degree in computing the trajectory of the space shuttle), whereas in other instances quite large differences are of no appreciable consequence (e.g., a mistake of one hundred pounds when measuring 100 metric tons of grain or a mistake of one mile when estimating the distance between New York and Los Angeles). This essential truth is often ignored in the sciences and humanities, where a premium is placed on achieving the greatest precision possible….failure to distinguish between important and inconsequential errors can be limiting and even self-defeating. It is not possible to assess interrater agreement adequately without considering the seriousness or cost of the various types of disagreement that can occur.*

Ultimately, judging the acceptability of reliability and validity indices will have to be based on the percentage of jobs potentially classified inconsistently or incorrectly into the categories that SSA intends to use in its ratings of job demands. Even medical tests often leave significant room for interpretation. This implies, for example, that neither incumbents nor supervisors and managers are sources of "true" information against which the other can be evaluated for accuracy.

Nevertheless, there are some cognitive and motivational issues of the survey situation that have predictable effects on the quality of measures. Common cognitive problems potentially relevant for ORS include:

- difficulty understanding survey questions
- recall error, particularly in the absence of specific prompts
- telescoping the timing of events or otherwise misestimating numerical quantities
- narrow frame of reference, particularly when questions use vague descriptors

In general, sources farther removed from the actual job are less likely to be familiar with intimate details (e.g., HR officials, middle and upper managers, formal job descriptions). There is commonly slippage between formal conceptions regarding how the work is performed and the way is actually performed.

Common motivational issues potentially relevant for ORS include desire to represent oneself in an overly positive light (self-enhancing bias) for ego satisfaction, material benefit, or self-protection. Thus, it would not be surprising to find workers over-estimate their autonomy. Ideally, survey questions are neutral and they do not motivate respondents to answer in specific ways because they potentially cast them or their employer in either a positive or negative light. The ORS will have to give careful consideration to how employer responses to questions regarding possible physical and environmental hazards in the workplace may reflect concerns such as potential liability in addition to any cognitive limitations such as limited personal knowledge of workplace practices and conditions. Persons in responsible positions within a firm have motives for downplaying aspects of jobs that involve hardship, risk, and danger.

III. LITERATURE REVIEW

The ORS seeks a review of the literature to ensure its items and method of data collection produce measures with high reliability, validity, and accuracy. It is relatively simple to specify the kind of information that is most useful. The ORS would be helped most by previous research using items on physical requirements, specific vocational preparation, cognitive job requirements, and workplace environmental conditions that are highly similar to its own items and administered under the various conditions that are being considered for the ORS itself. Such research would provide the clearest indication of best practices applicable to the ORS because the conditions closely approximate those of the ORS itself. Insofar as the content of survey items and other study conditions differ from those contemplated for the ORS they introduce additional sources of variance and uncertainty when trying to draw conclusions for the ORS regarding the merits of different sources of data and modes of data collection.

The first logical place to look for guidance are the large-scale government data collection programs most similar to the ORS in scope, the DOT, O*NET, and NCS, as well as trial phases of the ORS itself. There are other research streams that also can contribute important insights and perspective, such as IO psychology, but important differences from the ORS that limit their applicability need to be recognized. For example, it is much easier to control occupational coding unreliability as a source of measurement error for the kind of small-scale case studies of one or a few workplaces that are common in IO psychology. One can be more confident that recorded differences in job demands between big-city police departments and small-town departments represent genuine heterogeneity within occupations rather than measurement error compared to more impersonal data collection methods. Job requirements can be measured at even more detailed levels of occupational coding than those found in the DOT.

Likewise, personal visits may be nearly costless if the study sites are local employers, including the researcher's own university, and sample sizes are between 35 and 200 respondents. Smaller sample sizes often permit relatively long interview times, more comprehensive item batteries, and more labor intensive collection procedures. Employers may cooperate more readily out of institutional loyalty in the case of the university and due to the direct relevance of the study to their own specific jobs in the case of private employers. It is possible that some published studies are byproducts of consulting for which employers have contracted the researchers' services. Instruments that require hours to complete may be feasible in such situations because the employer also seeks to use the results to deal with some practical organizational matter of intrinsic concern, and the employees are permitted to complete the surveys during paid working hours. Such methods may generate very reliable scales based on detailed measures but are unlikely to "scale up" directly to a project like the ORS. Likewise, university personnel have a greater understanding of and appreciation for research so their homogeneity and cooperation levels with even normal-length instruments cannot be assumed to generalize to the ORS.

In addition, existing ORS practices permit semi-structured interviews balanced by intensive interviewer training to ensure common standards, use of detailed definitions of terms and instructions for field coding responses, probing of respondents to ensure consistency, and post-collection reconciliation and quality checks. The vast majority of existing studies use simple surveys completed by the respondent alone or in the course of a relatively standardized interview.

Given the goals of such studies it is entirely possible that their validity remains intact even if the workplaces selected constitute exactly the kind of haphazard convenience sample criticized by the NAS review of the DOT in 1980 and reiterated by APDOT in 1992. Likewise, because the occupations studied are often very specific and selected on an apparently opportunistic basis, it is difficult to judge the population that they may represent once they are pooled, e.g., all occupations, manual occupations, some composite of manual and non-manual occupations, etc. A population or whole-workforce perspective tends to be absent in IO psychology studies because their studies and overall professional orientation tend to be on the micro rather than macro level.

However, the goals of the ORS prevent it from sampling establishments or occupations on a purely convenience basis and it would be a mistake also to assume that all of the positive qualities of the smaller studies could be replicated easily in a reasonable time frame at reasonable cost for a national probability sample. In many, though not all respects a large-scale study is simply a different kind of project than a single-site or few-sites study, and it is easy to underestimate the difficulties of scaling up practices that are quite feasible in small-scale research to large, nation-wide random samples. High-volume data collection procedures are typically more impersonal and the projects have weaker and more impersonal ties to respondents. Insofar as the practices of these studies differ from what is possible in the ORS, their applicability to the ORS is qualified or limited.

Indeed, constructs and operationalizations often differ sufficiently *across* small-scale studies that one must be cautious in trying to aggregate results even aside from issues of their applicability to

larger projects; it is not easy to control for all sources of variance when there are many differences in theoretical focus, choice of measures, methods of data collection, number of scale items, and number of raters. The results of these studies are difficult to cumulate due to significant differences in design and reporting. When studies differ in so many ways from the ORS, it is difficult to assess the relevance of the reliability and validity of their measures for ORS procedures.

This section reviews available information on reliability and validity for the DOT, O*NET, NCS, and early phases of the ORS. This will be followed by a discussion of academic studies in IO psychology and other disciplines, such as occupational health. In general, this review will reinforce the conclusion that there is no substitute for the ORS conducting its own experiments and validation exercises to determine the best items, sources, and modes of data collection for its own distinctive requirements. Existing research needs to be judged in relationship to how closely it corresponds to likely ORS procedures.

A. DOT

SSA has relied on the DOT in its disability determination process for many decades, gaining legal acceptance and support from various stakeholders, many favoring creation of an updated DOT or comparable database. Nostalgia for the DOT is palpable, though not universal. The virtues of the DOT that are often cited are the fine detail of its occupational coding and its reliance on on-site job observations conducted by trained raters, both of which are believed to be sources of greater validity relative to O*NET and other program designs. The DOT is the main point of reference in many discussions of ORS, and the explicit or implicit benchmark in SSA's long process searching for a replacement. Because the DOT is treated as a gold standard, it is important to understand it more deeply. In fact, the DOT was the focus of an extensive study performed for the National Academy of Science (NAS) in the early 1980s whose results remain relevant.[25]

1. Background

The first edition of the DOT appeared in 1939 during the Depression as a tool to help the newly formed Employment Service match unemployed job-seekers to job vacancies. The second through fourth editions, some of which represented substantial changes in conception as well as simple updates to the ratings, appeared in 1949, 1965, and 1977. A partial update ("revised fourth edition") appeared in 1991. Throughout its history, the intended purpose of the DOT was to help match people to jobs, first in the context of mass unemployment and later as a guide for career planning and exploration for young people and other new labor market entrants. The DOT was designed initially for job placement among the unemployed, vocational education, and career counseling. Beginning with the third edition, SSA contracted with DoL to publish a supplement, Selected Characteristics of Occupations, containing the data elements that would be useful for disability determination and the agency became a large-scale user of the DOT, which led to significant concern as the revised fourth edition became increasingly outdated. However, the fourth edition of the DOT was not necessarily a sustainable model of data collection.

[25] I thank Pamela Stone, a primary author of the NAS study (as Pamela S. Cain), for discussions that were helpful in writing this section.

The DOT was a Federal-state collaboration. DoL's Division of Occupational Analysis in Washington, DC served as the DOT's national office, but all data collection and most other primary operational work occurred in at least ten field offices that belonged to state employment service agencies. The NAS panel reported there were 129 professional and support positions in the field centers and 15 in the national office, spending the vast majority of their time on the DOT (Miller et al. 1980, p.93). Although the national office was in charge of the DOT, the state agencies set hiring standards, and pay and promotion policies for the field job analysts. Many analysts were hired from state Employment Service offices or related agencies in which they had gained experience with the DOT as interviewers. Many analysts had bachelor's degrees in one of the social sciences, but there is no indication that they had degrees in IO psychology programs or were hired on the basis of any prior experience or training in job analysis specifically. Supervisors looked for general attention to detail and interpersonal skills that would be useful in gaining access to work sites, rather than formal training or credentials in job analysis. On the job training took 1-2 years, including informal mentoring (Miller et al. 1980, p.104). Unfortunately, the dual authority structure of the DOT operation contributed to chronic management issues, which were exacerbated by high turnover and poor leadership among top officials in the national office. The national office had five directors or acting directors between 1975 and 1978. It is quite possible that this reflected difficulties in meeting the DOT's work schedule. The number of staff in the national office had declined from 33 in 1966 to 10.5 FTE in 1978, of whom only 7.5 were professional level and responsible for coordinating the work of the 11 field centers (Miller at al. 1980, pp.94f.). Field center personnel "uniformly expressed a negative perception of the leadership of the national office" and "continually conveyed to us their feeling that the national office staff lacked a sense of direction and failed to maintain its leadership role adequately in coordinating the work of the field centers" (Miller et al. 1980, p.101, see also pp.112f.).

The DOT's mission was intrinsically difficult. In eleven years (1965-1976) the field centers produced more than 75,000 job analysis schedules for 12,099 occupational entries (Miller et al 1980, p.93). The number of individual job analyses was enormous, but because of the level of occupational detail it translates into only somewhat more than 6 reports per occupation on average. In fact, the coverage was much more uneven and haphazard than this average suggests. As is well-known, the DOT did not follow a probability sampling strategy and the representativeness of the cases selected for observation is unknown and probably unknowable. What is less well-known is that 16% of occupation entries were not based on a single job analysis, simply carrying over information from the DOT 3rd edition. As indicated in the frequency distribution below, another 29% of DOT entries was based on a single job analysis and 19% was based on two analyses. Even when there were multiple analyses per job, each report was usually based on analyst observations of only one or two workers, as well as interviews with the workers and others, including verification of details with immediate supervisors (Miller et al. 1980, p.125).

Percentage of Occupations by Number of Job Analysis Schedules, DOT 4th ed.

# JA schedules	Occupations
0	16%
1	29%

2	19%
3-7	24%
8+	13%

Source: Miller et al., 1980, p.158 (cf., Cain and Treiman 1981, pp.259f.)

The use of multiple sources of evidence and collection methods is a strength of the method, but given the commonly low ICCs for single ratings it is quite likely that there is significant random error in published DOT ratings that rest on one or two job analyses of one or two workers each. "Insofar as DOT occupations are internally heterogeneous, the heavy reliance upon descriptions of only one or two specific jobs would appear to lead to unreliable data," though it is not possible to assess the actual level of heterogeneity for occupations with so few ratings (Cain and Treiman 1981, pp.260; Miller et al. 1980, pp.157, 159). Even if one were to test for heterogeneity in SOC-type codes by aggregating DOT scores to the 3-digit Census occupational level using the April 1971 CPS supplement for which such a study is feasible, there is no guarantee that a high r_{wg} within Census occupations could be interpreted as evidence of a problem with heterogeneity given the uncertain reliability of the scores at the DOT-occupation level (discussed further below).

This spotty coverage of different occupations was part of a broader issue. In 1974, the national office, perhaps concerned about meeting the publication deadline, instructed analysts to abbreviate procedures "in order to speed completion of the fourth edition" (Miller et al. 1980, p.161). The field offices interpreted the directive in various ways and took different shortcuts that led to omission of various data elements and left certain kinds of jobs unobserved; 30% of the job analysis schedules were produced between 1974 and 1976 alone (Miller et al. 1980, pp.140ff). The quality and completeness of job analysis schedules from the field centers varied more widely after this change (Miller et al. 1980, pp.9, 98). Undoubtedly matters were not helped by the fact that the field guide for analysts, the latest edition of the Handbook for Analyzing Jobs, was not published until well into the process (1972) (Miller et al. 1980, p.145).

However, another point that bears mentioning is that the third edition of the DOT was not based directly on any actual fieldwork or job observations. A small study showed the average ratings of job descriptions by eight experienced analysts at the national office had good consistency with the average ratings of eight analysts observing similar jobs on site. The results were

> taken as evidence that ratings could be assigned using job descriptions only. Thus for the third edition DOT, ratings were assigned primarily by national headquarters personnel using job descriptions only, with some assistance from the field center staff. The fourth edition saw a change in the procedures used to rate jobs and occupations for the DOT…field center analysts not only collected job data and wrote descriptions but also rated each job…In addition, field analysts were responsible for assigning ratings to the occupational composites contained in the DOT, formerly a task of the national office (Miller et al. 1980, p.169).

In other words, the fourth edition of the DOT was unprecedented in its ambitions not only in terms of the number of occupations and data elements it covered, but also in terms of the labor intensity of its primary data collection method. The fourth edition's task was to complete a field study in approximately the same amount of time that had been allocated to the previous edition

to produce an office study. While the available evidence does not permit a firm conclusion, and poor planning and organizational structure were likely sources of problems, one suspects that a number of the management and substantive issues associated with the DOT resulted from its ambitious scope of work, and the resulting difficulty of meeting its schedule. It is not surprising that the DOT revision in 1991 was a limited update, and that it was followed by the establishment of an advisory panel to recommend alternatives.

By the time the Advisory Panel for the Dictionary of Occupational Titles (APDOT) issued its final report, the number of occupational analysts had shrunk to 30 spread across just five field offices (APDOT 1992, p.25). The APDOT panel concluded that the DOT's observation and interview methods produced accurate results "but at great expense of time and cost. Quite simply, it is a cost-prohibitive method for any realistic effort to describe all occupations in the American economy" (APDOT 1992, p.20). Likewise, it is not surprising that the interim report stated that the panel "believes that the resources needed to keep some 12,000 occupations current, valid, and reliable are likely to remain beyond the funding capability of any government agency. No other country even attempts this level of detail."[26] In an APDOT technical paper, a senior professional from the UN's International Labor Office noted that other countries attempting some similar kind of national occupational classification and dictionary (NOCD) found the costs mounted rapidly and unexpectedly, and were compelled by practical considerations to curtail their plans:

> In reviewing the work on different NOCDs one can easily get the impression that most of the organisations which have been able to develop a (reasonably) finished product, tend to collapse from the effort after crossing the finishing line, at least as far the NOCD is concerned. The general picture is that no systematic effort and hardly any resources are spent on the updating and maintenance of the NOCDs for a good many years after they have been published.[27]

In general, stakeholder comments regarding the appropriate replacement for the DOT seem to be unaware of its singular nature, as well as the corresponding difficulties it encountered. The high cost of replicating the DOT almost certainly explains why it was not updated and why the search for a substitute has been so long and frustrating. If there were an inexpensive way of updating the DOT it would have been done. The problem is that achieving even the level of quality and detail found in the DOT 4th ed. faces significant resource constraints.

Indeed, SSA's OIDAP panel conducted a review of international practice and found nothing comparable to the DOT. The study involved an extensive literature search and interviews with officials responsible for disability determination and occupational information systems in the United Kingdom, Canada, Australia, New Zealand, the Netherlands, as well as the ILO. None of those countries has a comparable database, and only the Dutch used any occupational information system for disability determination (Social Security Administration 2011, pp.71ff., 81). In interviews with foreign counterparts, SSA found, "for the couple of people who even acknowledged that they wanted to do something, there was just a practical limitation in terms of

[26] Accessed from the Federal Register, 57 FR 10588, March 26, 1992.

[27] Eivind Hoffmann, "Mapping the World of Work: An International Review of the Work with Occupational Classifications and Dictionaries," International Labour Office (1998) p.10. Revised version of technical report prepared for APDOT (APDOT 1992, p.47).

resources and the complexity of the task…and that seemed to limit them…" (Social Security Administration 2011, pp.88ff.). An extensive study of other OECD countries also found no database comparable to the DOT (Handel 2012). One of the closest was the Canadian labor ministry's Essential Skills database, which involved interviews with only 3,000 workers, covering 70% of SOC-type occupations and 80% of the Canadian workforce, before being discontinued (Handel 2012, p.88). In fact, the scope and scale of the DOT led some other countries to use its concepts and scores, as well as those of O*NET, for rating their own occupations (Miller et al. 1980, pp.85f.; Hoffmann 1998; Handel 2012; Social Security Administration 2011, p.85).

None of this should be taken to imply that the ORS can be bypassed on the basis of practices in other countries. Along with three other countries, the United States has "the most stringent eligibility criteria for a full disability benefit" among all OECD countries, "including the most rigid reference to all jobs available in the labor market."[28] Insofar as other countries have do not have the same policy need for an occupational information system because their policies are less stringent, their decisions with respect to the use of such databases do not generalize to the United States. The DOT is well-known among disability advocates in the U.S. as a tool for denying claims. If the bar for granting claims is to be set at a higher than average level, maintaining a comparable level of fairness in determining eligibility requires a correspondingly higher standard of evidence.

Although the DOT's level of occupational detail and the predominance of site visit methods cannot be replicated feasibly, this does not prevent some elaboration of recognized occupations and use of job observations for priority cases. **As a general recommendation of this report, it can be said that the ORS should try to accommodate stakeholders' legitimate concerns regarding the quality of information that will impact their lives to the extent feasible.** The fact that a prior model is not feasible *in toto* should not be taken to mean that its methods cannot be adopted in part. With thoughtful analysis and planning it may be possible to target priority occupations and industries for more detailed occupational definition and intensive study. Previous sections have noted the notable clustering of SSA's applicants in certain occupations and exertional and SVP levels, as well as the likely absence of a need to observe job activities of most office occupations, for example. Certainly, experiments and validation studies can be conducted to understand the implications of relying on standard as opposed to intensive methods for the final database.

Because the DOT remains the standard against which the ORS will be judged, it also makes sense to review the nature of its occupational classification system because occupational aggregation is an aspect of database validity, and evidence for the reliability and validity of DOT scores themselves, as their grounding in job observations, however imperfect, are considered the database's other core strength.

2. Occupation in the DOT

[28] Organisation for Economic Co-operation and Development (2010),"Sickness, Disability, and Work: Breaking the Barriers: A Synthesis of Findings across OECD Countries," OECD: Paris, p.89.

The validity of any system of job rating depends on the validity with which repeated measures are aggregated or reconciled at a higher level of categorization. The DOT appears to be singular in its level of occupational detail. Separate entries with descriptions and ratings for jobs such as "waiter/waitress, head," "waiter/waitress, captain," "waiter/waitress, formal," "waiter/waitress, informal," and over fifteen others varieties of waiter/waitress give some sense of the granularity of the database (Cain and Treiman 1981, p.254). Some indication of the difficulty of measuring very narrow occupations is reflected in the fact that the April 1971 CPS file with over 53,400 respondents coded into DOT occupations used only 4,517 DOT titles or about 37% of the total (Spenner 1980, p.245).

However, there does not appear to have been a clear set of principles governing the definition and selection of occupations to be rated or the shape of the overall system of occupational classification. The third edition appears to have been used as a sampling frame for occupations despite carrying forward a significant skew toward manufacturing jobs from previous editions. Although manufacturing accounted for only 21% of employment at the midpoint of data collection for the fourth edition (1971), 67% of all DOT 4th ed. entries describe jobs in manufacturing. By contrast, 12% of entries covered clerical, sales, and service jobs, despite accounting for 41% of employment in 1971 (Cain and Treiman 1981, p.257). There is no evidence that this lopsided degree of occupational detail reflects actual differences in job heterogeneity across these groups that might warrant such different treatment.

It is possible that an argument can be made that the DOT'S blue-collar, manufacturing bias provides disproportionate detail for occupations that are among the most heavily represented among SSDI applicants. However, empirical results presented earlier call this into question. Public discussion often seems to suggest that the DOT's impressive level of occupational detail is spread evenly across the occupational spectrum, but this is clearly not the case. If the level of detail for clerical, sales, and service were applied to the entire range of occupations, there would be 3,540 entries, not 12,100.[29] While the smaller figure represents finer occupational detail than found in the SOC or O*NET, it is a much lower standard than most people assume the DOT used based on its disproportionately elaborated classification of manufacturing jobs.

The grouping of job analysis schedules into occupations was also more art than science. With 75,000 job analysis schedules and far fewer occupations, deciding how to cluster jobs into occupations before determining the latter's scores was not straightforward because few schedules were likely to be identical. As the NAS report notes,

> …both "job" and "occupation" are theoretical entities. The central question in creating these entities is how to delineate the boundaries, by deciding how much heterogeneity should be tolerated within them.. Although combined jobs did not have to be identical, they were supposed to be similar (Miller et al., p.197, 143).

As part of their jobs, certain field analysts combined jobs into occupations according to prescriptions in a guide written by headquarters. The analysts grouped jobs initially based on the first three digits of the DOT code assigned by the analysts who performed the ratings, and then judgment was used to make further adjustments. The manual advised strongly against grouping

[29] This figure is derived from the following calculation using figures from the previous paragraph: [(0.12*12,100)/0.41].

jobs together that differed by more than two points on the Data-People-Things scales and more than one level on GED (Miller et al. 1980, p.144). Any remaining differences between job schedules were resolved using personal judgment to arrive at a composite description before a final review, at least for the 55% of occupations with more than one schedule (Miller et al. 1980, pp.116,143f., 159). This process was complicated by the varying number and quality of analysts' job analyses (Miller et al. 1980, p.143).

In summarizing the DOT's treatment of occupation, members of the panel wrote:

> This process is inherently arbitrary, depending heavily on analysts' judgments as to what constitute variants of the same occupation and what constitute different occupations. There are no procedures for ensuring comparability throughout the DOT in the level of aggregation of jobs into occupations…DOT occupations vary widely in their detail and specificity (Cain and Treiman 1981, p.260).
>
> …there exist no principles for determining the boundaries of occupations and hence no unambiguous procedures for aggregating jobs into occupations...What constitutes an "occupation"—and how much heterogeneity in the content of a set of jobs justifies a single occupational title—is a difficult question... Occupational titles are also used inconsistently in the DOT to define very specific or very heterogeneous groups of jobs[30] (Miller et al. 1980, pp.191, 194).

Although the NAS report called for the establishment of principles for defining occupations, it is not clear that this is possible in an objective sense. O*NET apparently has procedures for identifying new and emerging occupations, albeit relatively unpublicized.[31] However, there will always be judgment involved. As Harvey's review of job analysis methods explained in some detail, concepts like occupation are abstractions built up from information on specific positions.

> All higher-level organizational constructs—(jobs, job families, occupations)—necessarily involve aggregating across, and thereby ignoring and deeming unimportant, a potentially sizable number of both large and small behavioral differences that exist between positions.. It is easy to fall into the trap of treating jobs and job families as if they are monolithic "real" entities, and thereby lose sight of the fact that there is invariably a nontrivial degree of within-title or within-family variability in job behavior. In some

[30] Interestingly, the job "branch manager" was mentioned in the NAS report as an example of diversity (Miller et al. 1980, p.194) and this seemed echoed by a recent O*NET example of the need to account for diversity among "project managers" whose tasks differ dramatically between industries, e.g., job tasks of information technology project managers differ greatly from those of construction project managers. Thus, the title "project manager" without further specification may be insufficiently informative depending on the purpose (Social Security Administration 2011, pp.246f.). The key question for ORS is whether a level or method of ratings aggregation is acceptable or not for its purposes.

[31] A trade association or other source must indicate at least 5,000 employees and preliminary evidence must indicate "significantly different work from the existing SOC,…positive projected growth, training, certification program," and similar indication of existence as a distinct entity (Social Security Administration 2011, p.243). O*NET has identified 154 new and emerging occupations as of 2011. Associations, professional organizations, and users also provide feedback and input that contributes to identification of separate occupations (Social Security Administration 2011, p.244f.).

> cases, this within-classification variability may be unimportant; in others, it may be highly significant (Harvey 1991, p.80; cf. Buckley 2012).

Indeed, as the preceding implies, the issue goes beyond the need for judgment. While the desire for maximal occupational detail is understandable, the level of possible detail is effectively infinite and no source, including the DOT, can claim to have objective criteria or a clearly validated method for deciding the level of detail at which to stop. The Federal government recognizes eight levels of chemists (Buckley 2012). The *Alphabetical Indexes of Industries and Occupations* list over 31,000 occupations and 21,000 industry titles.[32] Perhaps BLS can be clearer on its procedures for defining distinct SOC occupations. However, it appears that there is no obvious objective or ironclad method for deciding which jobs are sufficiently similar on which specific dimensions to qualify for aggregation into a single occupation, while others require separate recognition as different occupations. The spirit of many outside comments suggests a common belief that "true" occupations can be identified easily. However, it is a fallacy to believe occupations are "real" entities, rather than more or less useful classificatory tools. There are no obvious standards for deciding how detailed a system of occupational coding must be (i.e., how many occupations it must recognize). At the limit, the level of detail in occupational coding is potentially infinite. No database can meet this standard, though ORS should aim for good coverage and sensible distinctions. **It would be worthwhile for ORS to study existing and best practice of recognizing distinct occupations, modify any procedures accordingly, and include the information in ORS public materials.**

One feasible method for avoiding this problem, perhaps the only method, is to include alternative formats or release microdata as part of the ORS database design. If the claim is that there is irreducible heterogeneity, i.e., each job differs in some respects from another such that any aggregation or occupational averaging is suspect, then the logical conclusion is to design ORS so that the same underlying data is viewable not only in terms of occupational categories but also in terms of the relevant trait categories (e.g., exertional demands, SVP, specific environmental conditions, and relevant combinations). If there is a fear that occupational aggregation washes out heterogeneity that is meaningful for the disability determination process, then the underlying data should be presented or available in alternative views. In principle, there is no reason why ORS could not make available in static or manipulable form information on the percentage of *jobs* that are sedentary and both unskilled and sedentary, rather than the percentage of jobs that are in *occupations* deemed sedentary or unskilled-sedentary after aggregation to the SOC level. In other words, statistics on the existence of jobs aggregated to the *trait* level could be presented, in addition to occupation-level means, percentiles, and category frequencies. If disability determination requires that the *names* of jobs be available to be cited in decisions as illustrative of other possible work along with trait prevalence, then *job* titles from original ORS interview schedules, anonymized and cleaned up to conform to the *Alphabetical Index* listings, could be included, as well. This database would be both an occupation- and job- or trait-level profile of available work in the U.S. economy. Indeed, for the many titles with few analyst site visits, the DOT is itself effectively a job-level database, not an occupation-level database.

[32] See https://www.census.gov/people/io/methodology/indexes.html and https://www.census.gov/people/io/files/overview2010.pdf (accessed 2/16/15).

3. Reliability and validity of DOT ratings

The reliability and validity of ratings depends not only on the quality of the raw ratings themselves, but also on how the ratings are processed afterwards and how cases are sampled, discussed in prior sections. The one point that remains prior to considering the raw ratings themselves is an aspect of sampling with implications for the nature of the job selected for rating. Many field analysts, as well as studies in the sociology of occupations and organizations, consider the total number of employees in the establishment, as well as industry, relevant for the structuring of outwardly similar-sounding jobs. Small establishments tend to require more versatility, while a larger workforce permits a finer division of labor and more task specialization (Miller et al. 1980, p.119).[33] Like the DOT, ORS might want to consider incorporating establishment size into its sampling design if it has not done so already, ensuring that the NCS sampling frame has good coverage of small establishments, in order to capture the full range of job diversity.

Evaluating the DOT's job ratings themselves is complicated by the fact that the NAS panel could uncover no reports or evidence on the validity or reliability of the ratings even though the method of data collection changed significantly between the third and fourth editions (Cain and Treiman 1981, p.261; Miller et al. 1980, p.169). Again, given the different policy context, the ORS is well-advised to avoid following the DOT's example in this respect.

As the panel noted, a further complication is the "absence of an external criterion of job complexity against which to assess the DOT ratings" (Miller et al. 1980, p.190)

> …most of the data contained in the DOT are unique, so no readily available bench marks exist against which to compare and assess them. In fact, a great deal of occupational research takes the DOT as the bench mark or standard of comparison (Miller et al. 1980, p.149)

In other words, there is no gold standard against which to evaluate the DOT because, given its singular nature, the DOT was the effective gold standard for job measurement.

Nevertheless, interviews with field analysts yielded some important information on specific variables relevant to the ORS. Job analysts did not use decibel meters or other objective measuring devices to measure physical demands and environmental conditions; indeed, employers often prohibited their use (Miller et al. 1980, pp.120, 137ff.). "Many of these factors could have been measured objectively," but instead the "analyst queried workers closely about the processes, machines, and materials they worked with in order to determine environmental conditions. To assess physical demands, job tasks were usually merely observed" (Miller et al. 1980, pp. 137, 139). The employer also controlled the work areas analysts were permitted to observe. **Again, this report recommends alternative methods of collecting information not dependent on employer participation for data elements that are potentially sensitive for employers, such as worker surveys, administrative records (EPA, OSHA, etc.), and occupational health and similar subject matter experts.**

[33] A presentation to OIDAP noted that small business owners and supervisors may also perform tasks of front-line employees, requiring use of multiple DOT codes to fully account for the requirements of these composite jobs (Social Security Administration 2011, p.279).

NAS interviews with field analysts also covered collection of SVP ratings:

> The SVP was reported to be difficult to rate because the frame of reference for measuring the amount of training was unclear. Training times were determined by considering a variety of data collected during the on-site study: employer's hiring requirements, union specifications, workers' qualifications, types of work aids used (e.g., calculators, gauges, etc.), and types of tasks performed (e.g., arithmetic calculations, writing, etc.). The Handbook cautions analysts not to rely too heavily on the qualifications demanded by the employer or union or on those that workers bring to the job but rather to assign GED and SVP primarily on the basis of skills or tasks intrinsic to job performance. According to most analysts' reports, however, employers' hiring requirements figured prominently in the assignment of these ratings, especially SVP (Miller et al. 1980, pp. 133).

Clearly, one should not expect *à priori* that employer interviews for the ORS will produce results that differ much compared to a replication of the DOT. From this description it appears DOT ratings were derived mostly from employers and other information that are well within the current design of the ORS. It seems unlikely that job observations played a large role in estimating training times, which are not directly observable by brief site visits in any case.

Finally, the NAS conducted two formal analyses of its own. Using the April 1971 CPS file that had been specially coded with DOT occupations and ratings from both the 3rd and 4th editions, the panel found roughly 95% of the workforce received the same ratings on each of the Data, People, Things scales across the two editions, despite the switch to on-site interviews and observations (Miller et al. 1980, p.192f.).[34] Despite the shortcuts noted earlier, including carrying forward of some 3rd edition ratings to expedite completion of the 4th edition, and the limited, small-scale studies indicating consistency-based convergence between desk ratings of job descriptions and ratings from on-site observations, this strength of this finding remains quite surprising. **It is strongly recommended that the ORS refrain from over-interpreting this result without conducting its own verification studies. One point to consider that may not be obvious is the possibility that jobs change more slowly than commonly believed and this contributed partly to the stability of the results.** The ORS will need to understand the pace of job change at some point for long-range planning regarding required frequency of database updates, which may also have implications for the level of budgetary resources available for any given edition. If less frequent updates are needed, then it is possible more resources will be available per version.

The NAS panel's second original analysis was a reliability experiment involving 42 experienced raters selected evenly across 7 field centers, each of whom rated two job descriptions for 12 randomly selected occupations representing a range of job complexity. The job descriptions were selected from high-quality job analyses conducted for the fourth edition. Analysts could consult the *Handbook for Analyzing Jobs* but not the DOT during the experiment. Nearly half completed a comments sheet afterwards, nearly all of whom indicated the job descriptions

[34] Note that the use of the CPS worker file to weight the data means that these figures refer to the percentage of *jobs* in the economy, which is the relevant concept for SSA, not the percentage of DOT *job titles*, which is a commonly-used but suspect indicator of the former.

contained insufficient information for rating physical demands and environmental conditions, while some noted the same problem for SVP. Nevertheless, analysts completed the task leaving very few missing data fields (Miller et al. 1980, pp.169, 318f.).

The experiment found substantial agreement for the physical demands and environmental conditions variables. For 10 of the 11 variables, the percentage of raters giving the modal response averaged 91%, ranging from 84% to 98%; the rate of agreement was 68% for the outlier (Miller et al. 1980, p.331). These are the only results reported that use a concept of absolute agreement, rather than consistency, but they are also not corrected for chance like Cohen's Kappa. The high rates of agreement may reflect low trait prevalence, as well as the quality of raters' judgments that would be observed in other problem contexts. Also, most of the variables are dichotomies, which minimize opportunities for disagreement, all else equal, because only differences in judgments around a single threshold are counted as errors. Nevertheless, taken together with the comments regarding the insufficiency of the job descriptions, it is equally likely that analysts were using shared heuristics and general impressions to fill gaps in factual details that the rating materials left open. This possibility needs to be kept in mind in evaluating results from any study using desk ratings. What appears to be rater convergence on objective characteristics could reflect shared stereotypes and halo effects. **It is recommended that ORS refrain from using desk ratings of job descriptions, which were the basis for the DOT 3rd edition and O*NET's Abilities and Skills ratings.** Although many comparisons of desk ratings and direct job ratings appear encouraging, the NAS experiment calls into question a simple interpretation of those results.

The NAS analyses also used ANOVA procedures to estimate reliabilities for SVP (0.80), Data (0.85), People (0.87), Things (0.46), Strength (0.54), and Location (0.66), as well as the three GED scores, Reason (0.82), Math (0.61), and Language (0.74). The different job descriptions within DOT occupations had a particularly strong impact on reliabilities for Things and Strength, perhaps due to the problems identified in the analysts' comments. Inter-rater differences after controlling for other sources of variation also remained substantial for all variables except Data, People, and GED-Reason (Miller et al. 1980, pp.323ff.). The report concluded, "ratings are substantially affected by the idiosyncrasies of individual analysts" and that to obtain reliable ratings "it will be necessary both to use more raters and more descriptions per occupation and to average the sets of ratings thus obtained." Fortunately, the estimated number of raters that would be needed to obtain acceptable reliabilities did not exceed four using the Spearman-Brown formula (Miller et al. 1980, p.326). ORS should make a rational determination of the number of raters per occupation it will use to achieve a target level of reliability and avoid the influence of idiosyncratic rater effects on final database values. Effective and consistent rater training, clear instructions, and reconciliation of divergent results are also critical to minimize idiosyncratic rater variance.

Although the NAS panel's assessments were quite disappointing in several respects, sociologists have had quite positive results using the DOT scores and close analogues from other instruments before and after the report appeared.

Kohn and Schooler (1973) interviewed a sample of 3,100 male workers in detail about their jobs, as well as collecting self-completed survey data, both of which drew heavily on DOT skill

concepts. Using factor analysis they obtained an index of substantive complexity, which they used extensively in widely-respected substantive research on work and personality. The variables and loadings for the index were complexity of work with data (0.85), people (0.82), and things (- 0.26); overall complexity of the work (0.80); and worker-reported times spent working with data (0.65), people (0.57), and things (- 0.68). The multiple correlation between this index of substantive complexity and the DOT's Data, People, and Things ratings was 0.78 (Kohn and Schooler 1973, pp.104, 106). Although the issue of shared stereotypes may be raised in this case as well, the large body of high-quality results produced with the data is strong evidence of their validity (see Kohn and Schooler 1983, Spenner 1988).

Another result relevant to the ORS' level of occupational aggregation is the finding that 584 3-digit 1970 Census occupations accounted for 72-77% of the variance in DOT scores for Data, People, Things, GED, and SVP (Spenner 1980, p.243f.). The somewhat finer codes used in the SOC could well account for a larger proportion of the variance of the more consistently concrete ORS items.

Handel (2000) conducted extensive validation exercises for GED and SVP using closely related questions in worker surveys: the Quality of Employment Survey (1977), and the Panel Study of Income Dynamics (PSID) samples of heads and wives (1976, 1985) and heads only (1978). The job complexity items asked in the PSID 1976 wave were repeated for household heads in the 1978 wave. Using reported job tenure and three-digit occupation and industry to code respondents as either job stayers or leavers, Handel found the rate-rerate correlation was 0.83 (n=1,356) for the job's required formal education and 0.60 (n=1,446) for the time required for an average person to become fully trained and qualified. As with intra-rater reliabilities of this type, one cannot exclude the possibility that true change over time is depressing values (e.g., internal promotions). The corresponding correlations for people who changed jobs, occupations, and industries were 0.51 (n=228) and 0.22 (n=257), substantial reductions of 0.32 and 0.38 (Handel 2000, pp.189f.). (All statistics cited from this work are simple Pearson correlations, not agreement-based ICC or Kappa statistics, as would be optimal.) The results for job stayers give a plausible sense of the magnitudes of rate-rerate correlations that ORS can anticipate, at least for incumbents. The contrasting results for job changers is evidence of divergent validity and suggest strongly that the ratings have information content beyond halos or social stereotypes.

At the level of Census occupation means, the rate-rerate correlations for job stayers were 0.95 for required education and 0.85 for training time, increases of 0.12 and 0.25 relative to the job-level correlations (Handel 2000, p.210). This could be taken as a substantial improvement in reliability as a result of averaging, though Harvey (2009) has expressed concern that occupational means may produce inflated reliabilities due to aggregation bias.

Some sense of the relationships between job-level and occupation-level measures and the relationship between incumbent and expert ratings can be gleaned from the correlation table below (Handel 2000, pp.280f.). In both panels the first three variables are job-level measures from the incumbent surveys, followed by three corresponding occupation-level means calculated from incumbent responses, and finally two parallel measures from the DOT aggregated to the level of 1970 Census 3-digit occupations. For both the PSID and the QES, correlations between job- and occupation-level measures are about 0. 80 for the job's required education ("GED") and 0.65 for training times ("SVP") (see cells [5,2] and [6,3] in each panel). For comparison, the

correlation between personal education and its occupational mean varies slightly around 0.76 (cell 4,1). Correlations between incumbent job-level and DOT scores are about 0.67 for GED in both surveys, and 0.57 for SVP, rising to about 0.84 for GED when occupation-level incumbent scores are used, a gain of 0.17. When occupation-level incumbent scores are used for SVP the correlation with DOT scores rises to a similar level for the PSID, a 0.25 increase, but rises much less to 0.65 in the QES, increasing 0.10.

These results show 3-digit 1970 Census occupation codes, somewhat coarser than SOC codes, explain two-thirds of the variance in required education measured at the job level and about 42% of the variance in reported training times. These magnitudes are substantively significant and give some sense of the values ORS may expect to find. Though they are below unity, the rate-rerate correlations from the PSID point to random error variance as the source of some

Correlations of Individual- and Occupation-Level Skill Measures from Panel Study of Income Dynamics (1976), Quality of Employment Survey (1977), and Dictionary of Occupational Titles (1977)

	Panel Study of Income Dynamics (PSID76)						
Variable	1	2	3	4	5	6	7
A. PSID Measures							
Individual-level:							
1. Own education							
2. Education required ("GED")	.68						
3. (ln) Training time ("SVP")	.36	.46					
Occupation-level:							
4. Own education	.73	.65	.37				
5. Education required ("GED")	.68	**.81**	.47	.81			
6. (ln) Training time ("SVP")	.41	.49	**.66**	.49	.62		
B. DOT Measures							
Occupation-level:							
7. GED	.60	**.69**	.53	.72	**.85**	.76	
8. SVP	.47	.56	**.58**	.57	.69	**.83**	.88

	Quality of Employment Survey (QES77)						
Variable	1	2	3	4	5	6	7
A. QES Measures							
Individual-level:							
1. Own education							
2. Education required ("GED")	.70						
3. (ln) Training time ("SVP")	.29	.42					
Occupation-level:							
4. Own education	.79	.69	.35				
5. Education required ("GED")	.68	**.80**	.44	.86			
6. (ln) Training time ("SVP")	.29	.42	**.65**	.40	.53		
B. DOT Measures							
Occupation-level:							
7. GED	.57	**.66**	.51	.73	**.83**	.61	
8. SVP	.42	.52	**.55**	.54	.65	**.65**	.88

Note: DOT measures merged onto PSID and QES data sets using the extended occupational classification scheme developed for mapping (DOT) codes into Census occupation and industry codes. All occupation-level measures from the PSID and QES are means of individual-level measures, where "occupation" refers to the extended occupation classification. "Own education" is respondent's actual level of education (measured in nine categories in the QES), "education required" is formal education required for job, "training time" is length of time required to learn job/become fully trained, "GED" (General Educational Development) is a DOT variable measuring level of formal education required on the job, and "SVP" (Specific Vocational Preparation) is a DOT variable measuring the amount of time a typical worker needs to achieve average performance in a specific job-worker situation (see U.S. Department of Labor 1991 for further details on DOT variables). The PSID includes all household heads and wives coded with 3-digit Census occupation and industry codes and the QES includes all persons working at least twenty hours per week. Sample sizes are approximately 3,500-4,000 for the PSID and 1,100 for the QES.

Source: Handel (2000, pp.280f.)

individual-level variation, which averaging is intended to eliminate. Reanalyses using absolute agreement and other concepts more appropriate to the ORS program would provide more direct evidence on the consequences of using job- vs. occupation-level measures for estimating prevalence, which is SSA's focus of interest.

The correlations between incumbent and DOT scores give some indication of the consequences of aggregating first the latter, "gold standard" ratings to the Census occupational level and then aggregating the job-level incumbent scores to the same level. These results show reasonable to strong convergent validity and no obvious sign that aggregation of either job-level or DOT-level scores to levels similar to SOC codes will cause problems, though further study is needed using metrics more appropriate to ORS criteria.

One of the limitations of these correlations is the absence of a criterion against which the job- and occupation-level measures can be evaluated. Interestingly, Handel found that job-level and occupation-level skill scores from incumbents predicted wages equally well in regression analyses, suggesting that true and error variation within occupations canceled out almost exactly (2000, p.189 fn.3). Using measures of cognitive and physical job demands based on Handel's survey of Skills, Technology, and Management Practices (STAMP), Autor and Handel (2013, p.S85) also found occupation-level means and job-level values reported by incumbents produced very similar coefficients and R^2 in wage regressions. When the model was expanded to include both kinds of variables, the two coefficients for cognitive demands fell but remained nearly equal and remained significant. By contrast, only the coefficient for the job-level physical demands variable remained significant in the joint model. From these results, it appears that occupational means have strong criterion validity but do not capture all meaningful variation found in job-level scores.

Finally, Handel (2015) correlated STAMP measures of job autonomy/discretion and cognitive, interpersonal, and physical jobs demands with parallel measures from the DOT and O*NET. The following two tables show these SOC occupation-level correlations, as well as their correlations with occupation mean wages from the CPS and measures of occupational prestige and socio-economic status commonly used in sociological research. The correlations in bold indicate convergent validity coefficients. STAMP required education and SVP-like variables follow O*NET's format. The correlations between wages and required education for STAMP (0.81), O*NET (0.77), and the DOT (0.79) are nearly identical, even though the DOT ratings are much older than the wage criterion, which is from 2005-2008. Both the DOT and O*NET GED-type measures are closely and almost equally strongly related to the STAMP measure, as well. Although there are individual variations, what is striking is how closely the DOT variables are related to both STAMP measures of similar constructs (convergent validity) and to recent wages (criterion validity), and the similarity between these values and both the STAMP-O*NET correlations and the latters' correlations with recent wage levels (convergent validity). The convergent validity coefficients are also often, though not invariably, greater than the correlations with constructs that differ on their face from the focal variable (divergent validity), though the clustering of related constructs in the tables mutes this effect. For all the problems identified by the NAS report, these results affirm its conclusion that "the DOT worker functions and worker traits constitute one of the richest sources of occupational data available anywhere" (Miller et al. 1980, p.173).

Occupation-Level Correlations between STAMP and Alternative Measures: Skills, Autonomy, and Authority

	Education	Experience	Training	Math	Verbal	Problem	(ln) Wage
DOT							
GED	**0.89**	0.46	0.66	0.65	0.90	0.79	0.79
SVP	0.79	**0.65**	**0.70**	0.69	0.83	0.78	0.81
Numerical	0.67	0.54	0.47	**0.64**	0.69	0.66	0.71
Verbal	0.84	0.46	0.57	0.57	**0.88**	0.71	0.74
Data	0.74	0.62	0.61	0.67	0.80	**0.74**	0.76
O*NET							
Education	**0.92**	0.42	0.63	0.54	0.86	0.67	0.77
Experience	0.59	**0.78**	0.56	0.57	0.66	0.66	0.77
Training	0.49	0.52	**0.63**	0.58	0.52	0.57	0.61
OJT	0.52	0.57	**0.66**	0.64	0.54	0.61	0.70
Math	0.60	0.59	0.51	**0.73**	0.61	0.63	0.72
Verbal	0.86	0.50	0.60	0.55	**0.91**	0.76	0.78
Cognitive	0.87	0.63	0.71	0.64	0.89	**0.82**	0.89
Other							
Occ. prestige	0.88	0.39	0.67	0.61	0.86	0.76	0.79
SEI	0.90	0.37	0.70	0.62	0.85	0.73	0.83
(ln) Wage	0.81	0.71	0.76	0.69	0.83	0.78	

	Index	Autonomy	Decision	Supervision	Repetitive	(ln) Wage
DOT						
DCP	**0.76**	0.46	**0.78**	**-0.50**	-0.53	0.64
REPCON	-0.51	-0.31	-0.38	0.45	**0.42**	-0.53
O*NET						
Autonomy	**0.75**	**0.50**	0.64	-0.49	-0.63	0.69
Management	**0.75**	0.38	**0.75**	**-0.51**	-0.62	0.81
Repetitive	-0.57	-0.32	-0.51	0.37	**0.53**	-0.60
Other						
Occ. prestige	0.56	0.26	0.40	-0.56	-0.53	
SEI	0.58	0.27	0.40	-0.60	-0.55	
(ln) Wage	0.74	0.41	0.61	-0.54	-0.70	

Note: Correlations weighted by sample size within STAMP occupations. Correlations indicating level of convergent validity in bold. The STAMP variable 'Index' is an additive composite of autonomy, decision-making, closeness of supervision, and task repetitiveness. *Source: Handel (2015)*

Key for DOT, O*NET, and other variables

GED=General Educational Development
SVP=Specific Vocational Preparation
Numerical=numerical aptitude quantile
Verbal=verbal aptitude quantile
Data=workers' relationship to data
Education=required education

Experience=length of related work experience in other jobs
Training=length of employer-provided classroom study
OJT=on the job training
Cognitive=general cognitive ability scale combining variables for (1) analytical thinking; (2) critical thinking; (3) complex problem solving; (4) active learning; (5) analysing data or information; (6) processing information; (7) thinking creatively; (8) updating and using relevant knowledge; (9) deductive reasoning; (10) inductive reasoning; (11) fluency of ideas; and (12) category flexibility (α=0.97)
Math=scale combining variables for (1) mathematics skills; (2) mathematics knowledge; (3) mathematical reasoning; and (4) number facility (α=0.92)
Verbal=scale combining variables for (1) reading comprehension; (2) writing skills; (3) writing comprehension; (4) writing ability; (5) knowledge English language rules (spelling, grammar, composition); and (6) frequency of using written letters and memos (α=0.95)
DCP=temperament suitable to accept responsibility to direct, control, or plan an activity
REPCON=temperament suitable to perform repetitive work or continuously perform same, fixed task
Management=scale combining variables for (1) judgment and decision-making; (2) management of financial resources; (3) management of human resources; (4) knowledge of business administration and management; (5) decision-making and problem solving; (6) developing objectives and strategies; (7) coaching and developing others; and (8) guiding, directing, and motivating subordinates (α=0.93)
Autonomy=scale combining variables for (1) freedom to make decisions without supervision and (2) freedom to determine tasks, priorities, or goals (α=0.89)
Repetitive=time spent making repetitive motions
Occ. prestige=occupational prestige scores
SEI=socioeconomic index scores
ln (wage)=log mean occupational hourly wages (October 2004–January 2006)

Occupation-Level Correlations between STAMP and Alternative Measures: Interpersonal and Physical Requirements

	Interpersonal	Presentation	Freq	Level		(ln) Wage
DOT						
People	**0.75**	0.62	**0.69**	**0.69**		0.38
O*NET						
People	**0.86**	0.76	0.67	0.69		0.62
Speaking	0.71	**0.79**	0.51	0.56		0.62
Customer/public	0.73	0.47	**0.76**	**0.73**		0.37
Other						
Occ. prestige	0.65	0.71	0.47	0.49		
SEI	0.65	0.74	0.51	0.52		
(ln) Wage	0.63	0.77	0.43	0.42		
	Physical	Stand	Lift	Coord	Demands	(ln) Wage
DOT						
Dexterity	0.34	0.21	0.32	**0.44**	0.30	-0.28
Effort	**0.83**	**0.62**	**0.78**	0.78	**0.83**	-0.52
O*NET						
Craft skills	0.33	0.25	0.36	0.32	0.29	0.06
Fine motor	0.73	0.51	0.71	**0.72**	0.70	-0.32
Gross physical	**0.91**	**0.84**	**0.79**	0.79	**0.87**	-0.51
Other						
Occ. prestige	-0.42	-0.26	-0.49	-0.32	-0.45	
SEI	-0.47	-0.27	-0.54	-0.39	-0.52	
(ln) Wage	-0.50	-0.41	-0.45	-0.43	-0.53	

Note: Correlations weighted by sample size within STAMP occupations. Correlations indicating level of convergent validity in bold. In the top panel, "Interpersonal" is the STAMP scale for interpersonal job requirements from Table 3. "Freq" is frequency of contact with people other than co-workers, such as customers and clients. "Level" is self-rated importance of working well with people other than co-workers. In the bottom panel, "Physical" is a scale (σ=0.79) composed of job requirements to stand for 2 hours, lift 50 lbs., have good eye-hand coordination, and self-rated physical demands of job (0=not all physically demanding, 10=extremely physically demanding). *Source: Handel (2015)*

Key for DOT, O*NET, and other variables:

People (DOT): Scale combining (1) workers' relationship to people (People) and (2) temperament for dealing with people beyond giving and receiving instructions (DEPL) (α=0.74)

People (O*NET): Scale combining variables for (1) persuasion; (2) negotiation; (3) speaking skills; (4) instructing skills; (5) service orientation; (6) dealing with angry people; (7) dealing with physically aggressive people; (8) frequency of conflict situations; (9) dealing with external customers or public; (10) frequency of face-to-face discussions; (11) frequency of public speaking; (12) resolving conflicts and negotiating with others; (13) communicating with persons outside organization; (14) performing for or working directly with the public; (15) training and teaching others; (16) interpreting the meaning of information for others; (17) customer and personal service knowledge; (18) education and training knowledge; (19) social orientation; (20) and social perceptiveness (α=0.94)

Speak: Subscale of O*NET people scale, combining variables for (1) speaking skills and (2) frequency of public speaking (α=0.70)

Customer/public: Subscale of O*NET people scale, combining variables for (1) service orientation, (2) dealing with external customers or public, (3) customer and personal service knowledge (α=0.86)

Dexterity: Scale combining (1) workers' relationship to things (Things); (2) finger dexterity; (3) manual dexterity; and (4) motor coordination (α=0.85)

Effort: Scale combining variables for (1) strength, and the sum of responses to the dichotomous variables (2) climb, (3) stoop, and (4) reach (α=0.81)

Craft skills: Scale combining variables for (1) controlling machines and processes; (2) repairing and maintaining mechanical equipment; (3) repairing and maintaining electronic equipment; (4) equipment maintenance; (5) troubleshooting operating errors; (6) repairing machines; and (7) installing equipment, machines, and wiring (α=0.95)

Fine motor: Scale combining variables for (1) finger dexterity; (2) manual dexterity; (3) arm-hand steadiness; (4) multi-limb coordination; (5) rate control (ability to time movements); (6) operating vehicles, mechanized devices, or equipment; and (7) time spent using hands to handle, control, or feel objects, tools, or controls (α=0.94)

Gross physical: Scale combining variables for (1) handling and moving objects; (2) general physical activities; (3) static strength; (4) dynamic strength; (5) trunk strength; (6) stamina; and time spent (7) sitting, (8) standing, (9) walking, (10) twisting body, (11) kneeling, crouching, stooping, or crawling (α=0.98)

Occ. prestige=occupational prestige scores

SEI=socio-economic index scores

ln (wage)=log mean occupational hourly wages (October 2004-January 2006)

4. Conclusion

Stakeholders have expressed great concern that the DOT will not be replicated for use in SSA's disability determination process. While the DOT has numerous strengths, it has also been subject to a number of misconceptions and idealizations. Yet, despite the DOT's methodological problems, there is much evidence that its ratings have strong construct and criterion validity. Its analyst ratings also correlate well with similar measures from incumbent surveys from different eras. Evidence from those surveys, as well as the DOT, argue for caution regarding any assumption that aggregation to SOC occupations has obvious implications for data quality. Any loss in genuine detail has to be weighed against the gain in reliability from eliminating random error, which was not absent from the DOT given the small base of rating data for many occupations. Further knowledge of the nature and significance of the tradeoff between true and error variation in the context of the ORS program requires further investigation.

*B. O*NET*

The strengths and limitations of O*NET have been discussed in detail in the Historical Background section of this report and elsewhere (Handel *forthcoming*; Hilton and Tippins 2010). Ratings and descriptors for job activities are not tied to observable behaviors that are needed for determining whether SSA applicants are capable of performing work that is available. Indeed, dissatisfaction with O*NET was a primary reason SSA embarked on its own program to collect occupational information. Nevertheless, the previous section showed good convergent validity and O*NET is one of the few large-scale data collection programs dealing with some of the same topics as the ORS. Given the meager state of occupational databases and the comparability of target samples, unlike most IO psychology studies, O*NET measures should be one source of construct validation for ORS. O*NET's experiences may provide some indication of the reliability issues the ORS will encounter. The levels of variation within SOC codes are of particular interest.

O*NET microdata could be used to calculate r_{wg} to determine the level of within-SOC agreement for SVP, working conditions, and the more behaviorally concrete physical demand items on the Work Context questionnaire. The frequency distributions for these variables can be calculated at both the job-level and SOC-level to determine if estimates of trait prevalence are sensitive to occupational aggregation. **The ORS should consider studies examining the reliability and within-occupation heterogeneity in O*NET microdata for variables that are close in substance and form to those under consideration for the ORS.**

If there happen to be variables, such as overall exertional requirements, that are comparable to the DOT it may be possible to calculate agreement DOT field raters and O*NET incumbents for specific SOC occupations that one would expect have not changed greatly based on prior knowledge (e.g., taxi and bus drivers, wait staff, cleaners, electricians, etc.) and compare levels of agreement to those for occupations that research suggests have changed greatly.

Nevertheless, the limits of such comparisons must also be recognized given the large differences between ORS constructs and most other O*NET variables in terms of the latters' level of abstraction, use of jargon, holism, and vagueness. Unfortunately, most O*NET reliability studies seem to focus on the O*NET variables that suffer most from these qualities (e.g., the Ability survey variables), rather than the ones closes to ORS variables, such as required education, SVP, Work Context items, and specific tasks performed on the job.

Most O*NET documents found for this report dealt with variables from O*NET's Ability and Skills domains. Initially, job incumbents were to complete instruments for both domains but incumbents' difficulty understanding and answering the Ability items led to the use of job analysts to complete the surveys at their desks on the basis of job descriptions, task information, and O*NET ratings for other variables made by incumbents (Tsacoumis and Willison 2010, Appendix G). The method was extended to the Skills survey, as well, after many years of using incumbents and presumably for the same reason. The ratings are currently performed by two groups of eight professionally qualified job analysts with graduate degrees in IO psychology or similar fields and at least five years of work relevant experience (Lewis et al 2011, slides 50ff.). If the standard error of a rating mean ($SE_{\bar{y}}$) exceeded 0.51 the level of agreement across raters is considered insufficient because the upper and lower bounds of the confidence interval are more than one scale point distant from the observed mean. In such cases efforts are made to reconcile divergent results and increase the level of agreement (Tsacoumis and Willison 2010, p.13). The target is for median ICC (3, *k*) values to be at least 0.80 using a consistency criterion rather than the more stringent standard of absolute agreement required by ORS.[35] Indices were constructed for raters' consistency in differentiating occupations with respect to a construct's Importance and Level, as well as for their ranking of constructs within occupations, which is a relative measure that has no relevance for the ORS. Each cycle involves rating a subset of occupations and as project documents note:

> *…it is important to note that this reliability is dependent on the sample of occupations being rated. That is, all else being equal, the ICC(3, k) based on ratings of a sample of*

[35] "The type of reliability of most interest in this situation is the extent to which raters agree about the order of and relative distance between occupations on a particular scale for a particular construct. For example, is there consistency across raters in how they differentiate among occupations on the required level of the skill *Critical Thinking*?" (Tsacoumis and Willison 2010, p.9).

> *homogeneous occupations will be lower than the ICC(3, k) based on ratings of a sample of heterogeneous occupations. It is important to keep this point in mind when interpreting the reliability results* (Tsacoumis and Willison 2010, p.14).

It should also be noted that as occupations are re-rated over successive updates to the O*NET database the raters receive the same materials as before with additions presented in bold and with asterisks, and deletions indicated with strikethroughs. After making a final rating the consensus rating from the previous cycle is displayed.

> *Analysts were instructed to consider the previous average rating, along with changes to the occupational information since the previous time the occupation was rated (information in bold with an asterisk and crossed out information)… Once they gave full consideration to all relevant information, analysts entered their final importance rating* (Fleisher and Tsacoumis 2012a, p.7).

While there is no perfect way to balance competing considerations, such as both minimizing random error and maximizing the independence of ratings, it is possible that this method produces a downward bias in estimates of occupational change. This could be investigated using changes in incumbent average ratings across cycles for similar constructs in other O*NET surveys, given the significant redundancy in constructs across the surveys. On a more general level, Harvey (2009) argues that analyst agreement is likely biased upward even in the cross-section because the specific stimulus is a common, single job description per occupation and the vagueness of O*NET constructs leads analysts to fall back on general impressions (i.e., halos) to guide the ratings they assign to specific items, among other criticisms.

A recent reliability report found the $SE_{\bar{y}}$ rarely exceeded 0.51 for analysts' Skills ratings currently or in the past (Fleisher and Tsacoumis 2012b, p.3). The median standard deviation of analyst ratings was 0.46 and the $SE_{\bar{y}}$ was 0.16, which is interpreted as indicating high levels of inter-rater absolute agreement. Almost all ICC(3,8) were above 0.80 and most above 0.85, while medians for ICC(3,1) were 0.43 for Importance ratings and .052 for Level ratings for the same 35 Skill constructs (Fleisher and Tsacoumis 2012b, p.6ff.). Again, these figures refer to consistency in assessments of "the order and relative distance among occupations on particular constructs for importance and level," not absolute agreement regarding the values of the ratings (Fleisher and Tsacoumis 2012b, p.8). This appears to exclude cases where analysts agreed a construct did not apply to a particular occupation, avoiding a source of inflated reliabilities that Harvey identified discussed above. Summary figures reported for a subsequent cycle representing a different sub-sample of occupations are nearly identical (Reeder and Tsacoumis 2014, p.3, 6ff).

A comparison of O*NET ratings from two groups of eight analysts and 10,017 job incumbents for a diverse sample of 289 SOC occupations found the analysts showed much greater consistency in rating the relative importance of skills within occupations than did incumbents. The median single-rater ICC for incumbents was 0.44, while the ICC for analysts was 0.72. Of course, the incumbents were rating their actual jobs while analysts were desk-rating common stimulus materials (titles, job descriptions, other O*NET ratings, etc.). Because the extent to which the difference in medians across groups reflects heterogeneity among incumbent rating

targets is unknown, the comparison is not a clean one.[36] The analysts were not paired to incumbents, so there is no way to distinguish differences that are artifacts of differential rating leniency from true differences in the jobs rated by incumbents and analysts. **Clearly, any ORS comparisons between analyst and incumbent ratings based on newly collected data should avoid this design flaw and ensure both are rating the <u>same</u> jobs, if at all possible.**
Interestingly, the simple correlation between incumbent and analyst single-rater coefficients across 20 major SOC groups was 0.72, indicating that the same jobs tended to produce relatively greater or lesser consistency for both analysts and incumbents. Standard deviations also indicated greater absolute disagreement among incumbents ($mean_{sd}$=1.04) than analysts ($mean_{sd}$=0.56) on a 5-point scale, but the targets rated were more diverse among the incumbents, as well. The level of disagreement did not vary by major SOC group for either incumbents or analysts. Averaging the ratings within the two rater groups produced strong and identical ICC and $SE_{\bar{y}}$ values across the groups because the average incumbent sample size per occupation was much larger (n=35) than the analyst sample size (n=8). The authors conclude this shows high reliability is possible using either type of rater (Tsacoumis and Van Iddekinge, pp.4ff.).

Incumbents' mean ratings by major SOC group averaged 0.47 SD higher than analyst means major SOC groups, calculated using Cohen's *d*. The differences were greatest for construction (0.77), maintenance and repair (0.67), and production (0.79) occupations, which are likely to be of particular interest to ORS (Tsacoumis and Van Iddekinge, pp.13f.). "However, in the absence of some external criterion or 'gold standard,' we cannot conclude that the higher ratings that incumbents tend to provide necessarily represent rating 'errors'" (Tsacoumis and Van Iddekinge, pp.17). As the authors note, determining the significance of these differences would require an assessment of their impacts on the actual practices they were intended to inform. In addition, using a standardized effect size measure obscures the absolute size of the differences in the original units, which have no justification for analyses of a criterion-based database like the ORS.

Incumbents and analysts ordered the detailed SOCs on a given skill in very similar ways, yielding SOC-level correlations for the 35 skills that averaged 0.69 and ranged from 0.45 to 0.80. The report also concludes that some of the lower skill correlations, such as for *Systems Analysis*, "were likely due to the relative lack of variation in the importance of such skills across SOCs rather than to incumbent-analyst inconsistency" (Tsacoumis and Van Iddekinge, p.14). The correlations of ratings across skills within SOC codes averaged 0.80 and showed much less variation across major occupation categories, indicating high consistency in the ranking of skills within each SOC across the two rater types (Tsacoumis and Van Iddekinge, p.14ff.).

Walmsley, Natali, and Campbell (2012) took advantage of O*NET's decision to transfer Skill ratings from incumbents to analysts between O*NET 14.0 and O*NET 15.0. They compared mean incumbent and analyst ratings of the importance of 35 Skill items across approximately 700 occupations using the two versions of the database. Incumbent skill ratings averaged 0.73 SD higher than analysts using Cohen's *d*, but the median difference was rather less (0.54 SD). The size of the differences was negatively related to analysts' mean ratings (r = -0.34), i.e.,

[36] "It is also important to note that incumbents often represent different jobs within a given SOC. Therefore, the relatively lower interrater reliability/agreement estimates we found may reflect true differences in the importance of O*NET skills across similar but different jobs" (Tsacoumis and Van Iddeking, p.17).

incumbent ratings exceeded analyst ratings by a larger margin for skills that analysts rated relatively low. This can be seen in the graph below that plots the standardized difference in means on the y-axis against analyst means on the x-axis along with the regression line. Fortunately, this article provided enough information to calculate the raw differences, as well. O*NET's Importance scales have five points and the difference between rater group means across the 35 Skills was 0.44 scale points. About one-third of the differences were in each of the scale-point intervals, (-0.10 – 0.24), (0.26 – 0.55), and (0.59 – 1.34). The correlation between raw-score differences and analyst rater means is -0.49, as can be seen in the second graph.

Again, the level of absolute agreement discussed above was compatible with relatively high correlations between the two groups. The median correlation of skill ratings across detailed SOC occupations was 0.67, indicating reasonable consistency in ranking occupations. About one-third of the correlations were in each of the intervals, (0.34 - 0.60), (0.61 - 0.70), and (0.71 - 0.82) (Walmsley et al. 2012, pp.288ff.). The table on which all of these calculations were based is reproduced below (Walmsley et al. 2012, p.290).

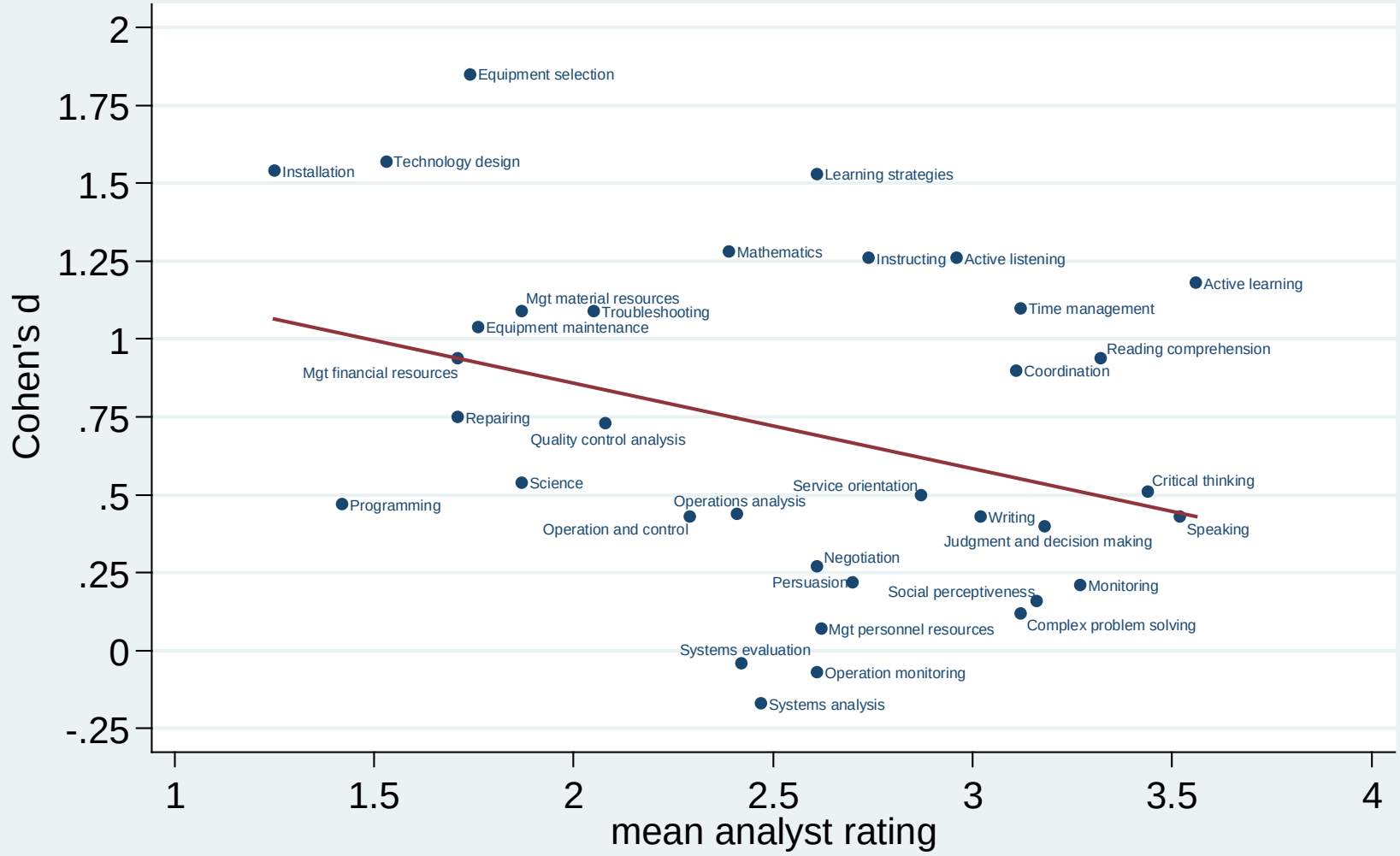
Obviously, the O*NET results for Skills can be only suggestive given the great difference in question format and measurement approach, but this section offers ideas for similar analyses using O*NET data on required education, SVP, Work Context items, and specific task statements, as well as ORS items.

Table 2. Correlations among O*NET 14.0 and 15.0 skill importance ratings across occupations

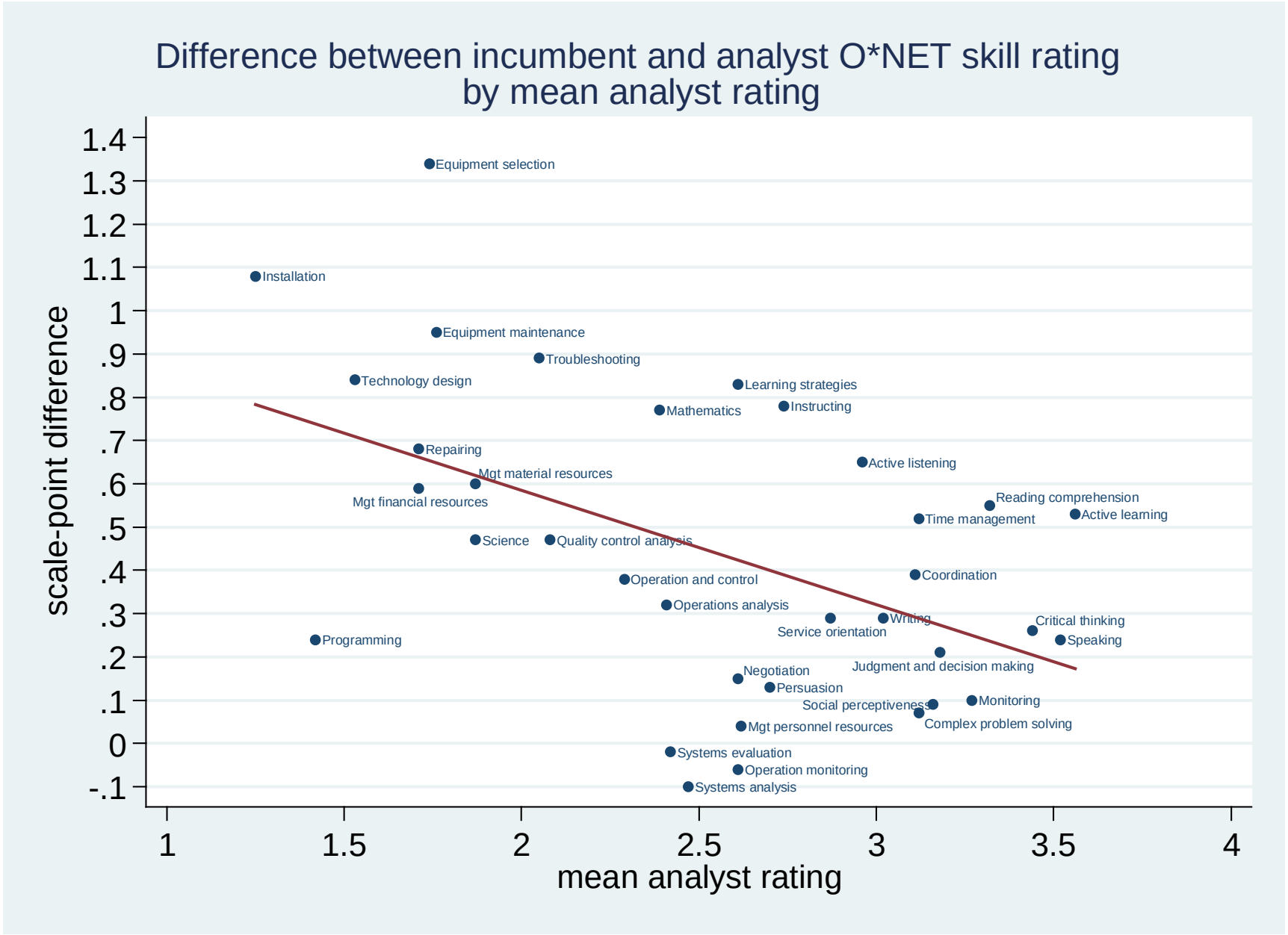
O*NET descriptor	Correlation 14.0 and 15.0	O*NET 14.0		O*NET 15.0		Cohen's *d*
		M	*SD*	M	*SD*	
Repairing	.82 (.86)	2.39	0.93	1.71	0.89	0.75
Writing	.81 (.90)	3.31	0.74	3.02	0.61	0.43
Science	.81 (.87)	2.34	0.91	1.87	0.83	0.54
Operation monitoring	.80 (.87)	2.55	0.95	2.61	0.82	–0.07
Equipment maintenance	.78 (.82)	2.71	0.95	1.76	0.88	1.04
Reading comprehension	.77 (.87)	3.87	0.61	3.32	0.56	0.94
Speaking	.75 (.86)	3.76	0.58	3.52	0.53	0.43
Operation and control	.75 (.81)	2.67	0.83	2.29	0.93	0.43
Troubleshooting	.75 (.82)	2.94	0.82	2.05	0.81	1.09
Critical thinking	.74 (.87)	3.70	0.59	3.44	0.43	0.51
Social perceptiveness	.72 (.86)	3.25	0.65	3.16	0.44	0.16
Complex problem solving	.72 (.87)	3.19	0.68	3.12	0.46	0.12
Active listening	.70 (.85)	3.61	0.53	2.96	0.50	1.26
Active learning	.70 (.84)	4.09	0.46	3.56	0.44	1.18
Persuasion	.69 (.85)	2.83	0.68	2.70	0.47	0.22
Negotiation	.68 (.86)	2.76	0.62	2.61	0.49	0.27
Service orientation	.68 (.82)	3.16	0.64	2.87	0.53	0.50
Installation	.67 (.71)	2.33	0.88	1.25	0.48	1.54
Instructing	.66 (.75)	3.52	0.62	2.74	0.62	1.26
Judgment and decision making	.64 (.79)	3.39	0.61	3.18	0.42	0.40
Learning strategies	.62 (.76)	3.44	0.50	2.61	0.58	1.53
Management of personnel resources	.62 (.75)	2.66	0.65	2.62	0.52	0.07
Mathematics	.61 (.71)	3.16	0.66	2.39	0.54	1.28
Equipment selection	.59 (.66)	3.08	0.72	1.74	0.73	1.85
Time management	.59 (.74)	3.64	0.56	3.12	0.37	1.10
Monitoring	.56 (.71)	3.37	0.56	3.27	0.37	0.21
Coordination	.55 (.72)	3.50	0.48	3.11	0.38	0.90
Technology design	.53 (.70)	2.37	0.65	1.53	0.40	1.57
Management of financial resources	.53 (.64)	2.30	0.72	1.71	0.52	0.94
Programming	.52 (.63)	1.66	0.54	1.42	0.48	0.47
Operations analysis	.50 (.60)	2.73	0.70	2.41	0.74	0.44
Quality control analysis	.50 (.59)	2.55	0.62	2.08	0.66	0.73
Management of material resources	.43 (.55)	2.47	0.60	1.87	0.50	1.09
Systems evaluation	.42 (.55)	2.40	0.56	2.42	0.54	–0.04
Systems analysis	.34 (.41)	2.37	0.62	2.47	0.58	–0.17
Mean	.64 (.76)	2.97	0.67	2.53	0.57	0.73[a]
SD	.12 (.12)	.57	0.13	.65	0.16	0.51[a]
Minimum	.34 (.41)	1.66	0.46	1.25	0.37	–0.17
Maximum	.82 (.90)	4.09	0.95	3.56	0.93	1.85

Notes: M = mean; *SD* = standard deviation. 14.0 skill ratings were made by incumbents and 15.0 skill ratings were made by analysts. Values in parentheses are correlations corrected for attenuation due to unreliability. Negative *d* values indicate that mean skill ratings are higher in O*NET 15.0 than 14.0. [a]Mean and standard deviation of the distribution of *d* values were calculated ignoring the sign for each *d* (i.e. all *d*s were positive for the M and *SD* calculation). This is done because the sign of *d* is contingent on which mean is entered first into the equation; essentially an arbitrary decision. This provides a summary of the magnitude and distribution of *d* values for O*NET 14.0 and 15.0 skills that is 'unbiased' by the sign.

Difference between incumbent and analyst O*NET skill rating by mean analyst rating



From Walmsley et al. (2012) Table 2


Difference between incumbent and analyst O*NET skill rating
by mean analyst rating
scale-point difference
1.4
1.3
1.2
1.1
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
-.1
1
1.5
2
2.5
3
3.5
4
mean analyst rating
Equipment selection
Installation
Equipment maintenance
Troubleshooting
Technology design
Learning strategies
Mathematics
Instructing
Repairing
Active listening
Mgt material resources
Mgt financial resources
Reading comprehension
Time management
Active learning
Science
Quality control analysis
Coordination
Operation and control
Operations analysis
Writing
Critical thinking
Service orientation
Programming
Speaking
Judgment and decision making
Negotiation
Persuasion
Social perceptiveness
Monitoring
Complex problem solving
Mgt personnel resources
Systems evaluation
Operation monitoring
Systems analysis

C. NCS

As far as can be determined, the semi-structured nature of the interview and some of the leveling dimensions are similar ORS (e.g., cognitive level, structured nature of work, supervisory responsibilities, contacts). However, it appears that there are no internal reports of reliability and validity available for the NCS, though it is believed they were conducted. Needless to say, for the reasons enumerated at the beginning of this report, the ORS should not only conduct careful validation studies but also ensure they remain part of the program's documents.

Because the ORS methods of data collection are most similar to the NCS, the absence of a validation record represents both a significant gap and an opportunity. It should not be difficult for the ORS to conduct studies using available data to compare NCS ratings with other sources of ratings, including any suitable ORS pilot results, for evidence of convergent and divergent validity. Naturally, the different units of measurement between the NCS and other surveys mean any validity coefficients (i.e., correlations) would reflect consistency and not absolute agreement on trait levels.

If the NCS data permit, it may be possible to study inter-rater reliability and method effects within the NCS, as well. Particularly helpful would be an investigation of the magnitudes of any differences between telephone interviews and in-personal interviews, and their relative consistency with job ratings from other sources. The NCS may also be useful for understanding the magnitude of within-occupation heterogeneity across establishments and employers. Ideally, many 3-digit SOC occupations would have multiple ratings from different employers. If occupation accounts for the vast majority of the variation, this would strengthen the general case for using SOC codes, bearing in mind previous comments regarding the problems with metrics such as proportion of variance explained and the utility of sub-occupation codes for particularly significant cases. Indeed, the level of residual variation after accounting for occupation should be examined to see if there are occupations that might require finer treatment in the ORS occupational scheme and data collection procedures.

D. ORS EXPERIENCE

Aside from the DOT, O*NET, and the NCS, the experience of the ORS to date is the most obvious source for information on the reliability, validity, and accuracy of different approaches to collecting the kind of data the ORS seeks to collect. The *Consolidated Feasibility Test Summary Report* (2014) provides much useful information but it is primarily impressionistic and omits key details at several points.

1. Interview format

One key point from the Report is that giving field economists greater flexibility and discretion in collection strategies significantly improves data quality and reduces respondents' perception of burden (p.3). Traditionally, most survey professionals have been leery of conversational interviews because they introduce additional, interviewer-related variance into measures due to varying interviewer skills, styles, and perspectives. Indeed, the report elsewhere concludes that "Well-defined and stable procedures are essential to ensuring consistency in the data collected

nationally" and that "inconsistent implementation of procedures" lowered inter-coder reliability and data quality (p.4). The report does not seem to recognize clearly that flexibility and consistency in procedures are usually in tension, if not contradiction, and finding the right balance can be difficult. Standardized interviews and surveys limit interviewer variance, including possible halo effects, but they also limit the interviewer's ability to probe for details, provide helpful prompts tailored to specific situations, and identify and correct respondents' misunderstandings and biases. Indeed, respondents themselves reportedly felt conversational interviews improved the quality of data and eased the interview process (pp.29, 36, 41).

There is not a great deal of research directly comparing the two methods (see Schober and Conrad 1997). If further study shows conversational interviews produce much better data than standardized interviews, this would seem to have the additional implication that relying on self-completed surveys from job incumbents, supervisors, and managers will lower data quality because even the availability of the interviewers to provide limited assistance, such as clarifying survey directions, will be removed.

If surveys are completed on the web drawings and photos may ameliorate common sources of confusion significantly and improve greatly the quality of self-reports. Clocks and calendars could be used to represent time spent on different tasks and photos or drawings could be used to represent body positions, physical actions, and weights of objects. These kinds of visual aids would also likely help improve information from managers collected by phone and email. However, web access and comfort levels are likely to be lowest for the less-skilled and manual workers that are of greatest interest, which is a potential limitation whose impacts would need to be examined.

2. Personal visits and job observations

Field economists also reported that "personal visits resulted in better data collection" (p.6) and job observations greatly helped formulate and target questions, as well as clarify and verify respondents' answers (p.7).

> *Respondents and field economists both expressed a definite preference for personal visit collection overall. Approximately half of the field economists reported that being required to collect the data remotely had a potentially negative impact on survey cooperation because respondents often expressed an unwillingness or inability to spend the time needed for full ORS collection by phone, whereas PV* [personal visit] *respondents were more amenable to the typical interview duration. A number of respondents also indicated that they ordinarily would delay response to or ignore requests for remote collection.*
>
> *The quality of data collected in this test also appeared to likely be affected by collection mode. A significant number of remote respondents (38%) cited time constraints that negatively impacted ORS collection; none of the personal-visit respondents reported feeling rushed. Thirty-seven percent of field economists reported that remote collection was less efficient than PV, and nearly 25 percent indicated that they felt that the quality of their remotely-collected data was worse than that obtained through PV. A number of*

reasons were offered by field economists for the perceived reductions in efficiency and data quality in remote collection, including: respondents' unwillingness to spend adequate time on the phone; field economists' inability to demonstrate and clarify ORS concepts; the absence of respondents' nonverbal cues to sense potential confusion and maintain engagement; and the inability to see the work environment.

Personal-visit respondents reported being less burdened than their remote-respondent counterparts. This finding likely reflects the true impact of mode on perceptions of burden, as well as mode differences affecting respondents' willingness to report on these perceptions. The most frequently cited reasons for feeling burdened (if some burden was reported) was time pressure, feeling rushed, or competing work demands. And although respondent engagement was generally good across modes, several field economists reported situations in which phone respondents' attention waned relatively quickly (e.g., after 15 minutes) and were obviously distracted by interruptions. Respondents cited similar concerns with remote collection (p.53).

A major limitation of this aspect of the Report is that the magnitude of the benefits associated with personal interviews is not addressed systematically. Given its prominence in the DOT and in the fifteen years of SSA discontent with the elimination of this aspect in O*NET, there needs to be more research on this subject.

3. Central office respondents

The Feasibility Test Report supports the suspicions of many that central office respondents often has problems due to their limited personal experience with the work tasks of jobs for which data is sought and because competing demands on their own time (pp.3, 14ff., 45ff.). Not surprisingly, central office personnel often find it difficult to answer questions on task duration that are of great interest to SSA (pp.48, 53f.).

Clearly, the larger the organization the less likely are central office personnel to have detailed and intimate knowledge of work requirements and tasks performed for highly dissimilar jobs performed in distant departments and facilities.

Possible exceptions to this are (1) industrial engineers, who conduct time and motion studies that can be quite detailed, (2) occupational health and safety personnel, and (3) environmental safety personnel (if different), who conduct internal research and oversight and may have to file reports with OSHA and EPA on workplace hazards of different kinds

4. Implications

The ORS faces major design choices. Central office HR respondents can provide useful orientation and written job descriptions can be an excellent source of ideas and prompts for conversational interviews, but the real question is the extent to which the ORS can use personal visits and the consequences of using other methods of large-scale data collection, such as telephone interviews, email, or self-complete questionnaires, like O*NET. Clearly, BLS needs to do a preliminary costing of personal visits to understand for its own purposes the proportion of data that can be gathered feasibly using this collection mode. For the remainder, the ORS will

probably have to conduct experiments to determine the reliability and validity of self-report surveys.

a. Self-completed surveys
If the ORS were to consider self-completed surveys the program will need to devote considerable attention to making the questions easily understood by all respondents. Illustrations and photos can help clarify ambiguities with concepts like "stoop," "crouch," and "kneel," but more complex concepts like "gross manipulation" and "fine manipulation" are likely to be more difficult. Communication of concepts like "Near visual acuity," "Far visual acuity," and "Hear and respond to auditory signals" is likely to be even more difficult.

The ORS needs to guard strenuously against following O*NET's approach, which was to use this kind of professional jargon in surveys distributed to lay persons along with extended definitions at a similar level of reading difficulty, violating the most basic principles of sound survey design. Existing interview materials that are meant to be used by trained staff in person-to-person interviews simply cannot be transferred to self-complete surveys without thoughtful modifications to ensure their clarity for diverse respondents. The Feasibility Report's findings that trained staff frequently required clarification from other program personnel is evidence enough to warn against this approach, as well as reports of respondent difficulty with certain words and concepts (p.42).

Likewise, the ORS is likely to run into difficulties if respondents are asked to self-complete items on task duration. "The most common negative reaction to the survey was that it was difficult to provide accurate duration estimates, either because the respondent did not know the job well enough, they reported that durations varied across people in the job, or because there was a wide range of job duties that were done intermittently" (p.29; see also p.35). Indeed, the duration concepts and level of detail indicated in the ORS Collection Manual are quite subtle and complex (pp.11ff.). It is very difficult to see how this kind of information could be collected without detailed guidance from trained interviewers. The previous quotation also reveals some of the problems involved using managers rather than job incumbents to answer very detailed questions about job tasks.

Other concepts used in the ORS Collection Manual, such as the distinction between work as generally performed and incidental activities (pp.10f.), are likely to be very difficult for respondents to understand and use on their own to the extent suggested by the Manual's illustrations used to guide field economists.

b. Data quality measures
There are very few numerical measures of reliability, validity, and accuracy in the ORS project documents available for this report. The Feasibility Report mentions various kinds of reviews of schedules (p.12f.) and an efficiency test (pp.14, 27f.), but not the results of these exercises. A mode of collection test found what appear to be high rates of coding disagreement between personal visits and remote collection by phone an email for several data elements, but do not mention whether the disagreements were large or small (p.14). A discussion of the accuracy of occupational coding is not entirely clear, though it indicates the greater difficulty of coding occupations at the fine levels of detail desired by many stakeholders (p.21). Potentially

significant, field economists found some respondents gave upwardly biased responses for some elements, including complexity, and that their training helped them to guide respondents appropriately and interpret their responses, but there is no description of the problem or its magnitude beyond this statement (p.42).

Tables of intra-class correlations show very high rates of agreement among 45 field economists scoring job tasks for four occupations after watching videos, with 3 of the 4 single-rater ICCs above 0.80 and all four ICCs for average ratings above 0.99. These ICCs were almost unchanged after participants viewed the videos again. However, it is not clear from the report whether the initial ratings preceded or followed a group discussion of the elements present, which would boost the agreement levels, nor are the specific elements rated described, so it is difficult to draw firm conclusions from this information (p.20). While encouraging, these results should be considered the beginning of more systematic testing.

Other tables indicated high rates of accuracy for SVP and required education, but often low rates for (1) prior work experience, (2) training, license, and certification, and (3) post-employment training (p.23).

c. Quality assurance

A short section of the summary report discusses quality assurance practices. Data from the DOT and O*NET are used to identify outliers in initial ORS data and other unexpected values based on relationships among variables in the other databases. Results are also compared to information in the NCS (p.12f.). It would be useful to know the percentage of initial values submitted by field economists based on interviews and job observations that are altered as a result of post-collection processing quality checks, and the magnitude of the changes (i.e., difference between initial and final values). Clearly, a data collection process that generates fewer anomalies and needs fewer corrections is more desirable. Fewer errors in the beginning mean less time and resources needed for quality assurance, fewer errors that might slip through the process, and greater confidence in the ORS primary data collection design. It may be noted that effective training for field economists and ongoing communication through group debriefings are essential for minimizing rater variance, but there are many other sources of ratings variance, as well, as noted in earlier sections of this report.

While ORS personnel undoubtedly recognize these points, it is not clear that they have been the subject of systematic study. The brief section on Data Review Procedures in the Feasibility Report contains no figures on how a sample of ratings changed at each step of the quality review process, so there is no way for this report to comment further on overall quality of current practices or the specific track records of different collection methods that may have been used (e.g., telephone interviews, on-site interviews, job observations). As noted previously, the absence of such information is a significant gap in the current ORS program design. **It is recommended that ORS initiates a formal and systematic study program to understand as deeply as possible the relative merits and data quality implications of different sources of information, modes of data collection, and instrument design.**

E. OTHER LITERATURE

1. IO psychology

SSA and BLS seek measures that apply to the entire workforce. The measures must be reliable and valid for all jobs. Most studies in IO psychology do not have this national, economy-wide scope. Studies using task inventories that must be customized for different occupations and workplaces may have desired concreteness but lack generality across occupations. Approaches that do seek to measure all jobs on common scales often resort to items that lack behavioral concreteness. Their track record may not apply to instruments like the ORS that attempt a different balance between concreteness and general applicability. Approaches that do aim for both concreteness and generality often validate their instruments on a convenience sample of occupations. If a reliability measure depends partly on observed total variation, then a representative sample of occupations, with greater total variation, may have different measured reliability. Likewise, regardless of the quality of the instrument, work measures tested on very narrowly-defined occupations (e.g., court reporters and security guards for a county government study) may have very different properties than the same measures used in a large-scale study like ORS due to the need to use broader occupational titles (e.g., *stenographers, all* and *security guards, all*).

SSA and BLS seek to gather measures that are behavioral and concrete, but potentially applicable to all jobs in the economy. One stated goal is to combine single indicators of behaviors, including quite precise duration measures, in pre-determined ways to derive measures of constructs such as "sedentary" and "heavy" work. Many existing approaches, though not all, use measures that are more general, lack duration precision, and use factor analysis or other empirical methods to combine indicators into scales before measuring the scales' reliability and validity. Conclusions applying to the latter do not necessarily generalize to the former.

In addition to the unsystematic sampling of occupations and establishments/organizations, IO psychology studies often test different instruments, so it is difficult to know how much of the variation in outcomes is due to the samples or the instruments. The instruments that have been subject to multiple testing (e.g., PAQ) are so different from NCS approach that it's not clear the results generalize to this project. Many of the studies also use job-specific items that are appropriate for specific work contexts or research purposes but not for the ORS. Task inventories can include 200 items on checklists specific to very narrowly defined jobs. By contrast, items on general purpose instruments will almost invariably have greater ambiguity or spottier coverage of different contingencies (e.g., difficulty in defining workload or mental demands consistently across occupations like teacher and pilot). A recent review captured this dilemma:

> *Industrial and organizational psychologists have long aspired to a "common metric" in the language of work through which work requirements could be compared across jobs. This aspiration led to the development of the DOT. In fact, one of the motivations behind the DOT's replacement, namely O*NET, was the DOT's reliance on occupation-specific tasks that interfered with cross-occupational comparisons. A review of current usages of O*NET, however, suggested that many of the psychologically worded items employed in*

> *O*NET* [are not helpful to users] (Sanchez and Levine 2012, p.414; references omitted, MJH).

The IO literature is itself cognizant of the difficulties with establishing standards for measurement quality, as well as determining whether they are met. As Morgeson and Campion observe, "The entire job analysis domain has struggled with questions about what constitutes accuracy. That is, how do we know that job analysis information is accurate or 'true?' The absence of a definitive answer to this question creates problems for both researchers and practitioners" (Morgeson and Campion 2000, p.819). They suggest a six-dimensional concept of "accuracy" corresponding roughly to two concepts of reliability (consistency, agreement) and four facets of validity (content, construct, divergent, and criterion validity relative to expert judgment) (p.821). However, they recognize that when different job ratings are averaged to reduce random error the aggregated score partly masks true differences because no two jobs are ever truly absolutely identical, "jobs [*and occupations—MJH*] are partly a social construction…jobs are really collections of demands with imprecise boundaries, making it difficult to definitively identify where one job stops and another starts" (p.822, 824).

> *If accuracy is viewed as convergence to a known standard, then speaking in terms of the `accuracy' of job analysis data is inappropriate, in part because there are rarely unambiguous standards against which to judge these data.… This points to a seemingly intractable problem: if one focuses on the accuracy of job analysis data, it is relatively easy to demonstrate how these data might not be stable or objective in an absolute sense. Because of this, we may never be certain in our knowledge about the data's accuracy* (Morgeson and Campion 2000, p.822).

Harvey and Wilson (2000) defend vigorously the possibility of objective measurement by focusing on observable behaviors and concrete tasks, which corresponds closely to the approach endorsed by this report and used in most ORS measures. However, it should be noted that this does not deal with the absence of absolute standards for delineating occupational boundaries, the problem of occupational coding errors that is exacerbated at finer levels of occupational detail, and the fact that complex constructs and mental activities, such as the ORS cognitive job requirements elements or SVP, are often not discrete, directly observable tasks.

Even so, the level of knowledge regarding measurement quality remains inadequate. In a major review of IO psychology research, Harvey found inter-rater reliabilities for relatively concrete items often range between 0.7 and above 0.9. However, some of these correlations are inflated by inclusion of cases in which a rated attribute does not apply (DNA) to a job: "the numerous DNA ratings provide a built-in level of 'reliability' that can mask serious disagreements on the items that *are* part of a job" (Harvey 1991, p.112, emph, orig.).

In a more recent review, Dierdoff and Wilson (2003) note that prior research has not included benchmarks of typical reliabilities. "To date, no such estimates have been available, and practitioners have had no means of comparison with which to associate the reliability levels they may have obtained" (p.635). The analysis divided studies according to kind of measure (task vs. descriptions of more *general work activities* or GWA), information source (job incumbents, supervisors and managers, job analysts), and scale type (frequency, time spent, difficulty, importance, PAQ). The mean for 119 inter-rater reliabilities from task-level job analyses was

0.77, compared to 0.61 for 95 reliabilities from GWA job analyses (p.638). The interpretability of these averages is complicated by the fact that the scale lengths and number of raters differed and the authors make statistical adjustments to account for these differences. They support the view that more concrete, task-based items have greater reliability but the value of the more detailed findings is limited by the difficulty of determining comparability between the original studies and the ORS. This is another illustration of the need for the ORS to conduct its own studies tailored carefully to address the particular issues identified in this report, as well as others identified internally by ORS.

The body of evidence in Sanchez and Levine's more recent review also indicates the greater reliability of task-based items relative to more general or holistic rating tasks, but some results continue to differ (2012, p.402). Not surprisingly, reliabilities for required personality traits are much lower than for tasks (Sanchez and Levine 2012, p.409). Agreement between incumbents and analysts tends to be higher for jobs that are less cognitively complex and within-occupation variation among incumbents is lower in jobs involving less complexity and autonomy, but there is otherwise little information on systematic sources of disagreement within job titles. Research seems to show that carelessness, insofar as it can be detected, is associated with elevated ratings from incumbents. Therefore, one concrete recommendation from these studies is to reduce respondent burden to avoid careless responding (Sanchez and Levine 2012, p.404). Sanchez and Levine offer a sobering view of the state of knowledge regarding the sources of rating variation among incumbents:

> *…continued "fishing" for differences observed among incumbents as a function of demographic breakdowns of dubious theoretical value (e.g., incumbents' race or sex) will simply replicate what we already know, namely that statistically significant differences among such groupings of incumbents are erratic, their effect sizes small, and their practical significance questionable… instead of trying to hide or eliminate disagreement, job analysis research should embrace it by looking more deeply into its causes. Legitimate disagreement represents unique ways in which incumbents experience their job, and a better understanding of their ideographic representations might increase our grasp on the various forms in which jobs can be crafted along with their requirements and consequences* (Sanchez and Levine 2012, p.407).

The picture that emerges from this review is one of fragmentation and study-specific conclusions rather than a body of consistent results that provides clear guidance on generally validated best practices that transfer easily to a program like ORS. The initial literature review for this report performed internally at BLS, reproduced in the Annex, reinforces this impression.[37] As that paper observes, most job analysis research involves incumbent-completed questionnaires and task lists, rather than interviews with human resources professionals, indicating, again, the distance from the current ORS design.

[37] "Research on Approaches and Methods for Obtaining ORS data – Annotated Bibliography and List of Job Analysis Tools/Methods" BLS (nd) in this report's Annex.

2. Health and, occupational health, ergonomics, and related fields

Given the ORS emphasis on measuring physical demands of occupations, it would seem useful to review results from fields in which physical activity, particularly in the workplace, is the central focus.

Based on a broad search of databases, Castillo-Retamal and Hinckson (2011) reviewed 11 studies published between 1990 and 2009 that studied physical activity and sedentary behavior at work with objective measures (motion sensors: accelerometers, pedometers), subjective measures (self-reports: interviews, surveys and questionnaires) and "standard criteria" (indirect calorimetry). The review reported validities for the objective measures between 0.82-0.92 for the accelerometer and 0.96 for the pedometer and 0.86-0.91 for a widely used survey, the Tecumseh Occupational Physical Activity Questionnaire. Test-retest reliability coefficients for both the Tecumseh survey and the International Physical Activity Questionnaire were 0.80 (Castillo-Retamal and Hinckson 2011, p.352f.).

The three studies using objective measures studied workers over 3-7 days. One study (n=181) found the number of steps taken during work hours ranged from 4,422 (sd=1,380) for university academics to 10,344 (sd=5,553) for blue-collar workers (mechanics, green keepers, dry cleaners) (p.347), a ratio of over 2.3 for these occupations used as contrasting groups.

Another study reviewed that used a pedometer (n=47) found cleaners spent 23% of their time sitting, while researchers and administrators spent more than 60% of their time sitting (p.347), a ratio of 2.65. These values illustrate the magnitudes that might be expected in ORS data.

Not surprisingly, the five studies in the review that used self-report survey responses had much larger sample sizes than those using physical instruments to monitor activity (n=90, 508, 526, 1579, and 6360). One drew a random sample of 1,579 full-time workers in two communities in Queensland, Australia (Mummery et al. 2005). Respondents were classified into three very broad categories: (1) managers, administrators, professionals, associate professionals; (2) clerical, sales and service workers, and (3) blue-collar workers. The sample divided roughly evenly between these groups (see table below). Respondents were asked the number of hours and minutes spent sitting during a normal working day and classified into quartiles: (1) 0 to 44 minutes; (2) 45 to 149 minutes; (3) 150 to 359 minutes; and (4) $\geq$360 minutes. Thus, 50% of the sample sat no more than 2.5 hours per day. The means (in minutes) are in the expected order, though there is no way to check their accuracy in an absolute sense from the research design, and the standard deviations are large. Interestingly, the figures for professionals and blue-collar workers represent 55% and 30% of a 7.5 hour day, respectively, not far off from the 60% and 23% figures recorded for the study using objective measures, though the professional/blue-collar ratio is closer to 1.8.

Variable	*n*	Sitting time mean (SD)
Total sample	1579	199.9 (177.8)
Professional	554	248.8 (175.1)
White-collar worker	409	207.1 (169.1)
Blue-collar worker	429	136.1 (164.0)

Some of the same authors studied a smaller, similar group (n=90) of broad occupations with both survey and pedometer measures. Responses to the Tecumseh Occupational Physical Activity Questionnaire can be converted into estimates of energy expended after scoring.
Those in the high and low white-collar groups spent over 80% of their work time in light activities (<3 METS), while blue-collar workers spent 21% of their time in light activities, 58% in moderate activities (3-6 METS), and 21% of their time in heavy activity (>6 METS) (Steele and Mummery 2003, p.402). One of the most interesting findings was a correlation between total occupational physical activity reported on the survey and average step-counts over three days (r=0.62), as well as correlations between step-counts and survey-based estimates of light work activity (r = -0.62), moderate work activity (r=0.69), and heavy work activity (r=0.38) (Steele and Mummery 2003, p.403).

A large, nationally representative Dutch survey asked workers how many minutes they spent sitting at work the previous day (Jans, Proper, Hildebrandt 2007). The results are broadly consistent with the previous table and the pedometer study of the proportion of the work day spent in sedentary activity.

Minutes sitting at work previous day

Occupational group	Minutes	Hours	Pct. of day
Legislators and senior managers (n=654)	181	3.02	40%
Clerks (n=1728)	160	2.67	36%
Scientific and artistic professions (n=2143)	128	2.13	28%
Commercial workers (n=635)	96	1.60	21%
Trade, industrial, transport (n=1370)	83	1.38	18%
Service workers (n=796)	51	0.85	11%
Agricultural occupations (n=228)	74	1.23	16%
Total (N=7,720)	117	1.95	26%

Note: Final column defines a work day as 7.5 hours.

The results suggest feasible methods for large-sample incumbent surveys of physical activity on the job and workplace environmental conditions. A detailed list of sedentary behavior questionnaires, including reliability and validity coefficients can be found at: http://www.sedentarybehaviour.org/sedentary-behaviour-questionnaires/ (accessed 2/12/2015).

F. OCCUPATIONAL CODING ISSUES

Although the question of measurement quality naturally directs most attention to focus on the quality of the job ratings, the use of ratings in the form of occupational averages means that the reliability and validity of the ratings also depends on the reliability and validity of the occupational categories to which they are matched and within which they are aggregated. ORS plans to use 8-digit SOC codes. This is finer than most data collection programs but a number of stakeholders argue that more detailed codes should be used because they will capture "true" occupational requirements better. The spirit of many outside comments suggests a common belief that (1) "true" occupations can be identified easily, (2) individual jobs can be classified into higher-level occupations without error, and that having done so one can obtain (3) a database that easily distinguishes random rating error within occupations from true heterogeneity across occupations, which permits valid occupational averaging to purge measurement error from the former. However, as discussed previously, there are no fixed and definitive standards for deciding how similar and in what ways two jobs must be if they are to be classified as belonging to a single occupation.

In addition, although heterogeneity within SOC occupations is potentially problematic, finer coding generally reveals problems in assuming occupational coding is error-free. In fact, it is known that occupational coding, like job ratings themselves, involves error (see Bound Brown, Mathiowietz 2001, pp.3802ff., Abraham and Spletzer 2010, and references therein). Indeed, the rate of classification error is an increasing function of the fineness of occupational detail because errors tend to involve misclassifications among neighboring occupations, which cease to be errors when broader occupational categories that encompass them are used. SSA claims examiners have great difficulty acquiring sufficiently detailed occupational information from applicants and assigning DOT codes (Social Security Administration 2011, pp.302ff.). While many perceive finer codes will permit within-occupation heterogeneity to be distinguished from measurement error, some of the anticipated benefits will be offset by increased classification error because the task of discriminating group membership becomes more difficult. Finer occupations reduce definitional heterogeneity, but increase coding error. Lower coding reliability means that the heterogeneity whose removal is sought through purer occupational definitions reenters through elevated error rates in the assignment of jobs to the narrower occupations. Defining occupations more narrowly than 8-digit SOC is no guarantee that heterogeneity will decline greatly in practice given the increased difficulty of the finer coding that the proposed solution itself entails. While there may be net gains, it is clear that finer occupational coding will not control within-occupation heterogeneity or measurement error as easily as these suggestions suppose.

One possibility for addressing continuing concerns with the level of occupational detail is to make the ORS job-level database publicly available, in addition to the SOC-level database. Alternative database views by trait or microdata could include the original job titles that were used to assign SOC codes, perhaps standardized somewhat by computer-matching the job titles obtained respondents to entries in the Census Bureau's *Alphabetical Index of Occupations.*

On a practical level, it is clear that attention must be paid to ensuring reliable classification of jobs into occupations as well as reliable and valid job ratings. Misclassification of jobs into occupations defined at any level of detail will produce upwardly biased estimates of within-

occupation variation even if the individual jobs were rated with perfect reliability and validity. Stated another way, estimates of rating agreement will be biased downwards if based on comparisons across jobs that belong to different occupations but are mistakenly treated as belonging to the same occupation. In short, unreliability of occupational coding will place limits on the measured reliability of job ratings, and this needs to be built into ORS planning and validation efforts.

O*NET's method of working through employers likely reduces occupational coding error relative to household surveys, as does the OES and ORS, though it is possible to increase the number of questions regarding occupational identification on other surveys to address this issue. Indeed, **the ORS should consider enriching the items used to identify and code occupations as a means of enhancing the reliability of job ratings.** As an aside, one might add that similar attention to enhancing the standard two-item question sequence for occupation in the Current Population Survey and Census would be salutary, as well. It is quite possible that small additions could produce significant improvements in the reliability of occupational coding.

Another issue, relating more to validity than reliability, is the treatment of residual occupations (*not elsewhere classified,* or NEC). No one really believes the SOC codes represent "real" occupations in a substantive sense and O*NET refuses to populate them with data. In fact, they are most likely used because a catch-all category is needed for the many very small occupations that are insufficiently populated to warrant individual recognition, as well as insufficiently informative answers from respondents in recognized occupations. Some NEC occupations are rather sizeable and taken together the NEC occupations may account for a significant proportion of jobs. It is not known if they contain greater heterogeneity in job characteristics than occupations with more substantive titles, so the validity of the aggregate category is unclear. However, if SSA wants to know how many *jobs* have different characteristics, it would seem that jobs classified in NEC occupations, for which the validity of occupational averages is unknown, presents a problem. Whether or not jobs belonging to NEC occupations will be excluded from the ORS database or how they will be included is unclear from the information available. If they are included, the dilemmas surrounding the treatment of intra-occupation ratings variation will be even more acute than the situations discussed previously given the acknowledged artificiality of the occupational categories.

Somewhat related is the more general issue of occupational weights. Rating jobs or occupations requires a sampling scheme that produces a sufficient number of cases per occupation, which is an especially important consideration given their anticipated use in deriving averages. However, these sampling considerations are different from ensuring national representativeness. Weights will be needed to derive estimates of the prevalence of jobs with different demands. Whether the NCS sampling scheme or another source will be used to derive population weights is unclear. It is also possible that exact population weights are not required by controlling interpretations of the "*significant numbers*" criterion in the Social Security Act used in Step 5.

G. UNTAPPED SOURCES OF OCCUPATIONAL INFORMATION

ORS might want to consider using a range of data sources in addition to those mentioned previously in constructing and validating ORS measures.

The Current Population Survey's January 1991 supplement contains a unique set of self-reported information on cognitive tasks performed at work (Handel 2000, 2007). The large sample size and Concreteness of the task measures make this a useful source for measuring the convergent validity of ORS cognitive measures and the anticipated level of intra-occupation variation in nationally representative samples. **The CPS January 1991 also shows the viability of a worker survey of job characteristics, which may be an important consideration for a number of ORS data elements, particularly environmental conditions due to their potential sensitivity in an employer survey.**

BLS' Occupational Safety and Health Statistics and administrative data (e.g., OSHA 300 logs) may be a useful source of both objective data and validation information on physical job demands and environmental hazards, as well as methodological insight into eliciting information from the public on these topics. Prevalence by occupation and industry can be compared to ORS results. Given the advisability of using sources other than employers, this database may help meet a key need.

Among firm officials, ORS should consider using safety officers, ergonomics professionals, and industrial engineers responsible for time and motion studies. There is a large literature in industrial engineering on the measurement of physical job demands, particularly a field known as *Methods-Time Measurement*, which is conducts detailed measurement of tasks broken down into small units of effort and motion. It is not known whether they have reference works that would be useful for ORS.

Researchers in the occupational health field have estimated the energy requirements of different tasks, and levels and prevalence of working conditions (e.g., noise levels) for occupations. It is not clear whether there are useful reference works with systematic occupational coverage or simply a corpus of individual research articles using diverse methods (see, e.g., Ainsworth et al. 2011). If ORS has not already consulted with experts at NIOSH, NIH, and CDC, the agencies should be contacted for any resources they may have on exertional levels by occupation and the prevalence of relevant injuries and environmental conditions that may be used as primary data or validation data, or some combination.

It is likely that there is an enormous amount of occupational information in job descriptions maintained by OPM. Whether this information includes data elements directly usable for the ORS and accessible in standardized electronic form is unknown. Nevertheless, given the care that is used in writing and validating Federal job descriptions, the frequently large numbers employed, and the similarity of many jobs across public and private sectors, ORS should investigate the possibility of using information from OPM and other sources of administrative data, such as standards from industry sector skills councils, as primary or validation data. Although Big Data techniques are still evolving, ORS might also consider the utility of

harvesting information from online sources, such as Monster.com and other massive job posting web sites.[38]

Finally, given that Steps 4 and 5 often involve skill transferability analyses, it may be helpful to perform empirical job flow analyses using CPS, ACS, or other large databases with a short panel structure. Occupational similarity is suggested to the extent that workers reported leaving one and entering another occupation over a two-year time frame. Such analyses would exclude individuals reporting increased education in the second year to control for changes in worker skills. The effects of experience-related skill acquisition could be controlled by considering occupational pairs similar to the extent that the flows are approximately symmetrical controlling for wages, occupational size, and other relevant characteristics. Although the specific characteristics that make the occupations open to similar workers cannot be isolated in great detail using this method, it provides a useful empirical check on the identification of similar occupations using the ORS' job trait information. Within these and similar limits, an inventory of occupational cross-listings based on observed job flows would be a useful reference aid for SSA's skills transferability analyses and vocational experts serving as witnesses, as well as another validation tool for the ORS ratings.

IV. IMPLICATIONS AND RECOMMENDATIONS

The ORS is a large and important project that will be the focus of unusual scrutiny given its direct role in the administration of a large public benefits program that is itself the subject of intense interest and political contention. While existing relevant methodological research provides some useful information, the results differ from the information the ORS requires along too many dimensions (measures and scales, target population and samples, collection methods, broader goals) to provide clear guidance and address stakeholder concerns effectively. *This report recommends the ORS conduct its own experiments and validation studies in order to ensure that the final design adopted aims for the most reliable, valid, and accurate measures that can be expected reasonably given the feasible alternatives.* For reasons discussed previously, the studies should address concerns of stakeholders in addition to those identified by the agencies involved. Clearly, perfection is not a realistic standard for any database but careful program design is a reasonable expectation.

The issues that experiments and validation studies need to address are multiple and complex. Ambiguity regarding SSA's exact needs and priorities and BLS' choice set remain problems for even the construction of a well-developed plan of action. Variations in program design and variable definition can have significant implications for the design of methodological studies, analytical methods employed, benchmark values against which to assess results, and recommendations for the final program design, such as where to target resources to maximize quality. Some of this reflects complexity of the Grids, ambiguity surrounding actual operation of Steps 4 and 5 in practice, and the unequal but imperfectly known distribution of SSA cases across occupations and relevant job characteristics, such as physical demands and skill levels.

[38] See, e.g., Marc Anderberg (2012), "Job Content and Skill Requirements in an Era of Accelerated Diffusion of Innovation: Modeling Skills at the Detailed Work Activity Level for Operational Decision Support," Employment and Training Administration, U.S. Department of Labor, esp. pp56ff.

While it is conceivable that every data element is equally critical and there are no other opportunities to improve measurement quality by applying resources in a non-uniform manner, this is not clear from the materials reviewed.

Given these and other considerations detailed below, the first recommendation is that the ORS develop a ***coherent master plan for conducting methodological studies***.

- **Recommendation:** The ORS needs to develop a strategy document to guide methodological research. The plan should give a clear, compact summary and explanation of specific data needs and intended uses, and feasible collection procedures. Specifically, the document should include
    1. a list of all variables, their levels of measurement (dichotomous, polytomous, or continuous), and response options (categorical) or units (continuous)
    2. a section identifying how composite variables (e.g., sedentary/non-sedentary) are to be constructed from multiple individual variables
    3. a section identifying items and response options that have greatest importance/relevance and those with limited applicability or lower priority based on either à priori grounds or based on the relative frequency/rarity with which situations present themselves to SSA. This may permit targeting of resources to where they have the greatest impact.
        - This has particular relevance for extreme ends of scales for variables that are difficult to measure.
        - Example: Does distinguishing durations that are "Constant" from those that are "Frequent" have great practical consequences or do almost all cases hinge on whether or not duration is "Frequent or more?"
        - Same applies to distinguishing "Heavy" and "Very heavy" work if almost all cases hinge on whether or not work is "at least Heavy." Likewise, are nine SVP levels necessary given what appears to be the coarser Grid skill levels, etc.?
    4. a section identifying occupations that are particularly relevant because they represent the most common previous work of SSA applicants or those most commonly cited as ones into which they could transfer. This will also help target resources. Specifically,
        - occupations with the greatest need for validation could be the subject of the greatest validation and data collection efforts (e.g., personal visits by field economists); occupations that are infrequently seen or whose characteristics are known to involve little physical activity (e.g., white-collar office jobs) need not be considered for physical observations or other intensive data collection methods
        - occupations with the greatest need for validation could also be sub-divided into finer codes to meet concerns regarding within-occupation heterogeneity (true variance), i.e., a limited expansion of the level of occupational detail may be possible
        - Rationale: If 50 job titles or occupations account for nearly 50% of cases then directing scarce resources and attention to study them more intensively and define them more narrowly addresses the largest portion of the concerns expressed by stakeholders

5. a clear statement of the different data collection methods under serious consideration and rankings of their desirability and feasibility, including
    - kinds of interview subjects
    - kinds of interview methods
    - a rough estimate of the number of personal visits and job observations that are feasible
6. a description of the format and content of final data products
    - clarification of relative importance of detailed occupational profiles scoring all jobs in the economy on all variables vs. focusing on estimates of the number of low-skill sedentary jobs

This planning document is necessary both as a roadmap for the work that a validation phase must accomplish and because detailed study of reliability, validity, and accuracy requires prior specification of intended procedures and purposes, and required levels of precision. More detailed recommendations await completion of this document but a general framework and some specific studies can be described given existing information.

In an ideal but imaginary world one would have a database that was the product of omniscience against which one could compare the results of the various design options under consideration. The ways in which the real world departs from this ideal point to the challenges and issues methodological studies must address, which can be given a rough logical order.

1. Because omniscience is impossible, the studies will need to decide whether it is possible to order available methods such that one method can credibly be considered a benchmark or criterion for accuracy against which the performance of the other methods can be evaluated (i.e., gold standard). A reasonable choice is to consider field economists' personal observations of work involving physical demands and relevant environmental conditions as the gold standard for those data elements. Decisions regarding the number of personal visits and observations that are feasible in methodological testing and as a permanent part of data collection rest with BLS. Physical measuring devices for objective data collection may or may not be considered feasible for inclusion in the testing phase. By contrast, SVP and mental requirements, such as educational levels and cognitive elements, and perhaps other data elements, cannot be observed, and the existence of alternative gold standards is an open question.

2. If gold standards do not exist or cannot be implemented within existing resource constraints, then the strategy is to compare agreement among largely unorderable alternative methods of data collection as a test of convergent validity. If there is high agreement within required levels of precision among different sources of information (HR officials, supervisors, incumbents), modes of collection (fixed interviews, conversational interviews, in-person vs. remote interviewing, surveys, web surveys), and modes of post-collection processing (regular, intensive reconciliation among analysts vs. statistical outlier detection), then the fact that different methods produce very similar values gives one confidence they are equally well-focused on the same target. The ORS will have to decide which combinations of sources, modes, and processing will define the

"methods" that are to be compared because it is likely infeasible to compare all possible combinations of these elements.

3. If the different methods exhibit significant and systematic disagreement then the question becomes which is to be preferred in the absence of a gold standard. It may be possible to use the method of contrasting groups to rank them according to how well they discriminate among groups that are expected to differ based on prior information, such as occupational ratings in the DOT, NCS, and O*NET. Likewise, patterns of association between different variables within methods can be examined for their relative levels of consistency with expectation (convergent and divergent validity). It may also be possible to rank them according to their relative reliability, as well.

4. Because all measures contain error and most ORS measures are criterion-referenced, reliability should be examined using agreement-based measures for continuous or categorical variables, as appropriate. This involves assessing agreement within each method in Step 2, rather than agreement across them as in that step. Sources of variance to be considered can include

    a. difficulty of the survey item (e.g., task frequency and duration)
    b. characteristics of the mode of data collection (e.g., interview length, in-person vs. phone interview, use of job observations)
    c. characteristics of the respondent (e.g., position, tenure, level of familiarity with target job, recall ability, quality of judgment, interest level in the interview)
    d. characteristics of the job rated (e.g., major occupation, average rated trait level, task homogeneity, industry, incumbent demographics such as education, age, gender, tenure)
    e. characteristics of written materials provided or obtained (e.g., idealized hiring requirements aka "purple squirrel")[39]
    f. characteristics of the employing establishment and firm (e.g., urban/rural location, organizational size, management cooperation level)
    g. identity of the field economist (e.g., experience, training, judgment, idiosyncratic variation in data collection practices and rating styles)
    h. identity of the field office (e.g., differences in training and informal practices)
    i. occasion (i.e., purely random rate-rerate variation)

    The ability to estimate multiple sources of measurement error simultaneously may be particularly useful given the possibilities listed above. Generalizability theory, associated with Cronbach and others, may be worth exploring for guidance in this area. As no set of experiments can investigate all of these sources of variance, prior project reports should be mined and field economists and supervisors debriefed to identify those that are most likely to be the largest contributors. Clearly, rater-related variance is the traditional focus of reliability studies and should be a principal concern of ORS reliability studies. *In addition, given the known difficulty of the duration items and the social distance*

[39] Lance Haun, "Don't Hire the Perfect Candidate," Harvard Business Review, January 14, 2013.

separating many firm officials in large organizations from physically demanding front-line jobs, special attention should be given to these items and situations.

5. One source of variation that involves both questions of validity and reliability concerns the occupational classification system. It is reasonable to assume that some variation within SOC codes reflects random measurement error, which can be mitigated substantially by averaging, but it is also likely that some within-occupation variation reflects heterogeneity. The absolute and relative magnitudes of occupational heterogeneity are unknown, as are the practical consequences of its concurrent removal when averaging. Addressing these issues will require further consideration, particularly when gold standards are absent; interim suggestions are included below.

6. The final issue is to evaluate observed estimates of accuracy, validity, and reliability relative to their consequences, if possible. For the numerous categorical variables that the ORS will produce, the key question is, ***What percentage of jobs is misclassified and how seriously when using different methods and judging against a gold standard or preferred method, and what percentage are variably classified and by how much when methods cannot be ranked?*** This reflects the idea that the standard for judging the magnitude of departures from perfect accuracy, validity, and reliability should be substantive importance. Because some variables are derived from combinations of others, the validity and reliability of the compounds as well as the components need to be measured. It is likely that there will be relatively few cases of extreme disagreement; how many cases close to a threshold are variably classified is less predictable. If variations in the final scores of interest are not large enough to have much impact on the characterizations of jobs or the decisions made using them then they are well within acceptable standards for the purposes for which the ORS is intended.

Implications of the preceding can be summarized in the following recommendation.

- **Recommendation:** The ORS needs to design and conduct a well-planned series of experiments and validation exercises to evaluate the efficacy of different approaches to data collection and optimize the final design of the program. This involves
    1. identifying all gold standards that might serve as criteria validation for accuracy
    2. defining a reasonable range of other methods to assess convergent validity when gold standards are unavailable
    3. identifying likely significant sources of error variance for assessing reliability, including duration items and respondents relatively distant from the front-line job
    4. considering methods for distinguishing error variance and heterogeneity within critical occupations and measuring their absolute and relative magnitudes
    5. relating standard measures of validity and reliability to rates of classification disagreement and the magnitude of disagreement to assess their practical implications

Some of this research may be possible using ORS data collected previously, while the greater part will almost certainly require careful design of new studies to address these particular concerns in a systematic manner. The ORS should contract with an experienced IO psychologist

to help with the design of these studies, after making significant progress on the master plan that will guide them and conducting internal discussions to gain at least some clarity on the specific issues summarized immediately above. The IO psychologist's level of experience with job analysis and experimental design, and respect within the field will have a significant influence on the quality not only of the studies themselves but also the communication of results and reception among stakeholders. Given the unease that has sometimes arisen during SSA's search for a replacement to the DOT, transparency and external expertise in this phase of the project are important considerations. BLS validity studies should simulate as closely as possible the various feasible methods of data collection to determine their relative quality.

There are a number of large-sample studies that may be useful to consider using existing databases.

- Records in the April 1971 Current Population Survey (CPS) contains both 1970 Census occupational codes, which are a somewhat coarser version of SOC codes, and fourth edition DOT codes that were assigned based on the original narrative responses to the relevant CPS questions. This is the only source of population weights for DOT occupations. It is likely possible to use this data to calculate variances and ranges of DOT scores within Census codes, and to calculate rates of misclassification by comparing the incidence of job characteristics using the DOT scores assigned to CPS respondents to the incidence calculated when those scores are replaced by Census occupation-level means. Although the passage of time is a limitation, the magnitude of effects would provide some empirical evidence regarding a hitherto untested assumption that the level of occupational detail in the DOT had a large impact on performance relative to what is possible using a more aggregated occupational classification scheme.

- More current data using non-DOT measures can also be used to address this question and to get a preliminary sense of the level of within-occupation variation. The detailed NCS measures from the previous leveling scheme differ from the DOT in their level of concreteness, but the data collection platform is very close or identical to what the ORS will use, and the greater subjectivity of the ratings means any test is likely to be conservative relative to ORS items. It would be useful to calculate ICC, r_{wg}, and other applicable statistics to measure the magnitude of within-occupation variation within the NCS and to identify the sources of variation among those listed above (e.g., respondent, establishment, rater).

- The magnitude and patterns of variation within occupations can also be analyzed using O*NET microdata for the Work Context variables (and perhaps Education and Training). Unlike almost all other O*NET items these variables use relatively objective wording and objective response options (e.g., frequency), like the ORS, and include measures of physical demands and environmental conditions. Early versions of the Education and Training questionnaires collected information from incumbents that are very similar to ORS items and the occupation codes are nearly identical. Examining within-occupation score variation and comparing the incidence of different job characteristics using individual responses and occupational means would be very valuable for understanding the patterns one might anticipate for the full-scale ORS.

- Although the different scaling of variables precludes using measures of absolute agreement, some indication of convergent validity between ratings derived from ORS-type respondents and job incumbents can be gained by occupation-level correlations between NCS and O*NET variables dealing with physical demands, environmental conditions, skill requirements, autonomy, and other parallel concepts.

The level of similarity between the other databases and the ORS in terms of item type, survey procedure, and the age of measures is summarized in the following table, where the number of symbols (**+**) indicates level of similarity:

	Item type	Procedure	Vintage
DOT	+ + +		
NCS		+ + +	+ + +
O*NET	+ +		+ + +

- **Recommendation:** The ORS should consider mining existing large-sample databases for the insights they can provide regarding the validity and reliability of measures and procedures similar to the ORS, and the likely magnitude and practical significance of within-occupation variation in job ratings.

Stakeholder concerns regarding job observations and level of occupational detail also need to be addressed.

- **Recommendation:** ORS needs to perform rough costing to understand the potential for using job observations for key jobs given foreseeable resource constraints.
- **Recommendation:** SSA needs to provide ORS with a reasoned list of critical jobs and data elements that reflect the most frequent situations it faces so ORS can conduct more intensive study of those cases in the form of observational data collection, finer occupational detail, and additional probes and other forms of evidence. SSA must identify priority needs in order for stakeholder concerns to be addressed within the limits of feasibility.

REFERENCES

Abraham, Katharine G. and James R. Spletzer. 2010. "Are the New Jobs Good Jobs?" Pp.101-143 in Labor in the New Economy, Katharine G. Abraham, James R. Spletzer, and Michael Harper, eds. Chicago: University of Chicago Press.

Agresti, Alan. 2012. Categorical Data Analysis. Hoboken, NJ: Wiley.

Alonso, William and Paul Starr. 1989. The Politics of Numbers. New York: Russell Sage.

Anderson, Margo J. and Stephen E. Fienberg. 2001. Who Counts? The Politics of Census-Taking in Contemporary America. New York: Russell Sage

Ainsworth, B.E., W.L. Haskell, S.D. Herrmann, N. Meckes, D.R. Bassett, Jr., C. Tudor-Locke, J.L. Greer, J. Vezina, M.C. Whitt-Glover, A.S. Leon. 2011. "2011 Compendium of Physical Activities: a second update of codes and MET values." Medicine & Science in Sports & Exercise. 43:1575-81.

Autor, David H. and Mark G. Duggan. 2006. "The Growth in the Social Security Disability Rolls: A Fiscal Crisis Unfolding." Journal of Economic Perspectives. 20:71-96.

Autor, David H. and Michael J. Handel. 2013. "Putting Tasks to the Test: Human Capital, Job Tasks, and Wages." Journal of Labor Economics. 31: S59-S96.

Bishop, John H. 1992. "Is a Skills Shortage Coming? A Review of BLS Occupational Projections to 2005." CAHRS Working Paper #92-04. Ithaca, NY: Cornell University, School of Industrial and Labor Relations, Center for Advanced Human Resource Studies.

Bishop, John H. 1997. "Is an Oversupply of College Graduates Coming?" National Center for Postsecondary Improvement, School of Education, Stanford University, Stanford, CA.

Bland, J. Martin and Douglas G. Altman. 1990. "A Note on the Use of the Intraclass Correlation Coefficient in the Evaluation of Agreement Between Two Methods of Measurement." Computers in Biology and Medicine. 20:337-40.

Bland, J. Martin and Douglas G. Altman. 2003. "Applying the Right Statistics: Analyses of Measurement Studies." Ultrasound in Obstetrics and Gynecology. 22:85-93.

Bland, J. Martin and Douglas G. Altman. 2012. "Agreed Statistics: Measurement Method Comparison." Anesthesiology. 116: 182-185.

Bound, John, Charles Brown, Nancy Mathiowietz. 2001. "Measurement Error in Survey Data," Pp.3705-3843 in Handbook of Econometrics (vol. 5), James J. Heckman and Edward Leamer, eds. Amsterdam: Elsevier.

Bowler, Mary and Teresa L. Morisi. 2006. "Understanding the Employment Measures from the CPS and CES Survey." Monthly Labor Review. February:23-38.

Bridgeman, Brent, Judith Pollack, and Nancy Burton. 2008. "Predicting Grades in Different Types of College Courses." College Board Research Report No. 2008-1, New York.

Buckley, John E. 2012. "The Relevance of Occupational Wage Leveling." Monthly Labor Review (April). http://www.bls.gov/opub/mlr/cwc/the-relevance-of-occupational-wage-leveling.pdf (accessed 12/13/14).

Cain, Pamela S. and Bert F. Green. 1983. "Reliabilities of selected ratings available from the Dictionary of Occupational Titles." Journal of Applied Psychology. 68:155-165.

Cain, Pamela S. and Donald J. Treiman. 1981. "The Dictionary of Occupational Titles as a Source of Occupational Data." American Sociological Review. 46:253-278.

Carnevale, Anthony P., Nicole Smith, and Jeff Strohl. 2010. "Help Wanted: Projections of Jobs and Education Requirements Through 2018, Technical summary." Center on Education and the Workforce, Georgetown University, Washington, DC.

Castillo-Retamal, Marcelo and Erica A. Hinckson. 2011. "Measuring physical activity and sedentary behaviour at work: A review." Work. 40: 345–357.

Cox, Nicholas J. 2004. "Speaking Stata: Graphing Agreement and Disagreement." Stata Journal. 4:329-349.

Cronbach, Lee J. 1984. Essentials of Psychological Testing, 4 ed. New York: Harper & Row.

Dierdoff, Erich C. and Mark A. Wilson. 2003. "A Meta-Analysis of Job Analysis Reliability." Journal of Applied Psychology. 88:635-646.

Dorans, Neil J. 1999. "Correspondences Between ACT™ and SAT® I Scores." New York: College Entrance Examination Board.

Ferguson, Renee. 2010. "Evaluation of 2008 Occupations Held by SSDI and SSI Disability Claimants." Presentation. Social Security Administration, Office of Program Development and Research.

Fleisher, Matthew S. and Suzanne Tsacoumis. 2012a. "O*NET® Analyst Occupational Skills Ratings: Procedures Update." HumRRO, Alexandria VA.

Fleisher, Matthew S. and Suzanne Tsacoumis. 2012b. "O*NET® Analyst Occupational Skills Ratings: Analysis Cycle 12 Results." HumRRO, Alexandria VA.

Greenlees, John S. and Robert B. McClelland. 2008. "Addressing Misconceptions about the Consumer Price Index." Monthly Labor Review. August:3-19.

Handel, Michael J. 2000. *Models of Economic Organization and the New Inequality in the United States*. PhD dissertation, Department of Sociology, Harvard University.

Handel, Michael J. 2007. "Computers and the Wage Structure." Research in Labor Economics. 26:155-196.

Handel, Michael J. 2012. "Trends in Job Skill Demands in OECD Countries." OECD Social, Employment and Migration Working Papers, No. 143. Paris: OECD.

Handel, Michael J. 2015. "Measuring Job Content: Skills, Technology, and Management Practices," in Oxford Handbook of Skills and Training, John Buchanan, David Finegold, Ken Mayhew and Chris Warhurst, eds. Oxford: Oxford University Press.

Handel, Michael J. (*forthcoming*) "O*NET: Strengths and Limitations." Job Tasks, Work Skills and the Labour Market, edited by Francis Green and Mark Keese, Organization for Economic Cooperation and Development.

Harvey, Robert J. 1991. "Job Analysis." Pp.71-163 in Handbook of Industrial and Organizational Psychology, Marvin D. Dunnette and Leaetta M. Hough, eds. Palo Alto, CA: Consulting Psychologists Press.

Harvey, Robert J. 2009. "The O*NET: Do too-abstract titles +unverifiable holistic ratings + questionable raters + low agreement +inadequate sampling + aggregation bias = (a) validity, (b) reliability, (c) utility, or (d) none of the above?" Paper provided to Panel to Review the Occupational Information Network (O*NET) (see Hilton and Tippins 2010 *infra*).

Harvey, Robert J. and Mark A. Wilson. 2000. "Yes Virginia, there *is* an objective reality in job analysis." Journal of Organizational Behavior. 21:829-854.

Hilton, Margaret L. and Nancy T. Tippins (Eds.). (2010). A Database for a Changing Economy: Review of the Occupational Information Network (O* NET). Washington, DC: National Academies Press.

Hubley, Nathaniel O. 2008. "The Untouchables: Why a Vocational Expert's Testimony in Social Security Disability Hearings Cannot be Touched." Valparaiso University Law Review. 43:353-406.

Hyde, Serah. 2014. "Is Disability Insurance Used as a Form of Extended Unemployment Insurance?" Monthly Labor Review (November).

Jans, Marielle P., Karin I. Proper, and Vincent H. Hildebrandt 2007. "Sedentary Behavior in Dutch Workers: Differences Between Occupations and Business Sectors." American Journal of Preventive Medicine. 33:450-454.

Johnson, David S. and Timothy M. Smeeding. 2012. "A Consumer's Guide to Interpreting Various U.S. Poverty Measures." *Fast Focus* No. 14–2012. Institute for Research on Poverty. University of Wisconsin—Madison.

Kane, Michael. 1996. "The Precision of Measurements." Applied Measurement in Education. 9:355-379.

Karman, Sylvia. 2009. "SSA's Challenge: The DOT." PowerPoint presentation to the Inaugural Meeting, Occupational Information Development Advisory Panel (February 23).

Kaye, H. Stephen. 2010. "The Impact of the 2007–09 Recession on Workers with Disabilities." Monthly Labor Review (October).

Kottner, Jan, Laurent Audigé, Stig Brorson, Allan Donner, Byron J. Gajewski, Asbjørn Hróbjartsson, Chris Roberts, Mohamed Shoukri, David L. Streiner. 2011. "Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed." Journal of Clinical Epidemiology. 64:96-106.

Kohn, Melvin L. and Carmi Schooler. 1973. "Occupational Experience and Psychological Functioning: An Assessment of Reciprocal Effects." American Sociological Review. 3897-118.

Kohn, Melvin L. and Carmi Schooler. 1988. Work and Personality. Norwood, NJ: Ablex.

Kuncel, Nathan R. and Sarah A. Hezlett. 2007. "Standardized Tests Predict Graduate Students' Success." Science. 315:1080-1081.

LeBreton, James M. and Jenell L. Senter. 2008. "Answers to 20 Questions about Interrater Reliability and Interrater Agreement." Organizational Research Methods. 11:815-852.

Lewis, Phil M., David R. Rivkin, and Pam Frugoli. 2011. "Overview of O*NET Data Collection and Activities." OIDAP Meeting presentation (May 4).

Lieberman, Trudy. 2013. "Disability, Social Security, and the Missing Context." Columbia Journalism Review. (May 31). Acccessed 1/10/15.

Lipsky, Michael. 1980. Street-Level Bureaucracy: Dilemmas of the Individual in Public Service. New York: Russell Sage.

Maier, Mark H. and Jennifer Imazeki. 2013. The Data Game: Controversies in Social Science Statistics (4 ed.). Armonk, N.Y.: M.E. Sharpe, Inc.

Miller, Ann R., Donald J. Treiman, Pamela S. Cain, and Patricia A. Roos. 1980. Work, Jobs, and Occupations: A Critical Review of the *Dictionary of Occupational Titles*.

Committee on Occupational Classification and Analysis, National Research Council. Washington, DC: National Academies Press.

Morgeson, Frederick P. and Campion, Michael A. 2000. "Accuracy in Job Analysis: Toward an Inference-based Model." Journal of Organizational Behavior. 21:819-827.

Morton, William R. 2014. "Primer on Disability Benefits: Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI)." Congressional Research Service Report RL32279.

Mummery, W. Kerry, Grant M. Schofield, Rebekah Steele, Elizabeth G. Eakin, and Wendy J. Brown. 2005. "Occupational Sitting Time and Overweight and Obesity in Australian Workers." American Journal of Preventive Medicine. 29:91–97.

National Research Council. 2013. Principles and Practices for a Federal Statistical Agency, Fifth Edition, Constance F. Citro and Miron L. Straf, editors. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nestoriak, Nicole and Brooks Pierce. 2009. "Comparing Workers' Compensation claims with establishments' responses to the SOII." Monthly Labor Review. May:57-64.

Popham, W. James. 1997. "Consequential Validity: Right Concern—Wrong Concept." Educational Measurement: Issues and Practice. 16:.9-13.

Prewitt, Kenneth. 2003. "Politics and Science in Census Taking." New York, NY: Russell Sage Foundation and Washington, DC: Population Reference Bureau.

Reeder, Matthew C. and Suzanne Tsacoumis. 2014. "O*NET® Analyst Occupational Skills Ratings: Analysis Cycle 14 Results." HumRRO, Alexandria VA.

Rindskopf , David. 2002. "Reliability: Measurement." International Encyclopedia of the Social and Behavioral Sciences. Neil J. Smelser and Paul B. Baltes, eds. Amsterdam: Elsevier.

Ruser, John. 2008. "Examining evidence on whether BLS undercounts workplace injuries and illnesses." Monthly Labor Review. August:20–32.

Sackett, Paul R., Dan J. Putka, and Rodney A. McCloy. 2012. "The Concept of Validity and the Process of Validation." Pp. 91-118 in The Oxford Handbook of Personnel Assessment and Selection, Neal Schmitt, ed. Oxford, UK: Oxford University Press.

Sanchez, Juan I. and Edward L. Levine. 2012. "The Rise and Fall of Job Analysis and the Future of Work Analysis." Annual Review of Psychology. 63:397–425.

Schmitt, Neal. 2014. "Personality and Cognitive Ability as Predictors of Effective Performance at Work." Annual Review of Organizational Psychology and Organizational Behavior. 1:45-65.

Schmitt, Neal, Jessica Keeney, Frederick L. Oswald, , Timothy J. Pleskac, Abigail Q. Billington, Ruchi Sinha, and Mark Zorzie. 2009. "Prediction of Four-Year College Student Performance Using Cognitive and Non-Cognitive Predictors and Impact on Demographic Status of Admitted Students." Journal of Applied Psychology. 94:1479-1497.

Schober, Michael F. and Frederick G. Conrad. 1997. "Does Conversational Interviewing Reduce Survey Measurement Error?" Public Opinion Quarterly. 61:576–602.

Social Security Administration. Occupational Information Development Panel. Quarterly Meeting transcript. May 4, 2011.

Sommers, Dixie and Laurie Salmon. 2011. "Sampling and Collection in the Occupational Employment Statistics (OES) Program." Presentation to the Occupational Information Development Advisory Panel (OIDAP) (May 4).

Spenner, Kenneth I. 1980. "Occupational Characteristics and Classification Systems: New Uses of the Dictionary of Occupational Titles in Social Research" Sociological Methods & Research. 9:239-264.

Spenner, Kenneth I. 1988. "Social Stratification, Work, and Personality." Annual Review of Sociology. 14:69-97.

Steele, Rebekah and W. Kerry Mummery. 2003. "Comparative Analysis of Occupational Physical Activity Across Occupational Categories." Journal of Science and Medicine in Sport. 6:398-407.

Tinsley, Howard E.A. and David J. Weiss. 2000. Pp.95-124 in Handbook of Applied Multivariate Statistics and Mathematical Modeling. Howard E.A. Tinsley and Steven D. Brown, eds. Amsterdam: Elsevier.

Trapani, Mark and Deborah Harkin. 2011. "Occupational and Medical-Vocational Claims Review Study." Social Security Administration (May).

Tsacoumis, Suzanne and Chad H. Van Iddekinge . 2006. "A Comparison of Incumbent and Analyst Ratings of O*NET Skills." HumRRO, Alexandria VA.

Tsacoumis, Suzanne and Shannon Willison. 2010. "O*NET® Analyst Occupational Skill Ratings: Procedures." HumRRO, Alexandria VA.

Uebersax, John (a). "Kappa Coefficients: A Critical Appraisal." http://john-uebersax.com/stat/kappa.htm accessed 12/20/14.

Uebersax, John (b). "Statistical Methods for Rater and Diagnostic Agreement" http://www.john-uebersax.com/stat/agree.htm accessed 12/20/14.

Vallas, Rebecca and Shawn Fremstad. 2014. "Social Security Disability Insurance: A Bedrock of Security for American Workers." Washington, DC: Center for American Progress.

Walmsley, PhilipT., Michael W. Natali, and John P. Campbell. 2012. "Only Incumbent Raters in O*NET? OhYes! Oh No!" International Journal of Selection and Assessment. 20:283-296.

Wolf, Jeffrey S. and David W. Engel. 2013. "Restoring Social Security Disability's Purpose." Regulation. Spring:46-54.

Annex: Initial literature review

Method	Evidence of validity / reliability of data collection method?	Evidence of validity/reliability of data collection source?
Archival	• Cane and Treiman (1981) had analysts rate written job descriptions. Estimated reliabilities of occupational characteristics, in general, were not very high, with an average minimum estimate of 0.63, and an average medium estimate of 0.70. The "things" and strength variables were very unreliably rated, and the authors cautioned that the strength rating should be used with caution. Ratings in manufacturing were higher than in service industries. Aptitude items were not reliably measured. • The National Research Council (1980) reviewed the DOT in terms of general usefulness, required research, and organizational changes. Several severe criticisms were noted including the use of incomplete job analyses, poor coverage of occupations, limited observations of jobs, and failure to meet expected standards.	• Cane and Treiman (1981) used factor analysis on 44 occupational characteristics to determine the key variables. Six significant factors emerged from the analysis, which accounted for 95% of the common variance. The factors were labeled substantive complexity (17 items), motor skills (10 items), physical demands (5 items), management (8 items), interpersonal skills (4 items), undesirable working conditions (3 items).

Method	Evidence of validity / reliability of data collection method?	Evidence of validity/reliability of data collection source?
Questionnaire	• Campbell et al (1997) in a small study found general agreement between questionnaire and observation of physical tasks. Lowest level of agreement for frequency and quantity lifted. • Stock et al. (2005) found that certain questions related to physical work demands and evaluated for reproducibility and validity performed well. • Handel (2010) developed the Survey of Workplace Skills, Technology, and Management Practices (STAMP). He found that question items that are more objective and assessed actual job requirements rather than subjective perceptions of job duties particular to the person performing the job minimized measurement error. His survey was validated on 2,000 job incumbents. • Morgeson and Humphrey (2006) developed the Work Design Questionnaire (WDQ). The WDQ was validated with 540 incumbents holding 243 jobs. The measure assesses primarily cognitive factors such as: autonomy, decision making, work methods, task variety, task significance, task identity, job complexity, information processing, independence, as well as some physical demands, work conditions, and information about equipment use. This measure is primarily useful in helping to identify the general nature and design of different types of occupations in the economy.	• Morgeson et al (2007) assert that questionnaires which are less task specific should be completed by subject matter experts/job analysts. • Morgeson et al. (2004) found that ability statements made by workers were more prone to bias than task statements. Supervisors and trained job analysts did not show this bias. • Barrero et al. (2009) summarized the validity of self-reports in three major databases: EBSCOhost, Web of Science and PubMed. They concluded that it is difficult to evaluate the validity of self-reported exposures based on currently available validity assessment studies. The reported data support the hypothesis that validity depends on study-specific factors that are often not examined. • Karasek et al. (1998) developed the Job Content Questionnaire. They found incumbents display some biased reporting of their job duties, including individual differences in of job perceptions of stressors. Common method variance inflated associations, suggesting incumbents exaggerated their abilities. • Green and James (2003) assessed the correspondence between employees and supervisors' ratings of job requirements. In general, there was a fairly good

Method	Evidence of validity / reliability of data collection method?	Evidence of validity/reliability of data collection source?
	• Pope et al. (1998) developed a questionnaire to quantify physical demands of different occupations. Weight estimation accuracy assessed on visual analogue scales and compared to direct observations made by the researchers. Results showed greatest accuracy of estimates were for dichotomous levels of work postures (e.g., do you ever lift objects with one hand?). Employees were fairly accurate at estimating amount of repetitive movement of wrists and arms. Employees tended to overestimate duration of physical demands slightly, approx. 5 minutes compared to direct observations. Repetitive movements of upper limbs overestimated to a greater extent but still below 10 minutes. Employees don't seem to distinguish well between "carrying" and "lifting." Jobs with greater variability might require longer observation periods. Recommended video-taping and comparing to self-reports. • Wall et al. (1995) developed new scales that assessed timing, control, method control, monitoring demand, problem-solving demand and production responsibility in different jobs. They found improvement of the problem-solving demand scale; tested replicability of measurement model by formal factorial invariance tests across four samples; additional information on scale reliability and construct validity; normative data for a wide range of shopfloor and related jobs.	correspondence between employee and supervisor ratings of job requirements. However, some discrepancies in reports of job autonomy were found, and employees tend to rate their skills at a higher level than supervisors. Job incumbents might inflate their abilities, or supervisors may have a broader perspective on skill level. • Landsberig & Theorell (2000) discuss several types of self-administered questionnaires to assess job requirements. The authors recommend using observations to validate self-report data, and that questionnaires are tailored to the specific job being assessed, rather than assessing global aspects of job requirements across many different types of jobs.

Method	Evidence of validity / reliability of data collection method?	Evidence of validity/reliability of data collection source?
Checklist/Task Lists	• Levine et al (1983) use 93 analysts and 7 job analysis methods and find position analysis questionnaires, functional job analysis, and task inventory are most effective.	• Manson et al (2000) find general agreement in ratings for task components between incumbents and supervisors. • Morgeson (2007) suggests that regardless of the source, paradata on respondents should be collected to test reliability. • Green and Stutzmer (1986) and Green and Veres (1990) find that incumbents tend to identify themselves as performing tasks that are not actually part of their job. • Stock et al. (2005) found that results of validity studies comparing self-reports with reference methods (structured interview, observation, or direct measurement) were mixed. • Morgeson et al. (2014) showed that job analysts are best at providing task lists that distinguish between holistic and decomposed job tasks when they are more experienced and cognitively focused on the task. Job incumbents provide better data and distinguish between decomposed and holistic job tasks when they have had more task experience, current job experience was related to the job, and previous job experience.

Method	Evidence of validity / reliability of data collection method?	Evidence of validity/reliability of data collection source?
Individual interview	• In a meta-analysis, Dierdoff and Wilson (2003) found that task data produced higher levels of inter-rater reliability than generalized work activity data, but lower levels of intra-rater reliability. After a statistical correction for scale length and number of raters, task data had both higher inter- and intra-rater reliabilities. Intra-rater reliability was higher than inter-rater reliability. Incumbents displayed the lowest reliabilities. Scales of frequency and importance were the most reliable.	
Observation interview	• Lysaght and Shaw (2010) describe key challenges associated with conducting reliable and valid job analyses. These include the psychometric properties of the rating scales, respondents' ability and willingness to provide information, quality of observational data that is affected by lack of rater familiarity and the sufficiency and representativeness of time samples of job (especially highly variable jobs), and rater competence. • Descatha et al. (2009) evaluated agreement between a questionnaire and an observational checklist for exposure assessment in the setting of an upper-limb work-related musculoskeletal disorders. The correlation between the two methods was low (r = .06). Self-assessment predicted incidence of MSD, but direct observation did not. The poor agreement between self-administered and observational methods suggests that these methods may be	• Stock et al. (2005) found that results of validity studies comparing self-reports with reference methods (structured interview, observation, or direct measurement) were mixed.

Method	Evidence of validity / reliability of data collection method?	Evidence of validity/reliability of data collection source?
	complementary, and should be used in combination. • Kilbom (1994) reviewed 19 different types of observational methods to assess job requirements. The conclusion was that these methods are a compromise between subjective, less reliable, cheap self-assessments; and expensive, but more precise direct measures. The reliability of observational methods depends on several factors, including the mode (video or direct observation), the complexity of the tasks observed, the number of observers, learning effects of the observers, number of categories observed, the scales used, real-time or time-sampling techniques. Some recommendations for improving observational techniques are outlined (e.g., body movements like sit, stand, kneel can be observed with a high degree of accuracy, head and neck postures are difficult to observe, more training of observers should increase validity/reliability, manual handling can be observed crudely.	

Method	Evidence of validity / reliability of data collection method?	Evidence of validity/reliability of data collection source?
Comparative methods	• Spielholtz et al. (2010) compared self-report questionnaires, observational video analysis and direct measurement for exposure to conditions that put workers at risk for MSD. Extreme posture duration, repetition, hand force (estimated from electromyography) and movement velocity were assessed for 18 subjects while performing each of three jobs processing tree seedlings. Self-reports were the least precise assessment method, which consistently overestimated exposures for each of the measured risk factors. Objective measures worked well for wrist extension, but not all movements. Direct measurement was the best method in terms of precision. Video analysis and direct measurement showed moderate disagreement. Self-reports were very imprecise compared to direct measurement.	
	• Nordstrom et al. (1998) Compared self-reported and expert-observed estimates of work-related physical factors using a wide range of occupations and industries. A trained ergonomist observed 60 workers performing their job duties for 1 hour each. Ratings on physical factors were assessed, and these were compared to the employees' self-ratings on the same elements. The measures agreed more than by chance, but agreed less for the elements bending at the waist and twisting of the forearm. The median percent agreement was 71%, and the median correlation was 0.46.	• Nordstrom et al. (1998) argued that combining self-reports with observational data provides an adequate assessment of job exposure. They also mention that a systematic interview with employees, rather than a self-administered questionnaire, provides more accurate data about job exposure. Without a gold standard of ergonomic exposure and other job requirements, validating this type of data will be difficult.