Occupational Requirements Survey Job Observation Report

November 16, 2015

Executive Summary

The Occupational Requirements Survey (ORS) is an establishment survey conducted by the Bureau of Labor Statistics (BLS) for the Social Security Administration. The survey collects information on the vocational preparation and the cognitive and physical requirements of occupations in the U.S. economy, as well as the environmental conditions in which those occupations are performed.

BLS conducted a job observation test during the summer of 2015 to provide validation for the ORS physical elements by comparing the data collected during pre-production to those collected through direct job observation, which is more typical among small scale studies of job tasks. As part of this test, Field Economists (FE) re-contacted establishments who had responded to the ORS pre-production survey and observed workers actually performing their jobs to obtain data on the physical requirements of the job. Our analysis has two goals. First, to assess the level of inter-rater reliability between the two FEs observing the job. Second, to compare the data obtained from observation to the interview method used in pre-production. The comparisons are performed by survey element and at the occupational level defined by the eight-digit Standard Occupational Classification (SOC) code.

We find relatively high levels of inter-rater reliability among the FEs, suggesting that any future observations could be done with single observers. We also find relatively high rates of agreement between observed and collected data for most physical requirements. A closer examination of the elements with lower, but still reasonable, levels of agreement leads us to find some evidence that missing duration of physical elements in pre-production can lead to underestimates of the duration of certain physical elements.

Overview of the Occupational Requirement Survey

In the summer of 2012, the Social Security Administration (SSA) and the Bureau of Labor Statistics (BLS) signed an interagency agreement, which has been updated annually, to begin the process of testing the collection of data on occupational requirements. As a result, BLS established the Occupational Requirements Survey (ORS) as a test survey in late 2012. The goal of ORS is to collect and publish occupational information that meets the needs of SSA at the level of the eight-digit standard occupational classification (SOC) that is used by the Occupational Information Network (O*NET).[1]

The ORS data are collected under the umbrella of the National Compensation Survey (NCS), which uses Field Economists (FEs) to collect data. FEs generally collect data elements through either a personal visit to the establishment or remotely via telephone, email, mail, or a combination of modes.

[1] The occupational classification system most typically used by BLS is the six-digit SOC (http://www.bls.gov/soc/), generally referred to as "detailed occupations". O*NET uses a more detailed occupational taxonomy (https://www.onetcenter.org/taxonomy.html), classifying occupations at eight-digits and referring to these as "O*NET-SOC 2010 occupations". There are 840 six-digit occupations and 1,110 eight-digit occupations.

For ORS, FEs are collecting occupationally-specific data elements to meet SSA's needs in the following categories:

- Physical demands
- Specific vocational preparation (SVP)
- Mental and cognitive demands
- Environmental conditions in which the work is performed.

In fiscal year 2015, the Bureau of Labor Statistics (BLS) completed data collection for the Occupational Requirements Survey (ORS) pre-production test. The pre-production test might better be described as a "dress rehearsal" as the collection procedures, data capture systems, and review were structured to be as close as possible to those that will be used in production.[2] Information on the results of this pre-production test are available at http://www.bls.gov/ncs/ors/preprod_coll.htm.

Background on Observation Test

The ORS job observation test was intended to assess whether the data collected through ORS interview collection methods are systematically different than data collected through direct observation. This test was conducted in response to both Federal Register Notice public comments and an external subject matter expert's recommendations for testing and validation of ORS survey design.[3]

The job observation test was conducted in summer 2015, running from June-September. The observation test involved re-contact of a subset of establishments that were interviewed as part of the pre-production test. Two FEs were sent to observe select jobs within the establishment and record data on the physical and environmental data elements during a one hour observation period.[4] The one hour observation period sought to achieve a balance between gathering data on as many quotes as possible and the respondent burden involved in conducting such a test.

As the goal of ORS is to produce estimates at the eight-digit O*NET SOC level, the observation test was structured to allow to us compare pre-production data to observed data at the eight-digit SOC level as well. Thus, a subset of occupations were chosen for inclusion in the test. The subset was chosen based on two criteria:

1. Roughly 40 "quotes" were collected at the eight-digit level in the ORS pre-production test. Quotes are the unit of collection in ORS and a quote is roughly equivalent to a job at an establishment.[5]
2. The jobs have been identified as a subset of occupations of particular interest to SSA due to their prevalence in the work histories of those applying for Social Security Disability Insurance or due to the physical requirements of the jobs.

[2] The sample design was similar that which will be used in production, but altered to meet test goals.
[3] A link to the subject matter expert's report can be found here: http://www.bls.gov/ncs/ors/pre-prod_estval.pdf
[4] The FEs used devices to collect data on a subset of the ORS environmental elements. Unfortunately, there were problems with the readings from some devices, resulting in us dropping the analysis of those elements from this report.
[5] For more information on "quotes" as measures in NCS (equivalent to their use as measures in ORS), see the BLS Handbook of Methods, http://www.bls.gov/opub/hom/pdf/homch8.pdf.

This resulted in the following occupations sampled for the observation test:

- Nursing assistants
- Cooks, institution and cafeteria
- Cooks, restaurant
- Waiters and waitresses
- Dishwashers
- Janitors and cleaners
- Maids and housekeeping cleaners
- Cashiers
- Retail Salespeople
- Receptionists and information clerks
- Team Assemblers
- Industrial truck and tractor operators
- Laborers and freight, stock, and material movers, hand

Procedures for the Observation Test

The sample consisted of 540 preselected quotes (456 from private industry, and 84 from State and local governments) from existing ORS pre-production test establishments. The test sample frame units were ordered by a combination of geography, industry and size class to ensure a good distribution of available establishments within each of the targeted occupations. The sample was drawn as two separate lists to allow occupations collected near the end of the pre-production collection to have a chance of selection.

For each of the sampled establishments and occupations, the first of two FEs secured the appointment and explained to the respondent the reason for the follow-up visit. Both field economists then simultaneously collected data via personal visit. If possible the FEs observed the same employee for the entire 60-minute observation period, but if for some reason that was not possible each FE was to observe an employee in the preselected job and document the situation in the remarks field. The FEs were instructed not to look at data recorded from the pre-production test for their establishments nor to discuss or reconcile their data with one another. Each FE independently recorded and coded their observations during the personal visit. FEs attempted to be as inconspicuous as possible and not to ask questions of the observed employee.

The FEs were instructed to code the duration in minutes and to code a duration of zero if the element was not observed. Some elements had additional questions such as whether it was performed with one hand or both, and for these elements the FE was to check the appropriate box and note the duration in minutes. FEs checked their data for accurate recording before marking the schedule and quote complete. Two collection debrief meetings occurred (one mid-test and one at the end of collection) to assess how the process worked.

Response Rates

Of those 540 jobs in the sample, FEs contacted 405 and observed 244, a 60% response rate. As shown in Table 1, the refusal rate varied considerably by occupation.

Table 1: Job Observation Response Rates

Occupation	Observed	Not contacted	Refused	Total Sampled	Response Rate
Nursing assistants	9	11	20	40	31%
Cooks, institution and cafeteria	19	9	12	40	61%
Cooks, restaurant	16	13	11	40	59%
Waiters and waitresses	19	11	10	40	66%
Dishwashers	13	15	12	40	52%
Janitors and cleaners	25	6	9	40	74%
Maids and housekeeping cleaners	20	12	8	40	71%
Childcare workers	6	16	10	32	37%
Cashiers	22	7	11	40	67%
Retail salespeople	17	10	13	40	57%
Receptionists and information clerks	23	6	11	40	68%
Team assemblers	17	4	7	28	71%
Industrial truck and tractor operators	17	8	15	40	53%
Laborers and freight, stock and , material movers, hand	21	7	12	40	64%
Total	244	135	161	540	60%

As expected, childcare workers and nursing assistants had very high refusal rates. These refusals largely stemmed from establishments' concerns about privacy under state and national laws. Some successful observations of both of these occupations did occur during the observation test, however, due to the small sample size we do not include them in our test analysis.

Measures of Inter-rater Agreement

The use of two FEs was intended to determine whether there was adequate inter-rater agreement. Requiring two FEs to observe the same job at the same time led to logistical difficulties in scheduling the observations and, in some cases, very close quarters for the FEs if the job being observed was confined to a small space (such as dishwashers). In addition to evaluating accuracy of the data collected, calculating inter-rater reliability might also lead to structural changes to any future job observations (i.e. some jobs for which only one rater is required).

To evaluate inter-rater reliability, we compared the duration data on a set of the physical requirements of jobs.

Table 2: ORS Physical Elements Analyzed

Physical Demand	Description
Climbing Ramps/Stairs	Ascending or descending ramps and/or stairs using feet and legs.
Climbing Ladders/Ropes/Scaffolding	Ascending or descending ladders, scaffolding, ropes, or poles and the like.
Communicating Verbally	Expressing or exchanging ideas by means of the spoken word to impart oral information.
Crawling	Moving about on hands and knees or hands and feet.
Crouching	Bending body downward and forward by bending legs and spine.
Fine Manipulation	Picking, pinching, or otherwise working primarily with fingers rather than the whole hand or arm.
Foot/Leg Controls	Use of one or both feet or legs to move controls on machinery or equipment.
Gross Manipulation	Seizing, holding, grasping, turning, or otherwise working with hand(s)
Keyboarding	Entering text or data into a computer or other machine by means of a keyboard or other device.
Kneeling	Bending legs at knees to come to rest on knee(s).
Lifting/Carrying	Lifting is to raise or lower an object from one level to another (includes upward pulling). Carrying is to transport an object usually by holding it in the hands, or arms, but may occur on the shoulder.
Pushing/Pulling	Exerting force upon an object so that it moves away (pushing) or toward (pulling) the force.
Reaching At/Below Shoulder Level	Extending hands and arms from 0 up to 150 degrees in a vertical arc.
Reaching Overhead	Extending hands and arms in a 150 to 180 degrees vertical arc.
Stooping	Bending the body downward and forward by bending the spine at the waist.

The duration of most physical elements for pre-production were classified into five categories:

1. Not present
2. Seldom – up to 2% of the day.
3. Occasionally – 2% up to one-third of the day.
4. Frequently – one-third up to two-thirds of the day.
5. Constantly –two-thirds or more of the day.

For the job observation test, we also classified duration using these categories based on the percent of the observed time spent in the activity. We then calculated inter-rater reliability by element. The inter-rater agreement is weighted to penalize for disagreements of higher magnitudes (e.g. If FE A records an element as occurring frequently and FE B records the same element as not present it is penalized more than one recording frequently and the other occasionally).[6]

Table 3: Inter-rater Agreement Measures, by Element

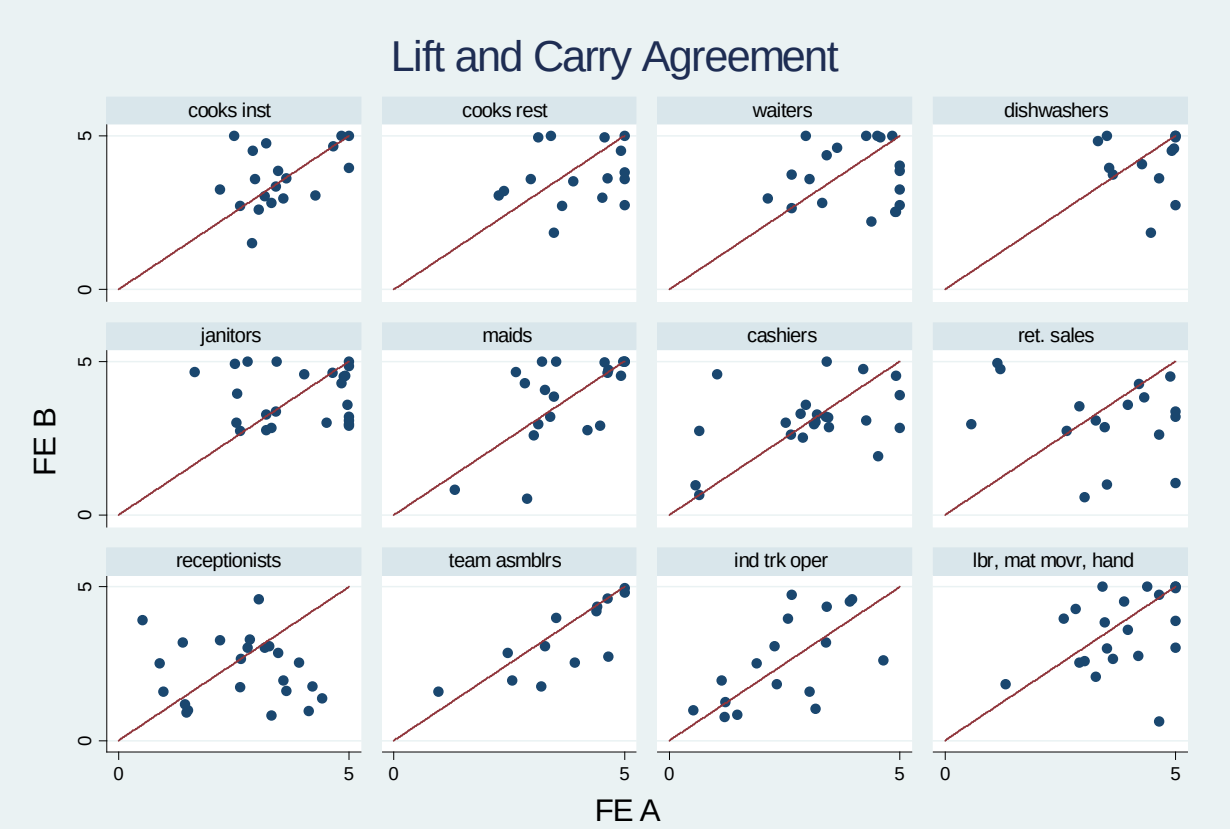
Physical Demand	Agreement
Climbing Ladders/Ropes/Scaffolding	98.2%
Climbing Ramps/Stairs	94.1%
Communicating Verbally	85.6%
Crawling	98.2%
Crouching	85.4%
Fine Manipulation	82.3%
Foot/Leg Controls	95.2%
Gross Manipulation	86.6%
Keyboarding	97.1%
Keyboarding- 10 Key	96.7%
Keyboarding- Other	95.5%
Keyboarding- Touchscreen	95.8%
Kneeling	94.8%
Lifting/Carrying	77.6%
Pushing/Pulling with Feet	95.0%
Pushing/Pulling with Feet and Legs	82.4%
Pushing/Pulling with Hands and Arms	78.9%
Reaching At/Below Shoulder Level	79.6%
Reaching Overhead	85.1%
Stooping	82.7%

To further examine the elements with lower levels of agreement, we produce scatter plots of the rankings of the two FEs separately for each occupation for the elements with inter-rater agreement below 80% - lifting/carrying, pushing/pulling with hands and arms, and reaching at or below shoulder level. Our goal in plotting agreement by occupation is to determine whether certain occupations appear to have higher levels of disagreement, which may be useful for training purposes.

[6] Statistical measures of agreement and reliability were also calculated. These are available in the report, **"**Agreement across Modes of Collection in the Occupational Requirements Survey: Results from a Pilot Job Observation Test," to be presented at the Federal Committee on Statistical Methodology in December 2015. This report will be posted to the ORS website in December.

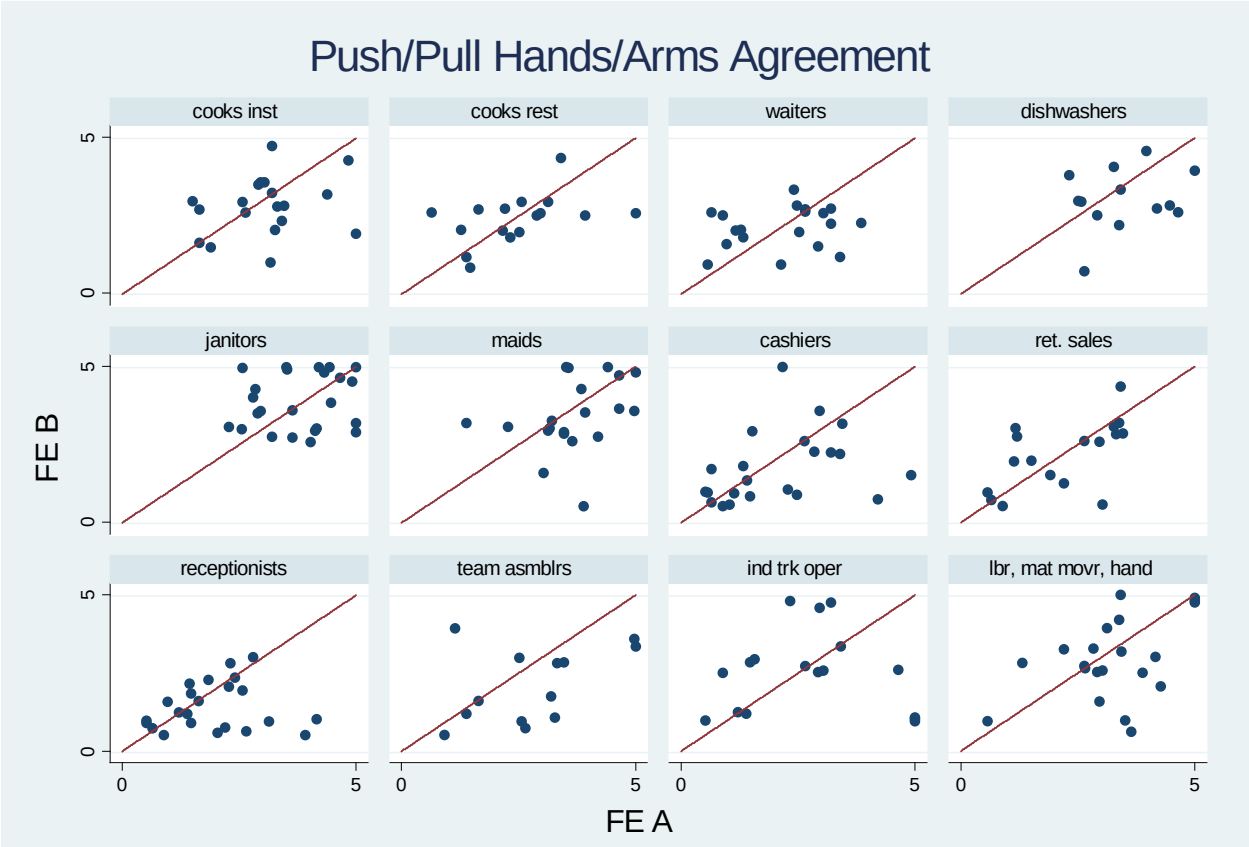
The 45 degree line provides a reference line for perfect agreement – both FEs' duration measures fall into the same category. Points substantially off the diagonal line represent major disagreements in the duration ranking.

Figure 1. Scatterplots of Lift and Carry Agreement, by Occupation



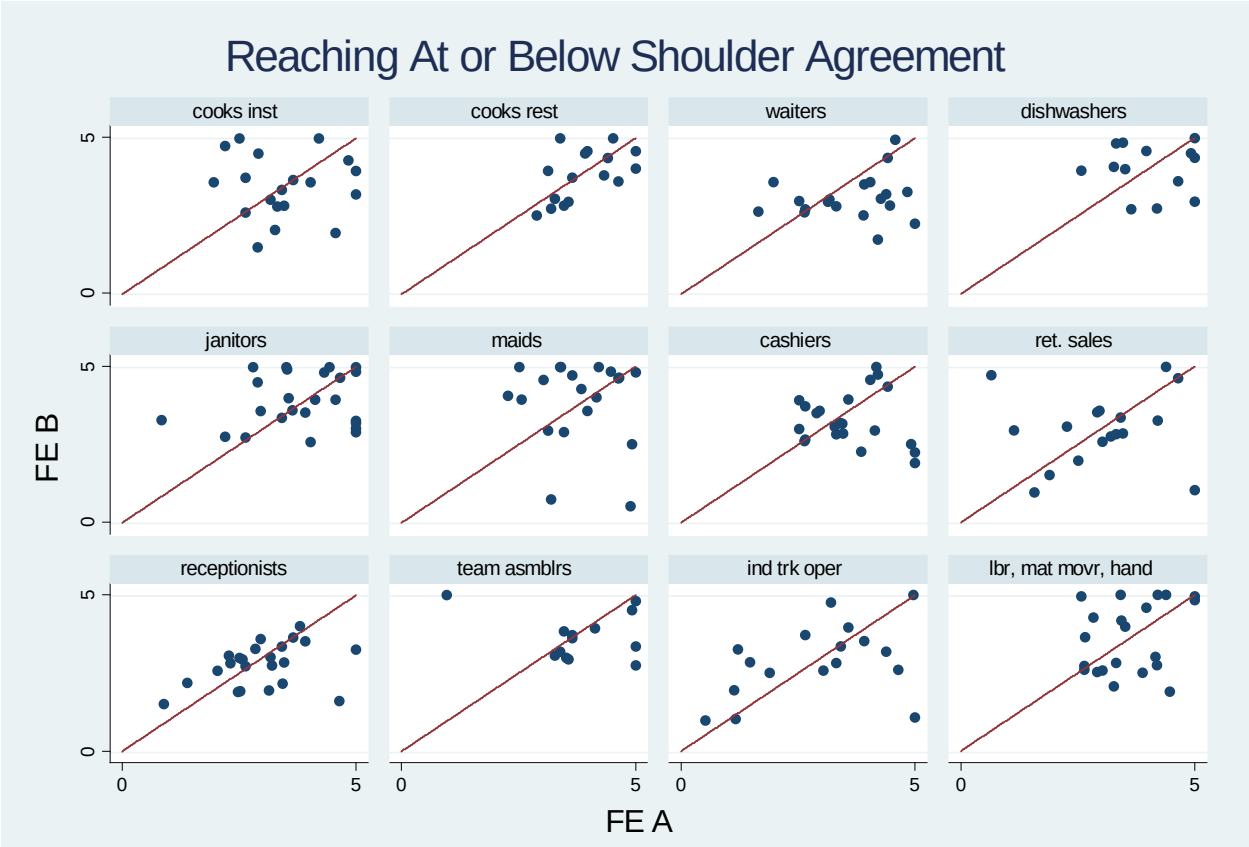
Turning first to the lift and carry element, we see that the levels of disagreement appear to vary by occupation. Team assemblers, for example, have durations that mostly lie on the diagonal line and when disagreements do occur they rarely are large in magnitude (i.e. it is never the case that FE A rates lift and carry as constantly present and FE B rates it as only occasionally present.) Receptionists, on the other hand, display several cases where the categories of rating are substantially different. In multiple cases one FE rated lift and carry as frequent or constant when the other rated it as not present. This suggests that additional training on this element is needed in particular occupations. For this element there was some confusion about the threshold required to classify an action as a "lift and carry", which may drive the disparity.

Figure 2. Scatterplots of Pushing and Pulling with Hands and Arms Agreement, by Occupation



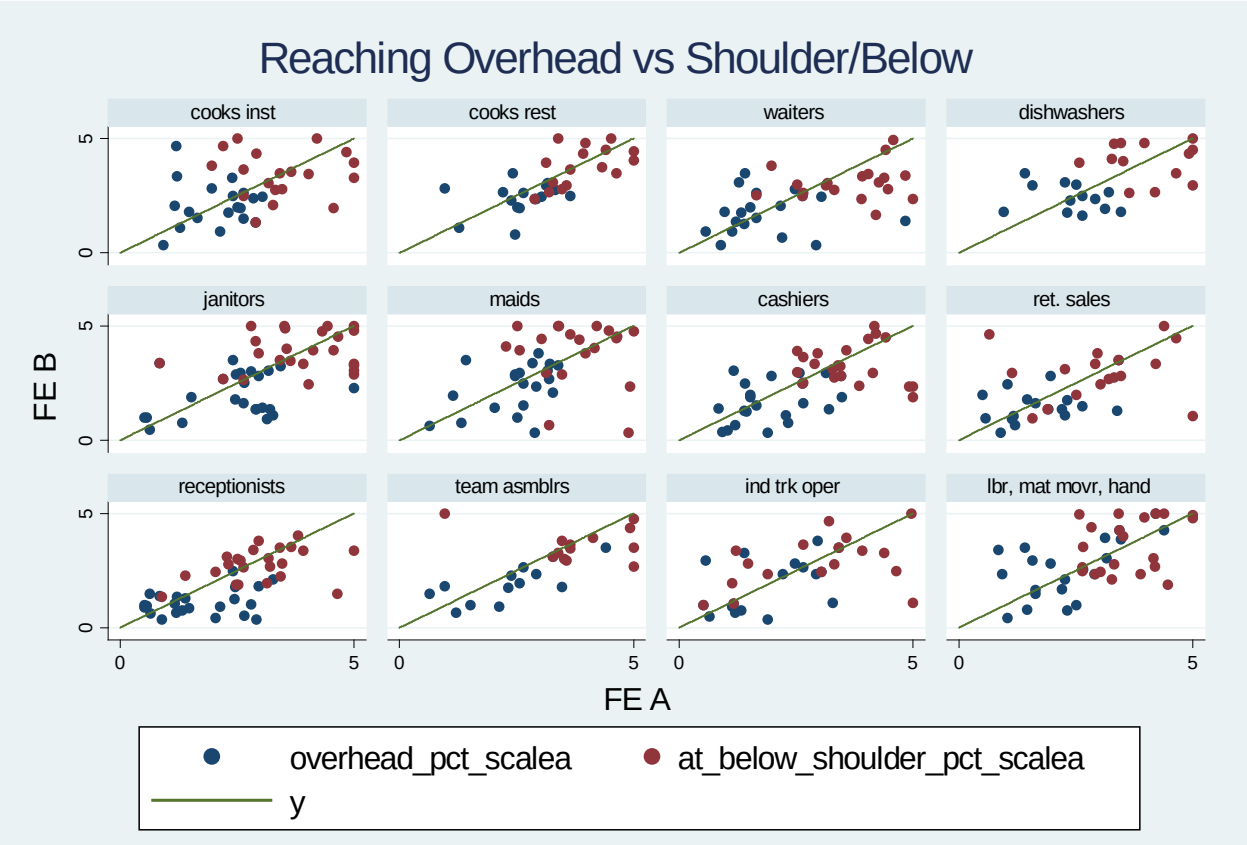
A similar result can be seen for the push/pull with hands and feet element. For most occupations there are few cases of the ratings differing by more than one category (ex. frequent versus constant), but among cashiers and receptionists, there were more substantial disagreements. This, again, suggests some additional guidance regarding the minimum threshold for pushing/pulling may be needed.

Figure 3. Scatterplots of Reaching At or Below Shoulder Agreement, by Occupation



Finally, we look at reaching at or below the shoulder (Figure 3). The overall agreement level for this element is roughly 80%, however we opted to break this element out by occupation to determine whether some of the disagreement between reaching may be explained by different interpretations by the FEs of reaching at or below the shoulder versus reaching overhead. To examine this we superimpose the reaching overhead into the reaching at or below the shoulder graph (Figure 4). Our expectation is that if the FEs are classifying the elements differently from one another (i.e. identifying a reach as overhead when the other identifies it at/below the shoulder) then the off-diagonal measures of disagreement should "cancel out" – that is, if many cases of disagreements are seen above the diagonal for reaching at/below the shoulder, then the disagreements for reaching overhead should be more likely to lie below the diagonal line.

Figure 4: Scatterplots of Reaching At or Below Shoulder and Reaching Overhead, by Occupation



We see this pattern in occupations such as laborers, cooks, waiters, and maids, suggesting that additional training on reaching would be helpful. Looking at maids, for example, one can see many cases where reaching at or below the shoulder (the red dot) is recorded as frequent or constant by FE B but only as occasional by FE A. When we superimpose reaching overhead on the same graph (the blue dot) we see that FE B is more likely to record not present/seldom when FE A records the duration as occasionally. This suggests that FE A and FE B may have different perceptions about the reaching thresholds – overhead versus at or below the shoulder. It is important to note that this distinction has been addressed through minor changes to the definitions for ORS. In the current production collection, overhead reaching involves a hand higher than the head. All other reaching is at or below shoulder level.

Observation and Pre-production Agreement

We next turn to an evaluation of the agreement between the observed values of the data elements and those collected in the pre-production test. We will refer to these as "observed" and "collected" values hereon. Measuring agreement between observed and collected data is complicated by two factors:

1. The observation test was of short duration, which may lead to discrepancies between the presence/absence of certain physical requirements. In particular, we expect high degrees of agreement in the presence or absence of physical requirements for those requirements with

durations that fall into the “frequent” or “constant” categories and lower levels of agreement for elements that occur “occasionally” or “seldom.”

2. In pre-production collection, roughly 20% of the physical requirements that were classified as “present” in the job had no duration provided by the respondent. The unknown duration is especially high in particular elements – in the sample of jobs that were observed, the collected data has missing duration in nearly 30% of the cases for communicate verbally and 25% of the cases for fine manipulation.

For most (but not all) observed jobs we have pairs of observations that we can compare with the collected data from pre-production on the same job at the same establishment. There are multiple ways to deal with having two observations on the same job – we consider the mean values across the two field economists, the max of the two values, and the min of the two values. The results were not appreciably different for these three approaches, which is not surprising since the inter-rater agreement was relatively high. In our analysis we use the maximum value approach to capturing the duration of the observed elements.

We measure agreement in the duration of the physical elements, using the weighted approach that was also used to generate inter-rater agreement in the earlier section and “penalizes” for higher increments of disagreement in the duration results.

Table 4: Agreement Between Observed and Collected Measures, by Element

Physical Demand	Agreement
Climbing Ladders/Ropes/Scaffolding	91.46%
Climbing Ramps/Stairs	84.68%
Communicating Verbally	80.61%
Crawling	95.41%
Crouching	70.17%
Fine Manipulation	78.18%
Foot/leg Controls	90.40%
Gross Manipulation	76.72%
Keyboarding	90.10%
Keyboarding- 10 Key	94.29%
Keyboarding- Other	93.16%
Keyboarding- Touchscreen	91.46%
Kneeling	81.15%
Pushing/Pulling with Feet	94.75%
Pushing/Pulling with Feet and Legs	73.40%
Pushing/Pulling with Hands and Arms	71.29%
Reaching At/Below Shoulder Level	77.39%
Reaching Overhead	79.86%
Stooping	75.00%

Again we focus on the elements with the lowest levels of agreement: crouching, stooping, reaching, manipulating, and pushing/pulling and look for patterns at the occupation level.

We look for evidence that observed durations are consistently higher than the collected durations, suggesting that establishment respondents may be understating the duration of certain physical requirements.

For the reaching elements, we do not see a pattern in the agreement plots (generally they cluster around the 45 degree line with points both above and below the line). For crouching we see higher durations for receptionists and maids in the observed than the collected data. Similarly, for pushing/pulling with feet and legs there are higher durations in the observed data for maids, waiters, and janitors. For the remainder of the elements – stooping, fine manipulation, gross manipulation, and pushing/pulling with hands and arms, durations for the observed data appear to be higher than the collected data across several of the occupations collected.

This provides some evidence that respondents underreport duration, when duration is actually reported. Recall that roughly 20% of the physical elements collected in pre-production were classified as "present, duration unknown". Can the observation data tell us something about this category among the elements whose duration appears to be understated in the collected data?

It appears so. We examine the observed frequency of the duration categories for stooping, fine manipulation, gross manipulation, and pushing/pulling with hands and arms. For stooping, 49 of the quotes that were observed were categorized as "present, duration unknown" during pre-production. Among these, 88% were observed with durations of occasionally or higher (37% were frequently or constantly). For fine manipulation, 58 quotes that were observed were categorized as "present, duration unknown" during pre-production and of these 91% were observed with duration occasionally or higher. The same patterns hold for gross manipulation and pushing/pulling with hands and arms.

From this, it appears that the "underestimate" of duration from the collected data is due to the missing duration being more likely to correlate with long duration observed. As a counter-example, 46 observed quotes had reaching overhead categorized as "present, duration unknown" in pre-production and only one of these was classified as frequently or constantly in the observation test.

Summary and Recommendations

The purpose of the job observation test was to provide validation for the ORS physical elements by comparing the data collected during pre-production to those collected from a different source – observation. Two field economists were assigned to observe the same job for 60 minutes and record the duration of each of the physical elements of the job.

Initial results show high levels of inter-rater reliability among the FEs, suggesting that any future observations could be done without pairs of FEs, provided that training is provided prior to the test that focuses on understanding the definitions of the elements, as well as the thresholds for the physical elements to be deemed present.

Comparing the observed data to that collected during pre-production proved somewhat more complicated due to the limited length of the observation resulting in some elements classified as "not

present" that were more likely present with very low frequency ("seldom"). The measures of agreement for duration are relatively strong, however, suggesting that the collected data and observed data have high levels of agreement across most elements.

Drilling down to the elements with lower levels of agreement, leads us to find some evidence that "present, duration unknown" classifications in pre-production can lead to underestimates of the duration of certain physical elements. The observation test suggests that for several elements the missing duration is distributed very differently than the collected duration, leading to estimates that may under- or overstate the frequency of a physical element.

One approach to dealing with this is to provide additional guidance to both field economists and the respondent to aid in estimating duration of the element. A second approach would be to continue to select a subsample of ORS occupations for direct observation.