Agreement across Modes of Collection in the Occupational Requirements Survey: Results from a Pilot Job Observation Test

Tiffany Chang
Kristen Monaco
Kristin Smyth[1]

U.S. Bureau of Labor Statistics
2 Massachusetts Ave., NE, Room 4130,
Washington, DC 20212

Abstract

The Occupational Requirements Survey (ORS) is an establishment survey conducted by the Bureau of Labor Statistics (BLS) for the Social Security Administration (SSA). 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 jobs are performed.

These data are collected by BLS Field Economists who conduct interviews with establishment representatives. A question that has been raised during the collection process is whether the data collected through this mode result in similar measurements as data that would be collected through direct job observation (which is more typical among small scale studies of job tasks). To answer this question, BLS conducted a job observation pilot test during summer 2015. As part of this test, Field Economists recontacted establishments who had responded to the ORS pre-production survey and observed workers performing their jobs to obtain data on the physical requirements of the job.

This paper presents the results of an analysis that compares data obtained from observation to those collected through interviewing an establishment representative. The comparisons are performed by survey element and at the occupational level (defined by the eight-digit Standard Occupational Classification (SOC) code[2]). We find relatively high rates of agreement between observed and collected data for most physical requirements.

Introduction

In the summer of 2012, the SSA and 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 SOC system that is used by the Occupational Information Network (O*NET).

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.

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

- Physical demands

[1] Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

[2] 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 SOC codes and 1,110 eight-digit SOC codes.

- 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 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.[3] [4]

Background on the Observation Pilot 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.[5]

The job observation test was conducted in summer 2015, running from June-September. The observation test involved recontact 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.[6] 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 us to 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 occupation level in the ORS pre-production test. A quote is a sampled job that has been matched with an eight-digit O*NET occupation. Quotes are the unit of collection in ORS and a quote is roughly equivalent to a job at an establishment.[7]
2. The jobs identified were a subset of occupations of particular interest to SSA.

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
- Childcare workers
- Cashiers
- Retail Salespeople
- Receptionists and information clerks
- Team Assemblers

[3] The sample design was similar to what will be used in production, but altered to meet test goals.
[4] The ORS pre-production report is available at: http://www.bls.gov/ncs/ors/pre-prod-report.pdf.
[5] A link to the subject matter expert's report can be found here: http://www.bls.gov/ncs/ors/pre-prod_estval.pdf
[6] 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.
[7] 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.

- 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, an FE 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 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 to not ask questions of the observed employee.

In this test, the FEs were instructed to code the duration in minutes during the 60-minute observation period and to code a duration of zero if the element was not observed. These durations were later fit to the duration scale used in pre-production, described later. 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 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 any childcare workers or nursing assistants in our test analysis.

Measures of Inter-rater Agreement in the Observation Test

The use of two FEs in the observation test was intended to determine whether there was adequate inter-rater agreement, or consistency between observers. In addition to evaluating accuracy of the data collected, calculating inter-rater reliability might also lead to structural changes in future job observation tests (i.e., if inter-rater reliabilities are high, only one rater would be required). However, 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).

To evaluate inter-rater reliability, we compared the recorded duration of the physical requirements of jobs. The physical requirements are shown in Table 2.

Table 2: ORS Physical Elements Analyzed

Physical Demand	Description
Crawling	Moving about on hands and knees or hands and feet.
Crouching	Bending body downward and forward by bending legs and spine.
Kneeling	Bending legs at knees to come to rest on knee(s).
Stooping	Bending the body downward and forward by bending the spine at the waist.
Reaching Overhead	Extending hands and arms in a 150 to 180 degrees vertical arc.
Reaching At/Below Shoulder Level	Extending hands and arms from 0 up to 150 degrees in a vertical arc.
Communicating Verbally	Expressing or exchanging ideas by means of the spoken word to impart oral information.
Keyboarding	Entering text or data into a computer or other machine by means of a keyboard or other device. This is measured separately for standard keyboards, touchscreen, 10key, and other keyboarding.
Fine Manipulation	Picking, pinching, or otherwise working primarily with fingers rather than the whole hand or arm.
Gross Manipulation	Seizing, holding, grasping, turning, or otherwise working with hand(s)
Pushing/Pulling	Exerting force upon an object so that it moves away (pushing) or toward (pulling) the force. This is measured separately for hands and arms, feet and legs, and legs.
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.

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 and then used a weighted version of Cohen's kappa to measure inter-rater reliability. The kappa statistic measures the agreement against a benchmark of the expected agreement, bearing in mind that if there are only a few possible categories the FEs could randomly enter data and agree simply by chance. We use a weighted kappa statistic that penalizes for disagreements of higher

magnitudes. For example, 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.

Kappa generally ranges from -1 to +1. Negative values of kappa indicate that the level of agreement is less than the expected agreement. Similar to correlation measures, kappa statistics close to one imply a higher level of agreement. While there exists some controversy in the literature regarding thresholds of kappa, Landis and Koch (1977a) have proposed the following standards: ≤0 is poor, 0.01–0.20 is slight agreement, 0.21–0.40 is fair agreement, 0.41–0.60 is moderate agreement, 0.61–0.80 is substantial agreement, and 0.81–1 is almost perfect agreement.

Table 3: Percent Agreement and Kappa Measure of Observation Inter-rater Reliability

ORS Element	Agreement	Expected Agreement	Kappa	Prob>Z	PABAK
Crawling	98.2%	97.8%	0.19	<.01	0.97
Crouching	85.4%	73.2%	0.46	<.01	0.71
Kneeling	94.8%	84.8%	0.66	<.01	0.90
Stooping	82.7%	73.4%	0.35	<.01	0.59
Reaching overhead	85.1%	75.2%	0.40	<.01	0.66
Reaching At/Below Shoulder Level	79.6%	72.2%	0.27	<.01	0.93
Communicating Verbally	85.6%	71.4%	0.50	<.01	0.90
Keyboarding	97.1%	79.0%	0.86	<.01	0.94
Keyboarding- Touchscreen	95.8%	85.6%	0.71	<.01	0.89
Keyboarding- 10key	96.7%	93.8%	0.47	<.01	0.59
Keyboarding- Other	95.5%	91.8%	0.46	<.01	0.69
Fine Manipulation	82.3%	76.3%	0.25	<.01	0.49
Gross Manipulation	86.6%	78.8%	0.37	<.01	0.58
Pushing/Pulling with Hands and Arms	78.9%	66.0%	0.40	<.01	0.49
Pushing/Pulling with Feet and Legs	82.4%	69.9%	0.42	<.01	0.58
Pushing/Pulling with Feet	95.0%	94.8%	0.03	0.28	0.88
Climbing Ramps/Stairs	94.1%	87.9%	0.51	<.01	0.88
Climbing Ladders/Ropes/Scaffolding	98.2%	96.3%	0.51	<.01	0.95

Table 3 presents multiple measures that allow us to assess agreement among field economists. Column 2 presents the level of absolute agreement among field economists. These values range from a low of 79% for pushing and/or pulling with hands and arms to 98% for crawling and climbing ladders. As can be seen in Column 3, however, the expected levels of agreement are relatively high and the kappa statistics are relatively low.[8] Averaging across the elements, kappa is 0.43, which falls in the "moderate" range. It is worth noting that all but one of the agreement measures is statistically higher than expected agreement in a 5 percent one-tailed test; the lone exception is pushing and/or pulling with feet.

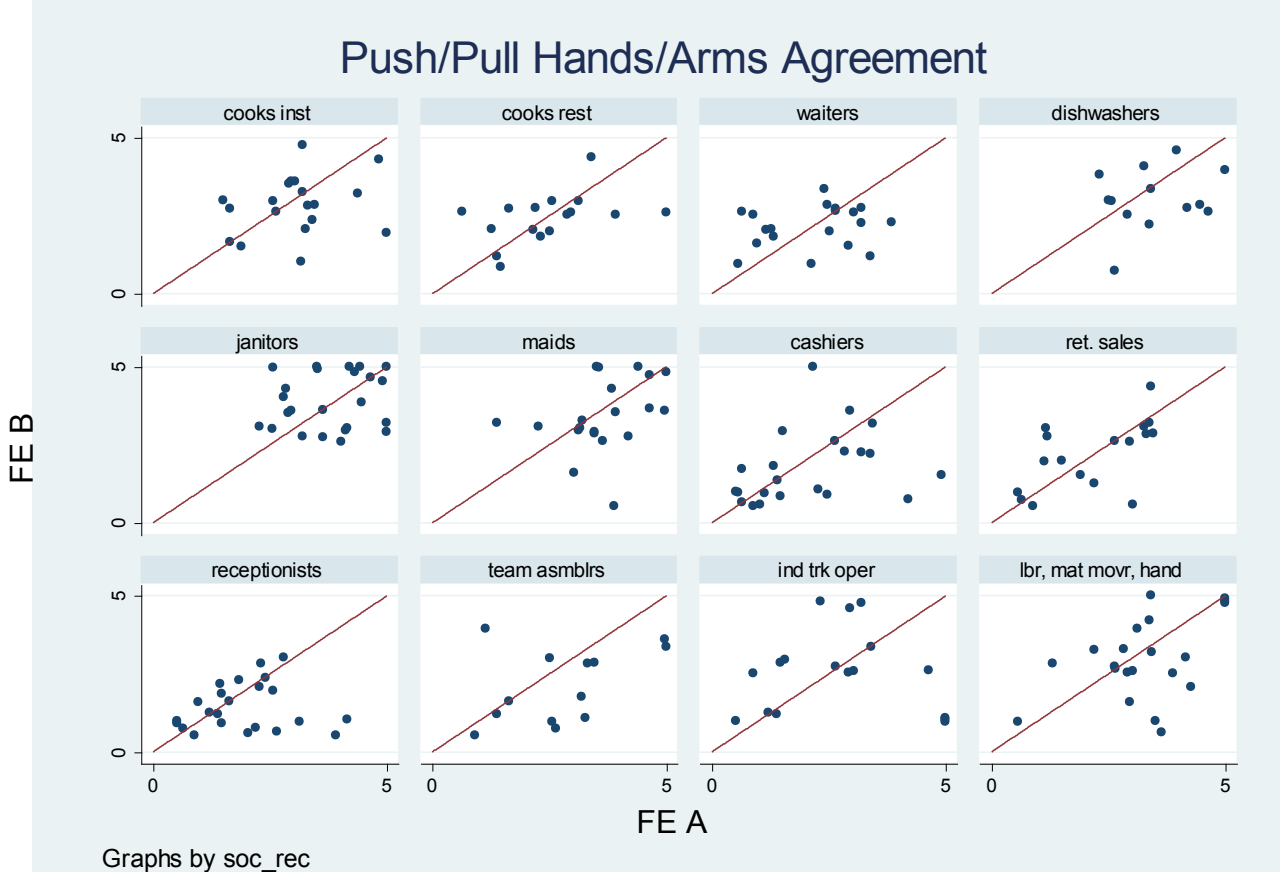
[8] Measures of expected agreement, kappa statistics, and the standard errors needed to compute p values were calculated using Stata (version 12). Stata's calculations are based on Landis and Koch (1977b) and standard errors are based on Fleiss, Nee, and Landis (1979)

A well-known issue with kappa is the influence of prevalence and bias on the kappa measures. Generally, categories with underlying uniform distributions will result in higher values of kappa. The distributions of the physical elements in ORS, however, tend to be highly skewed (both in terms of the underlying discrete values of the duration as well as the categorical measures). For example, in the jobs selected for observation, crawling is very uncommon in the sampled jobs while gross manipulation is present in almost all cases with durations in the frequent and constant ranges. Measuring kappa using data that have skewed prevalence can give rise to the "kappa paradox," where high levels of rater agreement have relatively low kappa statistics (Feinstein and Cicchetti 1990 and Cicchetti and Feinstein 1990).

To account for this, a measure of prevalence and bias adjusted kappa (PABAK) was used. The PABAK measure is presented in the final column of Table 3. Across all of the elements analyzed, the average value is 0.76, in the "substantial" range. We are particularly interested in the elements with lower PABAK values. These are reaching at or below the shoulder, fine manipulation, pushing and/or pulling with hands and arms, and pushing and/or pulling with feet and legs.

Given that there is a diverse set of occupations in our sample, we first examine whether the levels of agreement differ systematically by occupation. We explore this visually by creating plots of the FE duration measures by occupation. Figure 1 presents the graphs for pushing and pulling with hands and arms.[9]

Figure 1: Scatterplots of Inter-rater Agreement for Pushing or Pulling with Hands and Arms



[9] A full set of plots is available from the authors upon request.

The 45 degree line provides a reference line for perfect agreement, where both FEs' duration measures fall into the same category. Points substantially off the diagonal line represent major disagreements in the duration ranking. Since there are only five categories, the plots are rearranged so the individual points can be seen rather than superimposed within the category.

For most occupations there are few cases of the ratings differing by more than one category (ex. frequent versus constant), with some exceptions. Industrial truck and tractor operators and team assemblers had relatively few observations lying on the reference line. The plot of receptionists also displays an interesting pattern. There are multiple observations in the lower right hand of the chart, indicating multiple cases where FE B recorded no incidence of pushing or pulling with hands and arms and FE A recorded the duration as occasional or frequent. The debriefs with the FEs suggests that this may be due to different interpretations of the threshold of what constitutes pushing and pulling (e.g., does grabbing a piece of paper across the desk constitute pulling with hands and arms?).

Given that there appear to be differences in agreement by occupation, we run ordered logistic models of agreement for the four elements with low PABAK values with occupation as the explanatory variable. The dependent variable is a measure of the disagreement between FEs (ranging from 0 to 4). For pushing and pulling at or below the shoulder we find the highest levels of disagreement (evidenced by positive coefficients that were statistically significant in a 5% one-tailed test) among institutional cooks, dishwashers, and maids. For fine manipulation the highest level of disagreement is among retail sales workers. For pushing and pulling with hands and arms, the highest level of disagreement is among industrial truck and tractor operators. Pushing and pulling with feet and legs had relatively high levels of disagreement among all occupations except cashiers and receptionists.

The conclusion is that there does not seem to be a statistical pattern among occupations (or occupation types) that would explain the disagreement levels across all elements. Rather, additional training on the elements across all occupations is likely needed, with particular focus on the definitions and thresholds associated with each element.

The final issue we evaluate for inter-rater agreement is whether some of the disagreement found in reaching at or below the shoulder may be explained by different interpretations by the FEs of the level of reaching – whether it is above the head or at shoulder level and below. To examine this we superimpose the reaching overhead onto the reaching at or below the shoulder graph for maids and housekeepers. 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 2. Scatterplot of Reaching Overhead Versus Reaching At Or Below Shoulder Agreement

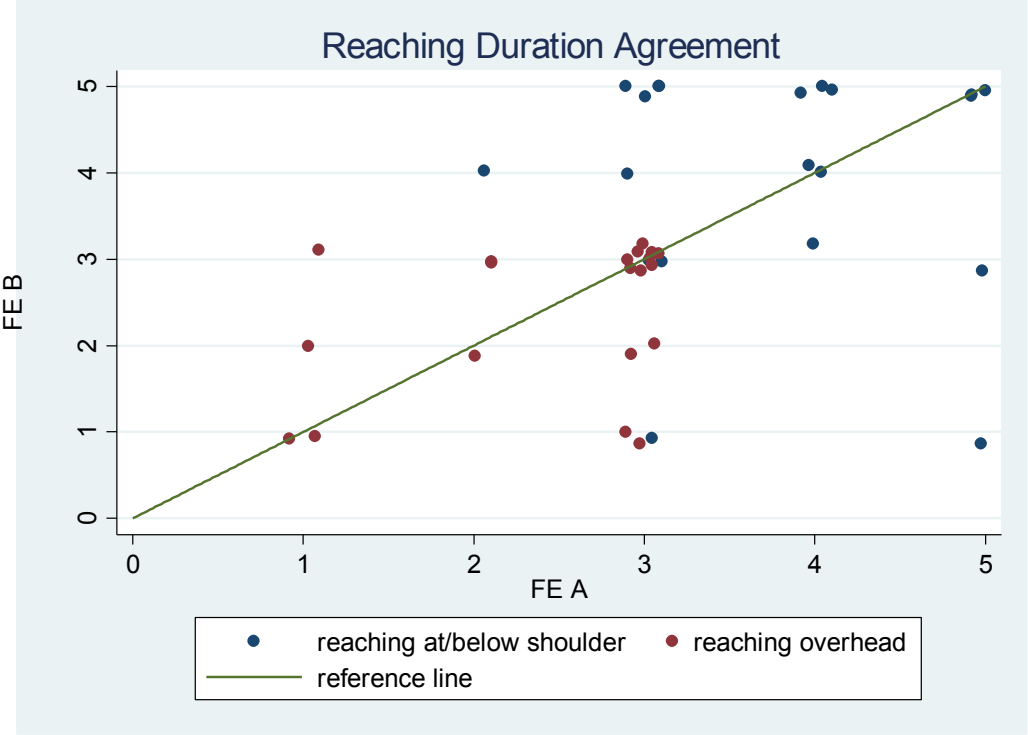


Figure 2 depicts several cases where reaching at or below the shoulder (the blue dots) is recorded as constant by FE B but only as occasional or frequent by FE A. When we superimpose reaching overhead on the same graph (the red dots) we see that FE B is more likely to record not present/seldom when FE 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.

In sum, we find high levels of "raw" agreement among raters and PABAK measures averaging 0.76, indicating substantial levels of rater agreement. For elements with lower levels of agreement we recommend additional training before future observation tests, with a particular focus on the element definitions and thresholds.

Agreement Between Pre-production and Observed Measures

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 "interview" values hereon. Measuring agreement between observed and interview 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 "almost never."
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 interview data has missing duration in nearly 30% of the cases for "communicate verbally" and 25% of the cases for "fine manipulation."

To deal with the challenges posed by the short duration of the job observation, we re-categorize the durations into four categories, aggregating not present and seldom into one category:

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

For most observed jobs we have pairs of observations (from two FEs) that we can compare with the interview 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 maximum of the two values, and the minimum 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.

Next, we move to measures of agreement in the duration of the physical elements. We use the weighted kappa 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: Percent Agreement and Kappa Measure of Agreement Between Observed and Interview Data for Duration

ORS Element	Agreement	Expected Agreement	Kappa	Prob>Z	PABAK	PABAK std error
Crawling	97.1%	97.1%	-0.01	0.58	0.96	0.03
Crouching	79.3%	77.8%	0.07	0.18	0.63	0.04
Kneeling	87.7%	85.4%	0.16	<.01	0.78	0.04
Stooping	74.0%	71.7%	0.08	0.02	0.38	0.04
Reaching overhead	84.3%	81.0%	0.18	<.01	0.62	0.04
Reaching At/Below Shoulder Level	71.2%	67.3%	0.12	<.01	0.31	0.04
Communicating Verbally	75.6%	67.3%	0.25	<.01	0.41	0.04
Keyboarding	92.1%	81.5%	0.58	<.01	0.81	0.04
Keyboarding- Touchscreen	93.9%	88.4%	0.47	<.01	0.85	0.03
Keyboarding- 10key	96.5%	94.9%	0.32	<.01	0.94	0.03
Keyboarding- Other	95.9%	94.8%	0.20	<.01	0.90	0.03
Fine Manipulation	76.7%	71.4%	0.19	<.01	0.44	0.04
Gross Manipulation	76.3%	70.5%	0.21	<.01	0.44	0.04
Pushing/Pulling with Hands and Arms	73.6%	66.6%	0.21	<.01	0.37	0.04
Pushing/Pulling with Feet and Legs	79.8%	76.0%	0.16	<.01	0.52	0.04
Pushing/Pulling with Feet	97.4%	97.4%	-0.01	0.57	0.94	0.03
Climbing Ramps/Stairs	89.8%	88.6%	0.11	0.05	0.82	0.04
Climbing Ladders/Ropes/Scaffolding	95.8%	94.9%	0.17	<.01	0.92	0.04

Much like the inter-rater agreement measures, we find high levels of agreement among the methods of data collection. The weighted kappa statistics vary considerably. With the exception of crawling, crouching, and pushing and/or pulling with feet, the agreement levels are greater than the expected levels of agreement in a 5% one-tailed test. The average value of the kappa statistic is 0.20, which denotes relatively low levels of agreement relative to agreement by chance.

As discussed previously, the kappa statistic is affected by prevalence and most of these measures have a skewed distribution (both in the categories and in the underlying measures of the continuous variables). We present the weighted PABAK measures in the final column of Table 4. The average is 0.68, which is considered "substantial;" however, there is considerable variation in the PABAK measure by data element. In particular, stooping, reaching at or below the shoulder, communicating verbally, fine manipulation, pushing and/or pulling with hands and arms, and pushing and/or pulling with feet and legs have low measures, even after adjusting for prevalence and bias.

We want to test whether the distributions of the variables are different between modes of collection. We first assess this using a Wilcoxon Rank Test, which tests the null hypothesis that both distributions are the same. These results are presented in column 2 of Table 5. It is a two-tailed test – rejection of the test rejects the hypothesis that the two variables come from the same underlying distribution. If we use a 5 percent threshold, the test is rejected for stooping, reaching at or below the shoulder, other keyboarding, fine manipulation, gross manipulation, pushing and/or pulling with hands, and arms and pushing and/or pulling with feet. These results are not surprising, given the PABAK statistics.

Of particular concern, given the potential uses of ORS data in the disability determination process, is whether the pre-production data appear to understate the duration of the physical elements. We evaluate this using a sign test. The sign test is a test of the difference in medians. We are particularly interested in elements where the sign test rejects the null hypothesis that the observed median is less than or equal to the interview median – rejecting this implies that the collected data are generally distributed with longer duration than the interview data (see column 3 of Table 5).

Table 5: Wilcoxon Rank Test and Sign Test.

	Wilcoxon Rank Test	Sign Test	
ORS Element	Wilcoxon p-value	Ho: p value observed ≤ interview	Ho: p value collected ≥ interview
Crawling	1.00	0.66	0.66
Crouching	0.86	0.50	0.59
Kneeling	0.08	0.97	0.05
Stooping	<.01	<.01	1.00
Reaching overhead	0.57	0.81	0.25
Reaching At/Below Shoulder Level	<.01	<.01	1.00
Communicating Verbally	0.07	0.95	0.07
Keyboarding	0.78	0.68	0.44
Keyboarding- Touchscreen	0.56	0.35	0.79
Keyboarding- 10key	0.41	0.29	0.87
Keyboarding- Other	0.01	<.01	1.00
Fine Manipulation	<.01	<.01	1.00
Gross Manipulation	<.01	<.01	1.00
Pushing/Pulling with Hands and Arms	<.01	<.01	1.00
Pushing/Pulling with Feet and Legs	0.21	0.22	0.84
Pushing/Pulling with Feet	0.02	0.02	1.00
Climbing Ramps/Stairs	0.07	0.98	0.05
Climbing Ladders/Ropes/Scaffolding	0.14	0.96	0.21

The variables with longer duration associated with observation are stooping, reaching at or below the shoulder, other keyboarding, fine manipulation, gross manipulation, pushing and/or pulling with hands and arms, and pushing and pulling with feet. When we measure the modes of these elements, only one shows a difference in mode between the collected and observed values – reaching at or below the shoulder. The value of the mode for this element is occasionally (2% up to one-third of the day) in the interview data and constantly in the job observation data (two-thirds or more of the day).

We noted earlier that missing duration was identified as an issue with ORS pre-production. In the case of reaching at or below the shoulder, 53 of the job observation duration measures were unable to be compared with interview duration data due to missing duration. It is notable that among the 53 missing quotes in pre-production, the job observation test recorded durations of frequently or constantly in 64% of the quotes. This is a common pattern among those elements where the sign test rejected the null of observation duration equal or below pre-production – the missing data in pre-production align with observed durations above the median and mode.

From this, it appears that the "underestimate" of duration from the interview 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 with unknown duration in pre-production and only one of these was classified as frequently or constantly in the observation test.

Summary and Conclusions

The purpose of the job observation pilot 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 (measured using prevalence and bias adjusted kappa) among the FEs, suggesting that future observation tests could be done with single observers, with adequate training on definitions and thresholds of the elements.

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 duration ("seldom"). The prevalence adjusted kappa measures of duration are relatively strong, suggesting that the interview 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 interview duration, leading to estimates that may under or overstate the frequency of a physical element. However, these conclusions must be qualified by the limited observation period of 60 minutes.

References

Cain & Green, 1983. Reliabilities of Selected Ratings Available from the DOT. *Journal of Applied Psychology,* 155-165.

Campbell L, Pannett B, Egger P, Cooper C, Coggon, 1997. Validity of a questionnaire for assessing occupational activities. *D. Am J Ind Med.*

Cicchetti DV and AR Feinstein, 1990. High agreement but low kappa II: Resolving the paradoxes." *The Journal of Clinical Epidemiology* 43: 543-549.

Descatha et al. 2008. Self-administered questionnaire and direct observation by checklist: comparing two methods for physical exposure surveillance in a highly repetitive tasks plant. *Applied Ergonomics,* 194-198.

Feinstein AR and DV Cicchetti. 1990. High agreement but low kappa I: The problem of two paradoxes." *The Journal of Clinical Epidemiology* 43: 543-549.

Fleiss, J. L., J. C. M. Nee, and J. R. Landis. 1979. Large sample variance of kappa in the case of different sets of raters. Psychological Bulletin 86: 974–977.

Foster, M.R., 1998. Effective job analysis methods. *Handbook of human resource management in government.*

Halgren, Kevin H., 2012. Computing Inter-rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology***,** 8(1), 23-34.

Kilbom 1994. Assessment of physical exposure in relation to work-related musculoskeletal disorders - what information can be obtained from systematic observations? *Scand J Work Environ Health,* 30-45.

Landis JR and Koch GG. 1977a. The measurement of observer agreement for categorical data. *Biometrics*. 33:159–174.

Landis JR and Koch GG. 1977b. A one-way components of variance model for categorical data. *Biometrics*. 33:671–679.

Lysaght, R & Shaw, L, 2010. Job Analysis: What it is and how it is used. *International Encyclopedia of Rehabilitation*. Available online: http://cirrie.buffalo.edu/encyclopedia/en/article/268/

Martin, McCabe, 2001. Lost time injuries: Demographic variables, self-reports, and observational assessment of occupational demands. *The Impacts of Social and Technological Change on Work, Health, and Safety*, 177-182.

Nordstrom et al 1998. Comparison of Self-Reported and Expert-Observed Physical Activities at Work in a General Population. *American Journal of Industrial Medicine*, 29-35.

Stock, Fernandes, Delisle, Vezina, 2005. Reproducibility and validity of workers' self-reports of physical work demands. *Scand J Work Environ Health*, 409-437.

Spielholtz et al. 2001. Comparison of self-report, video observation and direct measurement methods for upper extremity musculoskeletal disorder physical risk factors. *Ergonomics*, 588-613.

Winnemuller, Spielholtz, Kaufman, 2004. Comparison of ergonomist, supervisor, and worker assessments of work-related musculoskeletal risk factors. *J Occup Environ Hyg,* 414-422.