

Responses to Harvey's Criticisms of HumRRO's Analysis of the O*NET Analysts' Ratings

Suzanne Tsacoumis, HumRRO

April 2009

Robert J. Harvey, in his paper *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?*, criticized several aspects of the methods HumRRO has used to evaluate O*NET analysts' ratings of abilities and skills. The purpose of this document is to address some of his comments associated with HumRRO's work in this area.¹

Background

The information that populates the O*NET database is collected primarily from three sources: incumbents, occupational experts, and analysts. Targeted job incumbents provide ratings on occupational tasks, generalized work activities (GWA), knowledge, education and training, work styles, and work context. Importance and level information regarding the abilities and skills associated with these occupations are collected from analysts. There are theoretical and philosophical reasons for preferring one rater group to the other for collecting different types of data. For example, incumbents are generally more familiar with the day-to-day duties of their job; therefore, they are the best source of information regarding tasks and GWAs, as well as the context within which those tasks are completed. In contrast, it is likely that trained analysts understand the ability and skill constructs better than incumbents and therefore should provide the ability and skills data (Tsacoumis, 2007).

To evaluate the ratings that analysts provided, we performed four sets of analyses. First, we identified data that might be difficult to interpret because of limited agreement among raters or because there was an indication that the ability/skill level rating was not relevant for a specific occupation. Thus, we established a set of recommended suppression criteria that flagged any ability/skill having (a) a level rating deemed not relevant to an occupation because of low importance ratings, (b) too little agreement in importance ratings across raters for a particular occupation, [or](#) (c) too little agreement in level ratings across raters for a particular occupation.

The remaining three sets of analyses focused on computing measures of interrater agreement and interrater reliability. Poor agreement or reliability estimates might indicate confusion about the constructs, potentially due to either the nature of the construct definition or to rater training. Specifically, the second analysis involved computing the interrater agreement among the eight analysts for a given ability/skill. Next, the interrater reliability of the raters was computed to determine the similarity of the ratings with regard to the order of and relative distance among *constructs* on a particular scale (i.e., importance or level) within a particular occupation. This analysis provides information regarding the consistency across raters in terms of how they rated the 52 ability and 35 skill constructs with regard to the (a) level of ability/skill required to perform the occupation, or (b) relative importance of the ability/skill to performance in a particular occupation. The analyses of the ability and skill constructs were computed separately.

¹ Our choice not to comment on other elements of Harvey's paper should not be interpreted to mean that HumRRO agrees with the views expressed therein or the method used in other parts of his paper.

Finally, another interrater reliability estimate was computed to examine the consistency of ratings across occupations within constructs. This estimate of interrater reliability was computed for each ability and skill to determine the consistency with which raters rank-ordered occupations with regard to the importance/level of a given ability/skill.

Points of Contention

Issue 1

Harvey criticizes our use of ICC3:

*“Of course, one can question whether this arbitrary .80 cutoff is sufficiently stringent (given the findings reported for r_{wg} by Harvey & Hollander, 2002), and particularly, whether O*NET should have instead reported results using the ICC Case 1 or 2 formulas. That is, the Case 3 formula they used (which also tends to produce the numerically highest results; see Shrout & Fleiss, 1979, p. 425) assumes that raters are a “fixed effect” in an ANOVA sense; specifically, that there is no desire to generalize the reliability results to the larger population of raters who might use the O*NET scales, that the exact same set of raters is used to rate all occupations, and that those judges are the only ones of interest (Shrout & Fleiss, 1979, p. 421).*

*Unless O*NET indeed used a single set of judges to rate every one of the occupations rated in the nine waves of data collection described by Willison and Tsacoumis (2009), the Case 1 formula arguably should have been used (i.e., Case 1 assumes different judges may rate each target, whereas Case 2 assumes a randomly sampled set of raters, but that the same raters rate all OUs). ” (pg. 22)*

We stand by our choice of analytic techniques. ICC Case 1 is for a nested design; it assumes we have a unique, non-overlapping set of raters for each object of measurement (i.e., occupation). With O*NET ratings, raters are not nested within the objects of measurement (occupations). Rather, blocks of raters rate blocks of occupations (i.e., a unique, *non-overlapping* set of raters does not rate each occupation).

Similar to Case 1, Case 2 is also an *agreement* coefficient, whereas Case 3 is a *consistency* coefficient. We are interested in making statements regarding the consistency of the raters' relative rank ordering of (a) abilities/skills across occupations and (b) occupations across abilities/skills. We are not trying to assess absolute agreement among those raters. Hence, we are justified in our use of a consistency coefficient (i.e., ICC Case 3). ICC Case 2 yields a lower estimate of reliability than does Case 3 because the former treats rater main effects as error variance.

It is true that the ICC Case 3 formula treats raters as a “fixed effect.” However, the estimate we would get if we were to treat raters as a “random factor” (in an ANOVA sense) would be equivalent to the estimate we would get when treating raters as a “fixed factor” (this can easily be confirmed using SPSS). Specifically, we will get identical estimates of interrater reliability regardless of whether we calculated ICC(3, k), or whether we treated raters as a random factor and calculated what McGraw and Wong (1996) label ICC(C, k). Thus, we feel the results we report can be interpreted as informing the generalizability of ratings to other randomly sampled sets of k raters.

With regard to judges (raters) being the only ones of interest, this is the same issue raised regarding the treatment of raters as “fixed” vs. “random” factors. It is correct that formula for the Case 3 ICC assumes the exact same set of raters is used. In the instance in which we were interested in the consistency of raters’ rank-ordering of abilities or skills within each occupation, the design *was* fully crossed (i.e., each ability or skill within a given occupation was rated by the same set of eight raters). In the instance in which we were interested in the consistency of raters’ rank-ordering of occupations within each ability/skill, the design *was not* fully crossed; rather, within any given data collection “cycle,” two sets of eight raters rated two different blocks of occupations. The two sets of raters used within a cycle were not necessarily the same across cycles. This begs the question of whether using ICC(3,*k*) was appropriate in this latter instance given that occupations (the objects of measurement) were not fully crossed with raters. Putka, Le, McCloy, and Diaz (2008) recently illustrated a disconnect between traditional methods for estimating interrater reliability (e.g., Pearson correlations, ICCs) and the measurement designs often confronted in practice. They introduced a generalized interrater reliability coefficient [G(*q,k*)] that is appropriate regardless of whether one is dealing with a fully crossed, a fully nested, or what they termed an *ill-structured measurement design* (i.e., a design in which the sets of raters who rate each object of measurement vary in their degree of overlap). The publication of Putka et al. (2008) post-dates the bulk of the O*NET analysis work described above, but results of their simulation suggest that our use of ICC(3,*k*) would not result in an inflated estimate of interrater reliability in the case of the O*NET data (as Harvey suggests). Rather, for analyses where occupations are treated as objects of measurement, the resulting ICC(3,*k*) estimates slightly *underestimate* the interrater reliability of the mean ratings.

Issue 2

Harvey criticizes the fact that we look at the k-rater reliability estimate rather than just the single-rater estimate:

*“...however, after adjustment to estimate the reliability of a mean from eight raters (the approach used by O*NET; e.g., Willison & Tsacoumis, 2009), the median ICC(3,8) estimate increases to an apparently respectable .70” (pg. 10).*

*“O*NET prefers to interpret the stepped-up values produced with the ICC(3,8) formula (i.e., to estimate the reliability of a mean from 8 raters); using this Spearman-Brown type of correction, reliability estimates are considerably enlarged, ranging from .59 to .96 (median = .90)” p. 22.*

Harvey implies we did something inappropriate by focusing on the *k*-rater (*k*=8) reliability (i.e., the expected reliability of the mean rating based on eight raters) for the ability/skill ratings, as opposed to the single-rater reliability. Nevertheless, it is the reliability of the mean rating that is more relevant in this case. Users of the O*NET data will not make the decision regarding the importance of an ability or skill to an occupation based on the ratings of a single-rater, nor should they, given that the reliability of ratings provided by any single rater is going to be relatively low. This is precisely why we gather O*NET ratings from multiple raters (in this case, eight of them) and rely on the mean rating from those raters to estimate the importance of an ability or skill to an occupation. Given that O*NET databases report the mean ratings profile rather than the ratings from a single-rater, it is the reliability of the mean ratings that are most relevant to the O*NET user. Thus, the reported estimate is “preferred” because it is more relevant. In addition, it follows that rather than “stepping up” our single-rater estimate to obtain

the reliability estimate for each 8-rater mean profile (as Harvey states), we actually “stepped down” the 8-rater reliability estimate to obtain the single-rater estimate.

Issue 3

Harvey (in several forums) suggests that the resulting analyst agreement and reliability estimates are not acceptable.

First, it should be noted that there seems to be some confusion regarding interrater *reliability* versus interrater *agreement* (Tinsley & Weiss, 1975). Interrater agreement was computed to examine the level of absolute agreement among the analysts in ratings within a construct for a particular occupation. These indices identified the extent to which the eight raters provided the same numeric rating regarding the level of a particular ability/skill dimension (e.g., *Written Comprehension*) they deemed necessary to perform a particular occupation. To look at interrater agreement, we calculated the standard deviation (*SD*) of ratings across analysts for a given construct and scale for each occupation, as well as the *SE_M* of these ratings. For both indices, lower values indicate higher agreement, and vice versa.

In terms of the Cycle 9 ability ratings (the Willison and Tsacoumis [2009a] report to which Harvey refers), the importance ratings across all occupations had a median *SD* of 0.52 and a median *SE_M* of 0.18. The level ratings across occupations had a median *SD* of 0.67 and a median *SE_M* of 0.24. These results are consistent with the agreement results from Cycles 1-8. In terms of the skill constructs, the importance and level ratings across all occupations had a median *SD* of .52 and .71, respectively and a median *SE_M* of .18 and .25, respectively (Willison & Tsacoumis, 2009b). Overall, although the values are generally greater (indicating less agreement) for the level ratings than they are for the importance ratings, the results indicate that the ratings made by the analysts were quite similar in magnitude (i.e., strong interrater agreement) for both scales.

In terms of *interrater reliability*, as noted above, we computed two indices: (a) across constructs within occupations, and (b) across occupations within constructs. To examine the interrater reliability of the Cycle 9 ratings, we calculated the intraclass correlations (ICC[3,*k*]; Shrout & Fleiss, 1979) among the analysts' ratings to look at consistency across constructs within occupations. As mentioned previously, this calculation examines the similarity in the rank ordering and relative distance between the abilities on a particular scale within an occupation. The same analyses were conducted using the skill constructs. Although Harvey mentions a criterion of .70, our target level of interrater reliability is a median ICC(3,*k*) of .80 or greater. The value of .80 is judged to be a good rule-of-thumb that has been used previously in the O*NET context (e.g., McCloy et al., 1999).

The data revealed high levels of interrater reliability across the 106 Cycle 9 occupations (Willison & Tsacoumis, 2009a and 2009b). Specifically, the mean and median ICCs for importance ratings for the abilities (calculated across occupations) were .95 and .96 (*SD* = .03), respectively. Both the mean and median ICC for skill importance across occupations was .95 (*SD* = .02). The mean and median ICCs for the ability level ratings were .95 and .96, respectively (*SD* = .03) and were .95 for both the mean and median ICC for the skill level ratings (*SD* = .03). The reliability of both the importance and level ratings exceeded the median coefficient target of .80. Results also indicate that, for the most part, occupations with the lowest reliability estimates for importance had the lowest values for level ratings. This might be due to the skip pattern

which forces a “0” for level if the ability is rated “Not Important.” Overall, the results support a good level of consistency in the analysts’ ratings.

Another effective way to evaluate the reliability of the analysts’ ratings is to look at the consistency across occupations within constructs. This type of reliability is the extent to which raters agree about the order of and relative distance among 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 ability *Oral Comprehension*? To make this evaluation, we calculated ICC(3,*k*) for each construct on each scale (instead of for each occupation on each scale as described above). For example, each of the 52 ability importance scale ratings has a reliability value. The target level of interrater reliability for this coefficient is that the median ICC(3,*k*) across the construct ratings for a particular domain on a particular scale be .80 or greater (e.g., the median reliability across 52 ability level ratings should be at least .80).

The median ICC(3,8) across the ability ratings was .87 ($M = .84, SD = .11$) for importance and .90 ($M = .87, SD = .09$) for level (Willison & Tsacoumis, 2009a). In terms of the skill ratings, the median ICC(C,8) for importance was .85 ($M = .85, SD = .06$) and for level was .87 ($M = .87, SD = .05$) (Willison & Tsacoumis, 2009b).

Issue 4

Harvey suggests that aggregation bias impacts our reliability estimates.

*“...James (1982) showed that when individual rating profiles are correlated with an aggregate profile, even when substantial disagreement exists between the raters, correlations with the aggregate may be much higher than correlations between the rater-level profiles used to compute it. The relevance of this single-group aggregation bias issue to the problem of interpreting O*NET rater-OU rs is straightforward: if ratings of known poor quality can be aggregated to produce a situation where they correlate at an apparently acceptable level with the aggregate, it must be concluded that such correlations overstate the true quality of the ratings, and that a higher cutoff for assessing ‘adequate’ agreement must be set” (pp. 10-11).*

What Harvey is labeling as aggregation bias (at least in the context of reliability estimation) is a principal we have known in psychometrics since the early 1900s—as you add more replications to a given measurement procedure (e.g., items, raters, etc), the reliability of the mean of those replications is going to increase (Brown, 1910). Historically, we have not concerned ourselves with the reliability of any single replicate of a measurement procedure (e.g., a single item or rater), but rather have focused on the reliability of the mean, realizing each replicate in and of itself is an imperfect measure.

Harvey liberally cites Larry James’s 1982 paper, but he does not appear to be referencing the entire paper. First, it is critical to note that James was talking about whether it is sensible to calculate mean scores for organizational climate if individuals’ perceptions of climate disagreed with one another substantially. In essence, the argument is that with large levels of disagreement among individual raters/perceivers within an organization, there is no single, easily perceived climate at all. Thus, a low level of agreement (say, $ICC(1) = .05$), which might lead to a high $ICC(2)$ given enough raters (.94, if based on 300 raters), is irrelevant because it is nonsensical to think that the .94 gives you a meaningful climate construct, given the large disagreement among

the raters within each organization (i.e., they clearly do not share perceptions about a uniform organizational climate).

Below is an excerpt from James (1982) that appears to be omitted from Harvey's argument. We made substitution edits in red to parallel the current O*NET situation.

"It must be emphasized that the conclusions above do not necessarily generalize to perceptions of other types of variables. In fact, for some types of research a high level of agreement at the individual level may not be assumed or needed; reliability at the aggregate level [e.g., a high ICC(2)] may be all that is required. For example, suppose we ask nk individuals in each of K organizations {occupations} to describe an organizational {occupational} characteristic such as procedures for performance evaluation (e.g., frequency of evaluation and types of procedures used to evaluate performance) {importance of physical abilities for the occupation}. For illustrative purposes, it is assumed that (a) the procedures {abilities} used in a particular organization are the same for all individuals, and (b) the organizations {occupations} vary in regard to procedures used {importance of physical abilities} (an implicit assumption for any type of ICC statistic; cf. Ebel, 1951). One might now argue that the reliability of mean differences is the appropriate statistic on which to decide whether to aggregate perceptions. This argument is based on the logic that the mean perception per organization {occupation} contains less error variance (i.e., is more reliable) than the perception of a single rater. In other words, in comparison to a single rater, the mean of a set of perceptions is the more accurate representation of the true score for the target of perception (i.e., the organization) on the perceptual variable (i.e., procedures for performance evaluation) (cf. Jones & James, 1979; Lord & Novick, 1968).

Note that in the scenario above it is assumed that each organization {occupation} has a true score on the perceptual variable, which seems reasonable for an organizational {occupational} characteristic. Moreover, intraorganizational {intraoccupational} variation in perceptions contributes to the error term, but is not of substantive interest other than that the larger the error term, the lower the reliable differentiation among organizational means. This, however, may generally be compensated for by increasing nk [see equation for ICC(2)]" (James, 1982, pp. 222-223).

In sum, we believe Harvey is misapplying James's warning concerning aggregation bias by overgeneralizing to a situation where the reliability at the aggregate level (i.e., of the mean profile) is of paramount importance.

References

- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296-322.
- Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.
- James, L.R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology*, 67, 219-229.
- McCloy, R., Waugh, G., Medsker, G., Wall, J., Rivkin, D., & Lewis, P. (1999). *Determining the Occupational Reinforcer Patterns (ORPs) for the O*NET occupational units*. Raleigh, NC: National Center for O*NET Development.
- McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30-46.
- Putka, D.J., Le, H., McCloy, R.A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, 93, 959-981.
- Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420-428.
- Tinsley, H.E.A., & Weiss, D.J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22, 358-376.
- Tsacoumis, S. (2007). *The feasibility of using O*NET to study skill changes*. Prepared for the National Academies Center for Education, Washington, D.C.
- Willison, S., & Tsacoumis, S. (2009a). *O*NET analysis occupational abilities ratings: Analysis cycle 9 results (FR-08-58)*. Alexandria, VA: Human Resources Research Organization.
- Willison, S., & Tsacoumis, S. (2009b). *O*NET analyst occupational skills ratings: Analysis cycle 9 results (FR-08-63)*. Alexandria, VA: Human Resources Research Organization.