Harvey, R. J., & Hollander, E. (2004, April). *Benchmarking r_{WG} interrater agreement indices: Let's drop the .70 rule-of-thumb*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Chicago.

Benchmarking r_{WG} Interrater Agreement Indices: Let's Drop the .70 Rule-Of-Thumb

Robert J. Harvey and Eran Hollander
Virginia Tech

Variance-based interrater agreement indices in the r_{WG} family are often interpreted using rules-of-thumb derived for reliabilities (e.g., ≥ .70 = acceptable). Monte Carlo results suggest that far more stringent standards are needed, especially for maximum-variance r_{WG}, as values > .70 can routinely be obtained from totally *random* ratings.

Researchers have developed a range of indices sensitive to pattern-based ratings similarity (e.g., average interrater *r*) and/or level-based agreement (e.g., see Dunlap, Burke, & Smith-Crowe, 2003, Mitchell, 1979; Shrout & Fleiss, 1979; Tinsley & Weiss, 1975). The popular r_{WG} family of indices takes the latter approach, evolving from the original (e.g., James, Demaree, & Wolf, 1984) to include several derivatives (e.g., Brown, 2002; Lindell, 2001; Lindell & Brandt, 1997, 1999; Lindell, Brandt, & Whitney, 1999); all quantify agreement as a function of observed cross-rater variance relative to an error-term defining an *unacceptable* baseline of disagreement.

Original r_{WG} (denoted as r_{WGu}) uses a uniform distribution on *A* points as the error variance:

$$r_{WG} = 1 - \frac{s_x^2}{s_{EU}^2} \tag{1}$$

$$s_{EU}^2 = (A^2 - 1) / 12 \tag{2}$$

Here, r_{WGu} = 1 denotes perfect agreement, whereas 0 reflects a white-noise-style absence of agreement; however, because one can find situations in which *systematic* disagreement (e.g., half of raters rate high, half low) exceeds s_{EU}^2, negative r_{WGu}s can result. Although not a flaw in one sense (i.e., it simply indicates that disagreement exceeds the error baseline), because James et al. (1984) and others initially viewed r_{WG} as index of *reliability,* negative values were considered "improper" (e.g., see Lindell et al., 1999, pp. 131-132) and efforts were made to make r_{WG}'s range more closely match that of the classical test theory (CTT) reliability coefficient (i.e., 0-1).

The derivative index we denote as r_{WGmax} (e.g., Lindell & Brandt, 1997) was one result; r_{WGmax} increased the error term (relative to r_{WGu}) via the use of a *maximum-variance* estimate, s_{MV}^2 (i.e., the variance in a bimodal-extreme distribution in which half of the raters provide the lowest- and highest-possible ratings, which is clearly *not* an *unsystematic* type of disagreement):

$$s_{MV}^2 = .5(H^2 + L^2) - [.5(H + L)]^2 \tag{3}$$

(H = highest, L = lowest). Brown's a_{wg} (2002) is similar, using conditional maximum-variance to reflect the fact that as the mean departs from the midpoint, the maximum variance is reduced:

$$a_{wg(1)} = 1 - \frac{s_x^2}{[(H+L)M_x - (M_x^2) - (H*L)]*[(k/k-1)]} \tag{4}$$

(M_x = mean, *k* = raters). Conceptually, a_{WG} is similar to r_{WGmax}, but with a situationally-sensitive error term; unlike original r_{WGu}, however, both express agreement relative to a baseline of *worst-imaginable* disagreement. Multiple-item r_{WG}s were also developed (e.g., see Lindell, 2001).

Interpreting the Magnitude of r_{WG}

Controversy surrounds the use of rater agreement indices (e.g., see Cohen, Doveh, & Eick, 2001; Dunlap et al, 2003; Klein, Conn, Smith, & Sorra, 2001; Kozlowski & Hattrup, 1992; Murphy & DeShon, 2000a, 2000b; Schmidt & Hunter, 1989; Schmidt, Viswesvaran, & Ones, 2000); we focused on the issue of interpreting the *magnitude* of r_{WG}. Two main strategies can be identified: (a) rules-of-thumb designed to identify minimum-acceptable agreement (e.g., using cutoffs advanced for reliabilities); and (b) statistical significance tests (e.g., Dunlap et al., 2003).

Rules-of-thumb based on CTT. Given Equations 1-4 and prior research (e.g., see Lindell, 2001, pp. 93-94), it is indisputable that the choice of error term strongly influences the scale or *metric* on which r_{WG} is expressed, and that r_{WG}s (e.g., r_{WGmax} and a_{WG}) that use larger errors will be numerically larger than r_{WGu} using s_{EU}^2. It is therefore puzzling why many researchers and practitioners continue to rely on arbitrary rules-of-thumb to interpret r_{WG}, especially the popular rule-of-thumb stating that $r_{WG} \geq .70$ denotes acceptable agreement. Although not peculiar to r_{WG} (e.g., cross-rater *r*s in the .70's and up were deemed "substantial correlations" by Fleishman & Mumford, 1991, p. 543), the practice of viewing r_{WG} in the .70's and higher as representing acceptable

convergence is widespread. For example, Dirks (2000, p. 1007) noted that "because it is necessary to determine whether aggregation is empirically justifiable, …. I computed R_{wg} … which was an acceptable .87;" Zohar (2000, p. 593) cited r_{WG} values in the.70's and mid .80's as proof that judgments "were sufficiently homogeneous for within-group aggregation;" similarly, Judge and Bono (2000, p. 757) observed that "the mean r_{wg} statistic was .74. This *relatively high level of interrater agreement* appeared sufficient to justify aggregation" (emphasis added).

In addition to implicit references, examples of explicit use of the .70 cutoff are not difficult to identify; for example, George (1990, p. 112) explicated its roots in CTT reliability theory:

> "Thus, to determine if it is meaningful to speak of an affective tone of the group, it is necessary to ascertain if consistency of affect exists within groups. To this end, [r_{WG}] was used. This *interrater reliability coefficient* can be interpreted similarly to *other types of reliability coefficients*. For example, James (personal communication, February 4, 1987) suggests that *a value of .7 or above is necessary to demonstrate consistency* within a group; this is the same figure Nunnally (1978) provided as an acceptable level for an internal consistency reliability coefficient for this type of research. Based on the James et al. procedure, the estimates of within-group interrater reliability … generally show *a high level of agreement*. The average interrater reliability was .87; more specifically, (a) 50 of the 52 estimates … fall *above the .70 ballpark figure for a "good" amount of agreement* (L. R. James, personal communication, February 4, 1987)" (emphasis added).

Recently, Totterdell (2000) relied on the .70 rule, citing George (1990) as justification:

> "To test whether the players' moods were sufficiently consistent within teams to justify the use of aggregate scores …within-group interrater agreement (r_{wg}) estimates were .63 and .79 for individual and team happy mood, respectively. Values on this estimate *must normally be greater than .70 to demonstrate sufficient consistency* (see George, 1990), which implies that only the measure for team happy mood could be aggregated." (emphasis added).

The justification for the $r_{WG} \geq .70$ rule rests largely on the argument that (a) some researchers (e.g., James et al., 1984) viewed rater agreement as being similar to reliability, (b) others claimed that reliabilities as low as .70 are useful (e.g., Nunnally, 1978), therefore (c) $r_{WG} \geq .70$ implies acceptable "interrater reliability" and consensus. We question the use of any arbitrary cutoff. First, it has been argued (e.g., Kozlowski & Hattrup, 1992; Lindell et al., 1999, pp. 131-132) on theoretical grounds that r_{WG} should *not* be viewed as an index of reliability; likewise for interrater *r*s (e.g., Crocker & Algina, 1986; Murphy & DeShon, 2000a; Shrout & Fleiss, 1979). Second, the fact that r_{WG}s using different error terms lie on *different metrics* (e.g., Lindell, 2001, pp. 93-94) implies that no arbitrary cutoff could possibly be valid for all r_{WG}s; for example, even if the .70 cutoff is correct for original r_{WGu}, it would be too lenient using an index (e.g., r_{WGmax}) that yields larger values from identical ratings. Third, there is little empirical basis for a .70 cutoff, and few studies (Kozlowski & Hattrup, 1992, being an exception) have attempted to determine how various r_{WG} values equate with "real world" levels of interrater agreement.

Statistical tests. An alternative strategy for interpreting r_{WG} (e.g., Dunlap et al, 2003; Lindell et al., 1999) involves statistical significance tests. Dunlap et al. (2003) critiqued the chi-square test advanced by Lindell et al., in which "the null hypothesis tested by chi-square is that there is *no agreement among raters* in their rating of an item above and beyond what would be expected by chance or random responding." Dunlap et al. preferred an alternative test that evaluates "the null hypothesis that the actual distribution of responding is *rectangular*" (p. 356). We agree with Dunlap et al. (2003) that statistical tests of r_{WG} are useful if one's objective is to determine if *any* nonzero convergence exists. However, the logic of null hypothesis testing poses limitations on its usefulness, given that such statistics test whether *zero* agreement exists; although useful, this reflects a qualitatively different goal from determining if "reasonable consensus exists for a group to aggregate individual level data to the group level of analysis" (Dunlap et al., p. 357). Evidence of the former offers *necessary,* but certainly not *sufficient,* evidence of the latter.

The Present Study

To interpret r_{WG}, we agree with Kozlowski and Hattrup (1992) that "what is needed is a systematic investigation of the various indexes under varying levels of agreement … to allow researchers to … select methods appropriate for their situation" (p. 166). That is, to supplement statistical tests, researchers need a method for comparing obtained r_{WG}s against r_{WG}s computed in datasets having known "benchmark" levels of agreement or data quality. Toward that end, we used Monte Carlo methods to generate item ratings from simulated raters judging a common target that had *known true* quality (i.e., a given proportion of "true" variance based on the actual score of the target, and a given proportion of random error); we then benchmarked r_{WG} values from different equations, paying particular attention to the level of random error present in data producing r_{WG} =.70. As Kozlowski and Hattrup (1992) and Lindell (2001) noted, the answer to the question "how large an r_{WG} is large enough?" is contingent on the r_{WG} error term, as well as one's context and purpose (including factors such as number of rating points, number of raters, distribution shape). Obviously, a single study – even a massive factorial – cannot capture all possible situations in which benchmarks are needed; we therefore focused on finding a means to allow researchers to benchmark r_{WG} in any specific situation, with any desired agreement goal.

This study had two objectives. First, we sought to illustrate how r_{WG} benchmarking can be employed in a specific research situation, choosing parameters to model a rating process of substantive interest; second, to address the .70 rule-of-thumb, we chose parameters describing more generic, generalizable rating situations. In both cases, ratings quality was varied to range from what we viewed as a *realistic* bad-agreement baseline (i.e., *random* ratings sampled from a realistic population distribution) through perfect (i.e., each rater produced the target's known true score). Although contrived situations in which raters produce even more error variance exist (e.g., Equation 3), we felt this represented an understandable, yet unacceptable baseline (i.e., in what situation would *random* Normal or uniform ratings be deemed acceptable?).

For the application-specific benchmarks, we focused on the *Occupational Information Network* (O*NET; e.g., Peterson, Mumford, Borman, Jeanneret, Fleishman, 1999; Peterson, Mumford, Borman, Jeanneret, Fleishman, Levin, Campion, Mayfield, Morgeson, Pearlman, Gowing, Lancaster, Silver, & Dye, 2001), the replacement for the *Dictionary of Occupational Titles* (DOT). As with earlier research on the Fleishman ability-trait survey (which, like the O*NET, uses single-item *holistic* scales; e.g., Fleishman & Reilly, 1992), the research reported by Peterson et al. (1999, 2001) in support of the O*NET relies heavily on rater-convergence statistics to justify inferences regarding data quality. To evaluate the r_{WG}s produced by the O*NET raters, our Monte Carlo process closely paralleled the O*NET ratings (including its highly skewed distributions). For the latter objective, based on prior Monte Carlo research (e.g., Blunt, 1986; Harvey & Hayes, 1986) we hypothesized that .70 values for r_{WG} (especially using the maximum-variance error term) could be obtained even when ratings contained substantial amounts of error; this question was addressed using more generic rating situation parameters.

Method

Real-Data Sample

O*NET r_{WG} results were obtained from the rater-level judgments used to produce the national O*NET database; teams of 4-6 analysts rated 1,180 "occupational units" (OUs). We used the Level ratings (0-7 scale) for each of the 42 O*NET scales in the *General Work Activity* (GWA) survey (see Figure 1). Using Equations 1-4, r_{WGu}, r_{WGmax}, and a_{WG} were computed for each item in each OU, then aggregated by OU (corrected variances in numerators were used). Because the GWA survey posits that each scale describes *independent* content (Peterson et al., 1999), we did not use more complicated methods of aggregation (e.g., see Lindell, 2001) that assume a common underlying construct. That is, although the question of how the multiple-item generalizations of r_{WG} (e.g., Lindell, 2001) perform is an interesting one that should be addressed in subsequent research, we focused exclusively on benchmarking the single-item r_{WG} indices. Pairwise interrater *r*s were also computed in each OU, and aggregated by OU by computing the median. See Hubbard, McCloy, Campbell, Nottingham, Lewis, Rivkin, & Levine (2000) and Levine, Nottingham, Paige, & Lewis (2000) for further details regarding O*NET data collection.

Monte Carlo Procedure

The Monte Carlo analyses examined the values produced by each r_{WG} index in a range of benchmark datasets having *known true* properties with respect to ratings quality and rater behavior. Rather than manipulating rater covariance directly, we manipulated *data quality* by specifying a true-score target for each item/OU combination, then adding varying amounts of random distortion to each rater/item/OU combination using a CTT-like functional relation:

$$X_{ijs} = qT_{js} + (1-q)e_{ijs} \quad (5)$$

where T = true score for job, X = observed rating, q = quality weight (0-1), and subscripts i = rater, j = job, and s = scale. As in CTT, the OU's true scores were independent of error (for each OU, the same randomly-selected true profile was used for all raters and replications), and error was random and rater-scale-OU specific. To produce data of known true quality (*defined as the relative weighting of true versus error in Equation 5*) we varied q (e.g., q = 0.8 for 80% true, 20% error); to keep T and e on a comparable metric (i.e., to maintain the desired proportions, given that we modeled raw scores), the same random generators were used for both T and e.

Although the theoretical parallels between Equation 5 and CTT are obvious, we termed our independent variable *rating quality* rather than *reliability* to avoid the unnecessary (and in our view, *inapplicable*; e.g., Crocker & Algina, 1986; Murphy & DeShon, 2000a) baggage that the use of the latter term engenders, which in any event is irrelevant to the main task at hand – i.e., benchmarking r_{WG} in data having *known true qualities* (*not* known "interrater reliability"). Obviously, functions other than Equation 5 could have been used; we chose it because it offers an easily comprehensible way to express the amount of undesirable variance in a rating condition (the fact that Equation 5 also describes the way in which CTT postulates that raters make their ratings is not essential to our main objective). Indeed, given that our data-quality manipulation defines the degree to which each rater's item ratings are distorted from the known-true correct value for the target, it is arguably more similar to the notion of rating *accuracy* (defined as the convergence with a known-true standard or criterion) than *reliability*. In any event, the main issue concerned the levels of r_{WG} produced across the range of the quality variable, regardless of whether one chooses to view it as "reliability," "accuracy," or simply freedom from error.

For the O*NET analyses, we used rater groups of N = 5, with 42 integer (0-7) scales; q ranged from 0 to 1 by 0.05, with 200 simulated occupations and 200 replications of each. Given the non-Normal O*NET distributions, custom random-number generators were used for T and e to sample integers from the actual population distributions (i.e., generating uniform 0-1 values, then assigning the rating corresponding to that location in the cumulative distribution function for each item). For the generic analyses, we modeled 20 raters judging 20 7-point (0-6) scales with 50 different true profiles and 200 replications of each, varying the type of random process: (a) random real-valued T and e were sampled from a Normal(3, 1.3) population (to avoid skew), ranged-checked, combined via Equation 5, then rounded to integers; and (b) integer-valued T and e were sampled from a rectangular distribution, and similarly processed. A uniform distribution was examined given the parallels to the "bad baseline" error in r_{WGu}; the Normal distribution was examined given our view that it may represent a more realistic and easily generalizable rating situation (i.e., arguably far more traits that would be the subject of a Likert-type rating process exhibit population distributions that are Normal, as opposed to rectangular, in practice).

Results and Discussion

Table 1 presents descriptive statistics for the O*NET raters, Figure 2 depicts real-data r_{WG} and a_{WG} distributions, and Figure 3 shows the distribution of interrater *r*s. The skewed results in Table 1 reinforce the need to use custom random generators to draw conclusions regarding the O*NET database. As expected from Equations 1-4, Figure 2 shows that the choice of error term strongly affects r_{WG}: median (2.5th to 97.5th percentile) r_{WGu} = .83

(.60-.92), a_{WG} = .89 (.81-.94), and r_{WGmax} = .93 (83.-.97). Figure 2 also indicates that a_{WG} provides somewhat different results than r_{WGu} and r_{WGmax} (which are simply linear transformations of each other). Although the rater profile *r*s in Figure 3 are quite variable, and many are quite low (median = .63), if the .70 rule-of-thumb were applied (and valid) the r_{WG} results could be cited as evidence that O*NET raters exhibited strong convergence (e.g., all a_{WG} and r_{WGmax} > .70, with approximately half > .90).

However, the Monte Carlo benchmarks (see Figure 4) tailored to match the O*NET rating process lead to a very different bottom-line conclusion: namely, that the .70 cutoff is *far* too lenient for these r_{WG} indices. That is, (a) even in conditions of *zero* true data quality (i.e., q = 0: zero true score, totally random ratings sampled from the population distribution for each O*NET item), average r_{WG}s using maximum-variance error terms fall in the low .80's (i.e., a range that is widely viewed in the literature as representing good convergence); (b) for original r_{WGu}, on average the .70 rule is satisfied even with *80% random* item variance; (c) the average r_{WG}s produced by the actual O*NET raters correspond with disturbingly low Monte Carlo benchmarks (e.g., real-rater r_{WGmax} and r_{WGu} correspond with almost *60% random* variance, with median a_{WG} at *70% random*); and (d) the median O*NET interrater *r* corresponds with benchmark random percentages of approximately 50%. Thus, in situations like those seen in the O*NET GWA survey (i.e., 8-point scale, small rater groups, skewed items), the Figure 4 results indicate that the .70 rule-of-thumb for r_{WG} provides a highly exaggerated view of data quality. Although the amount of random error that one might consider to be tolerable in a given situation is open to debate, it is difficult to imagine situations in which ratings containing 50-100% random noise would be deemed acceptable. Indeed, the fact that a r_{WG} = .90 value of corresponds to over *70% random error* for r_{WGmax} and nearly *50% random error* for r_{WGu} clearly indicates the fallacy of attempts (e.g., George, 1990) to equate the metric of r_{WG} with the CTT reliability coefficient.

Although some may consider these results to be paradoxical (especially the q = 0 conditions), they can easily be explained via inspection of Equations 1-4 and Table 1. That is, even ratings sampled randomly from these skewed distributions will show much less variance than a *uniform* 0-8 or bimodal-extremes 0/8 random distribution, a fact that ensures sizable r_{WG}s even with no true agreement. Nonzero cross-rater *r*s can likewise be explained due to over-fitting in small samples (i.e., the *N* for these *r*s is the number of items), plus the fact that random values sampled from a population profile having different *means* across the profile items will tend to parallel the population pattern of high/low item-mean values, further ensuring positive interrater *r*s.

The generic benchmarks (see Figure 5 for uniform, Figure 6 for Normal) further support the conclusion that the .70 rule-of-thumb represents a far-too-lenient standard for interpreting r_{WG}. That is, the Figure 5 results are based on the uniform-random concept of disagreement assumed by r_{WGu}; as a validity check, we see that r_{WGu} does indeed approximate 0 for q = 0 (it is actually slightly negative due to the sample-versus-population variance issue; see Lindell et al., 1999, p. 133). However, even in these arguably somewhat unrealistic conditions (i.e., we think random raters would be more likely to look Normal than uniform), r_{WGu}=.70 is obtained with *over 50% random* variance, and a_{WG} and r_{WGmax} achieve the .70 cutoff with *80% random* variance. Using the Normal distribution that we find more realistic, the .70 rule is met with *85% random* variance for r_{WGu} and *100% random* for a_{WG} and r_{WGmax}. As with the O*NET results, the .70 rule leads to totally misleading and grossly exaggerated data-quality inferences for all r_{WG}s.

Interestingly, for all of the Monte Carlo conditions modeled above (including the very small samples of raters used by the O*NET), the amount of sampling error seen for the various r_{WG} indices across replications (i.e., the dashed lines around the median values) is often relatively modest (especially for the maximum-variance indices, and the more desirable ranges of rating quality), although clearly the choice of a uniform versus non-uniform population distribution exerts appreciable impact (i.e., Figures 4 and 6 versus 5). Given the general similarity of results for maximum-variance r_{WG} indices obtained in the radically different rating situations modeled in Figures 4 and 6, unless one has reason to assume a uniform distribution these results may be generalizable to a wide range of rating situations. Specifically, in both cases the Monte Carlo benchmarks indicate that for maximum-variance r_{WG}, values less than the low-to-mid .90's correspond to very questionable rater performance and high levels of random error. Although we do *not* recommend that researchers adopt a .90 rule-of-thumb to replace the .70 cutoff (i.e., in some situations even values in the low .90's may reflect unacceptable disagreement), these results suggest that r_{WGmax} values in the low .90's and below should be given close scrutiny.

Conclusions

Our findings show that the widely-used .70 rule-of-thumb for r_{WG} produces a grossly inflated view of ratings quality across a wide range of rating situations. We conclude that it is time to abandon the .70 rule, and instead judge r_{WG} using benchmarks that are appropriate to each rating situation; our generic benchmarks may be useful in many situations, and our Monte Carlo method can easily be adapted to produce benchmarks customized for any desired rating situation.

Regarding O*NET, given that the r_{WG}s from O*NET raters match benchmarks that reflect high levels of random rating, we view our findings as raising significant questions regarding the quality of the O*NET database. Because our Monte Carlo parameters were modeled directly on O*NET, there can be little question regarding realism or applicability. Given that much of the data cited in support of the holistic scales used in O*NET (e.g., Fleishman & Mumford, 1991; Peterson et al., 1999, 2001) involves rater convergence, we conclude that significant additional research is necessary before researchers can conclude confidently – as did Peterson et al. (2001) – that "the O*NET provides a *highly usable* and inexpensive methodology for analyzing jobs … [that] will have *a great impact on research and practice*. It is *certain* to provide many years of good service to the public, just as the DOT did" (p. 487, emphasis added).

In short, when median r_{WG}s for the raters who produced the O*NET database correspond to benchmarks having *60-70% random* variance (indeed, *half* of the rated occupations involve *higher* levels of error), the fundamental quality of this database is clearly called into question. Although it remains possible that averaged ratings based on profiles containing high levels of random noise may be found to be useful in practice and lead to valid occupational decisions, this is ultimately an empirical question that must be answered via additional research.

References

Brown, R. D. (2002). *Interrater agreement reconsidered: The role of maximum possible variance.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Orlando.

Campion, M. A., Medsker, G. J., & Higgs, A. C. (1993). Relations between work group characteristics and effectiveness: Implications for designing effective work groups. *Personnel Psychology, 46,* 823-850.

Cohen, A., Doveh, E., & Eick, U. (2001). Statistical properties of the rWG(J) index of agreement. *Psychological Methods, 6*, 297-310.

Crocker & Algina (1986). *Classical and modern test theory*. Orlando: Harcourt Brace Jovanovich.

Dirks, K. (2000). Trust in leadership and team performance: Evidence from NCAA basketball. *Journal of Applied Psychology, 85*(6), 1004–1012

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for rwg and AD interrater agreement indices. *Journal of Applied Psychology*, 88, 356-362.

Fleishman, E. A., & Mumford, M. D. (1991). Evaluating classifications of job behavior: A construct validation of the ability requirement scales. *Personnel Psychology, 44(3)*, 523-575

Fleishman, E. A., & Reilly, M. E. (1992). *Administrator's guide: FJAS: Fleishman job analysis survey.* Palo Alto, CA: Consulting Psychologists Press.

Geyer, P. D., & Hunter, J. E. (1992). *Issues of reliability and validity in ratings of occupational characteristics in the Dictionary of Occupational Titles.* Technical report prepared under contract to the US Department of Labor.

George, J. M. (1990). Personality, affect, and behavior in groups. *Journal of Applied Psychology*, 75(2), 107–116.

Hoffman, D. A., Morgeson, F. P., & Gerras, S. J. (in press). Climate as a moderator of the relationship between leader-member exchange and content specific citizenship: Safety climate as an exemplar. *Journal of Applied Psychology*.

Hubbard, M., McCloy, R., Campbell, J., Nottingham, J., Lewis, P., Rivkin, D., & Levine, J. (2000). *Revision of O*NET Data Collection Instruments.* Raleigh, North Carolina: National Center for O*NET Development.

International Association of Rehabilitation Professionals (2001, April 8). *Extending O*NET to its occupational roots*. Available at http://www.rehabpro.org/onet6.doc.

Levine, J., Nottingham, J., Paige, B., & Lewis, P. (2000, May). *Transitioning O*NET to the Standard Occupational Classification*. Raleigh, North Carolina: National Center for O*NET Development.

James, L., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*(1), 85-98.

Judge, T. A., & Bono, J. E. (2000). Five-Factor Model of personality and transformational leadership. *Journal of Applied Psychology, 85*(5), p 751–765.

Klein, K. J., Conn, A. B., Smith, D. B., Sorra, J. S. (2001). Is everyone in agreement? An exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology, 86*, 3-16.

Kozlowski, S. W. J., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77*, 161-167.

Lindell, M. K., & Brandt, C. J. (1997). Measuring interrater agreement for ratings of a single target. *Applied Psychological Measurement, 21,* 271-278.

Lindell, M. K., & Brandt, C. J. (1999). Assessing interrater agreement on the job relevance of a test: A comparison of the CVI, T, $r_{WG(J)}$, and $r_{*WG(J)}$ indices. *Journal of Applied Psychology, 84,* 640-647.

Lindell, M. K., & Brandt, C. J. (2000). Climate quality and climate consensus as mediators of the relationship between organizational antecedents and outcomes. *Journal of Applied Psychology, 85*(3), 331-348.

Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement, 23*, 127-135.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin, 86, 376–390. Murphy, K. R., & DeShon, R (2000a). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology, 53*(4), 913-924.

Murphy, K. R., & DeShon, R (2000a). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*(4), 873-900.

Murphy, K. R., & DeShon, R (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology, 53*(4), 913-924

Nunally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A.(1999). *An occupational information system for the 21st Century: The development of O*NET.* APA Books.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Levin, K. Y., Campion, M. A., Mayfield, M. S., Morgeson, F. P., Pearlman, K., Gowing, M., Lancaster, A. R., Silver, M., & Dye, D. (2001). Understanding work using the Occupational Information Network (O*NET): Implications for practice and research. *Personnel Psychology, 54,* 451-492.Schmidt, A. M. & DeShon, R. P. (2003). *Problems in the use of rwg for assessing interrater agreement*. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology.

Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology, 74*, 368–370

Schmidt, F. L., Viswesvaran, C., Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*(4), 901-912.

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.

Waldman, D.A., Ramirez, G.G., House, R.J., & Puranam, P (2001). Does Leadership Matter? CEO Leadership Attributes and Profitability Under Conditions of Perceived Environmental Uncertainty. *Academy of Management Journal, 44*(1), 2001: 134-143.

Zohar, D. (2000). A group-level model of safety climate: Testing the effect of group climate on microaccidents in manufacturing jobs. *Journal of Applied Psychology, 85*(4), 587–596.

Table 1. *Summary Statistics for Actual O*NET Ratings*

GWA	Mean	SD	0	1	2	3	4	5	6	7
1	3.326	1.500	0.4075	6.1736	37.6151	59.1245	76.8302	88.8302	98.3698	100.000
2	3.012	1.462	3.1698	12.8302	41.7208	63.8189	83.1245	94.7925	99.3057	100.000
3	2.833	1.450	7.2000	14.9736	42.3547	68.5132	87.6075	96.6792	99.3509	100.000
4	2.235	1.441	16.7094	28.4679	56.7849	80.4528	95.2906	98.8679	99.8491	100.000
5	2.123	1.240	11.0792	26.0528	66.6717	89.5698	95.4566	98.9132	99.8792	100.000
6	2.502	1.442	9.6302	21.6755	54.3547	76.1962	90.3245	97.8415	99.7434	100.000
7	2.476	1.338	8.0906	20.4679	53.1774	79.8340	92.3170	98.6264	99.8038	100.000
8	2.418	1.644	13.1170	28.9057	58.4755	76.4226	87.3962	94.4755	99.3509	100.000
9	2.517	1.676	11.5019	27.2302	57.5245	73.3736	85.6906	94.0981	98.8075	100.000
10	2.515	1.448	6.4453	23.9245	56.2717	76.1962	88.2717	97.6000	99.7283	100.000
11	1.916	1.579	21.6453	43.8189	69.6302	84.5585	92.0604	97.1321	99.4566	100.000
12	2.675	1.621	7.1547	25.1321	51.2755	71.3509	84.1962	94.1585	99.1698	100.000
13	1.283	1.484	44.5585	62.6264	77.9623	91.4717	96.3170	98.8830	99.8340	100.000
14	1.387	1.507	40.8000	58.0377	79.0642	89.0113	95.4264	99.0340	99.9094	100.000
15	2.468	1.474	7.1094	26.3245	58.0075	74.6717	91.0943	96.3623	99.5774	100.000
16	2.725	1.436	5.5094	14.8528	50.0679	74.2943	88.6943	94.9585	99.0340	100.000
17	3.254	1.355	2.4906	7.5774	31.4415	55.0340	83.4566	95.2755	99.2453	100.000
18	2.190	1.505	19.2784	31.2953	56.7784	80.5254	95.2144	98.2337	99.6530	100.000
19	1.209	1.401	45.2679	59.9396	85.2226	93.7660	96.4679	98.6113	99.7585	100.000
20	0.708	1.267	69.6755	76.2264	90.8377	95.8792	97.9623	98.9132	99.6830	100.000
21	1.108	1.494	53.8868	65.9472	83.8340	91.3208	95.7887	98.6113	99.7585	100.000
22	2.258	1.367	13.9623	24.8302	57.7660	84.8604	94.0528	98.7925	99.8642	100.000
23	1.402	1.551	42.8830	56.9208	77.8264	87.9698	95.8038	98.6113	99.7736	100.000
24	1.016	1.338	51.3660	67.9698	89.2679	94.0830	97.4038	98.5057	99.7736	100.000
25	2.222	1.444	14.6113	30.3547	60.0755	80.2264	93.7509	98.7925	99.9396	100.000
26	1.792	1.591	29.1623	46.3849	68.8453	84.8604	93.5849	98.0226	99.8642	100.000
27	2.758	1.635	7.3962	21.3434	50.8528	68.9509	83.6377	92.6642	99.3509	100.000
28	2.268	1.857	25.7208	38.7472	55.2604	72.8151	86.2792	94.9585	99.3358	100.000
29	2.377	1.530	15.3962	29.4340	54.6264	69.4642	94.4755	98.9585	99.9094	100.000
30	1.528	1.551	35.1698	52.7698	79.6075	88.6038	93.2981	97.9472	99.7434	100.000
31	1.251	1.455	44.7245	62.0226	81.9623	90.8226	96.3623	99.1094	99.8340	100.000
32	1.342	1.512	41.3887	59.8642	81.8566	89.6604	94.7774	98.4604	99.7736	100.000
33	1.411	1.706	48.3170	60.7396	73.7962	85.0717	92.6491	98.5962	99.7132	100.000
34	1.673	1.537	31.0189	47.7887	74.3245	87.4415	93.7208	98.5057	99.8792	100.000
35	0.968	1.320	54.3094	70.6113	87.2151	94.0226	97.5849	99.4264	99.9547	100.000
36	1.370	1.505	41.1170	57.4189	80.3774	91.0038	94.8377	98.3849	99.8189	100.000
37	1.037	1.547	60.0151	71.4264	81.6151	88.3170	95.6981	99.2604	99.9547	100.000
38	1.249	1.460	44.5283	61.6755	83.4264	91.0340	95.6226	98.8679	99.8792	100.000
39	1.823	1.830	35.3660	50.8075	67.1698	80.0755	88.6792	96.0302	99.5472	100.000
40	1.647	1.359	25.9170	44.2717	78.8377	91.2000	96.1358	99.0340	99.8189	100.000
41	0.570	1.219	75.9396	84.4830	90.9887	94.0528	97.9170	99.6075	99.9245	100.000
42	1.519	1.582	37.3283	55.0340	76.8906	87.0189	93.3736	98.6113	99.8340	100.000

Note. Mean and SD are from $N = 6{,}625$ ratings from the national O*NET database; columns 0-7 report the cumulative percentages of raters in this population for each of the 0-7 rating points on the given O*NET scale.

Figure Captions

Figure 1. Sample O*NET rating scale from the GWA survey. A zero rating for the Level scale is given if the rater judges the job to be a '1' on the Importance scale, creating an effective 0-7 Level scale.

Figure 2. Scatterplot of distributions of a_{WG} (vertical axis) by r_{WGu} (horizontal) computed from 6,625 O*NET ratings (1,180 occupations rated by teams of 4-6 raters on 42 GWA items); each plot represents the average (i.e., aggregated across items) for each occupation. Reference lines are drawn at the 2.5th, 16th, 50th, 84th, and 97.5th percentiles.

Figure 3. Frequency distribution of median interrater *r* values (i.e., median across all pairs of raters in each rater group for each occupation) for actual O*NET raters.

Figure 4. Values for interrater agreement indices (vertical axis) obtained from analysis of simulated O*NET raters (N = 5 per group) having known percentages of true-score variance (horizontal axis); R = results for median interrater *r* between profiles for each rater group-occupation pairing (orange lines), C = conventional uniform-variance r_{WG} (blue), M = maximum-variance $\underline{r}_{WGmax}$ (red), A = a_{WG} (green). Solid lines represent median values for each level of data quality (0 = zero true variance, 1 = 100% true), dashed lines around median represent 68 percent (i.e., 16th and 84th percentiles) and 95 percent bands (2.5th and 97.5th percentiles).

Figure 5. Values for interrater agreement indices (vertical axis) obtained from analysis of generic 7-point scale using uniform distributions, 20 items, 20 raters, 50 jobs, 200 replications.

Figure 6. Values for interrater agreement indices obtained from analysis of generic 7-point scale using Normal distribution (M=3.0, SD=1.3), 20 items, 20 raters, 50 jobs, 200 replications.

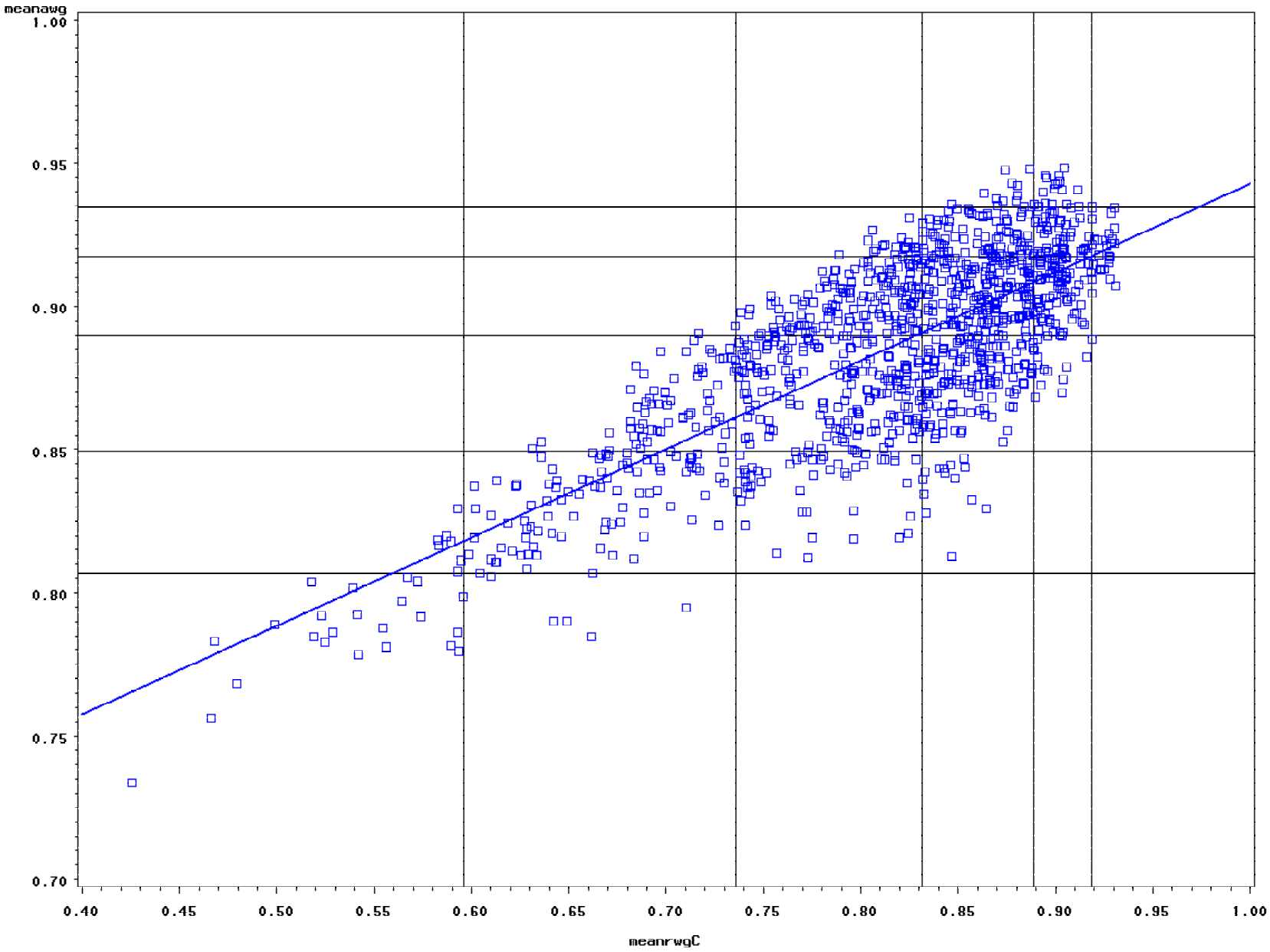
4. Inspecting Equipment, Structures, or Materials

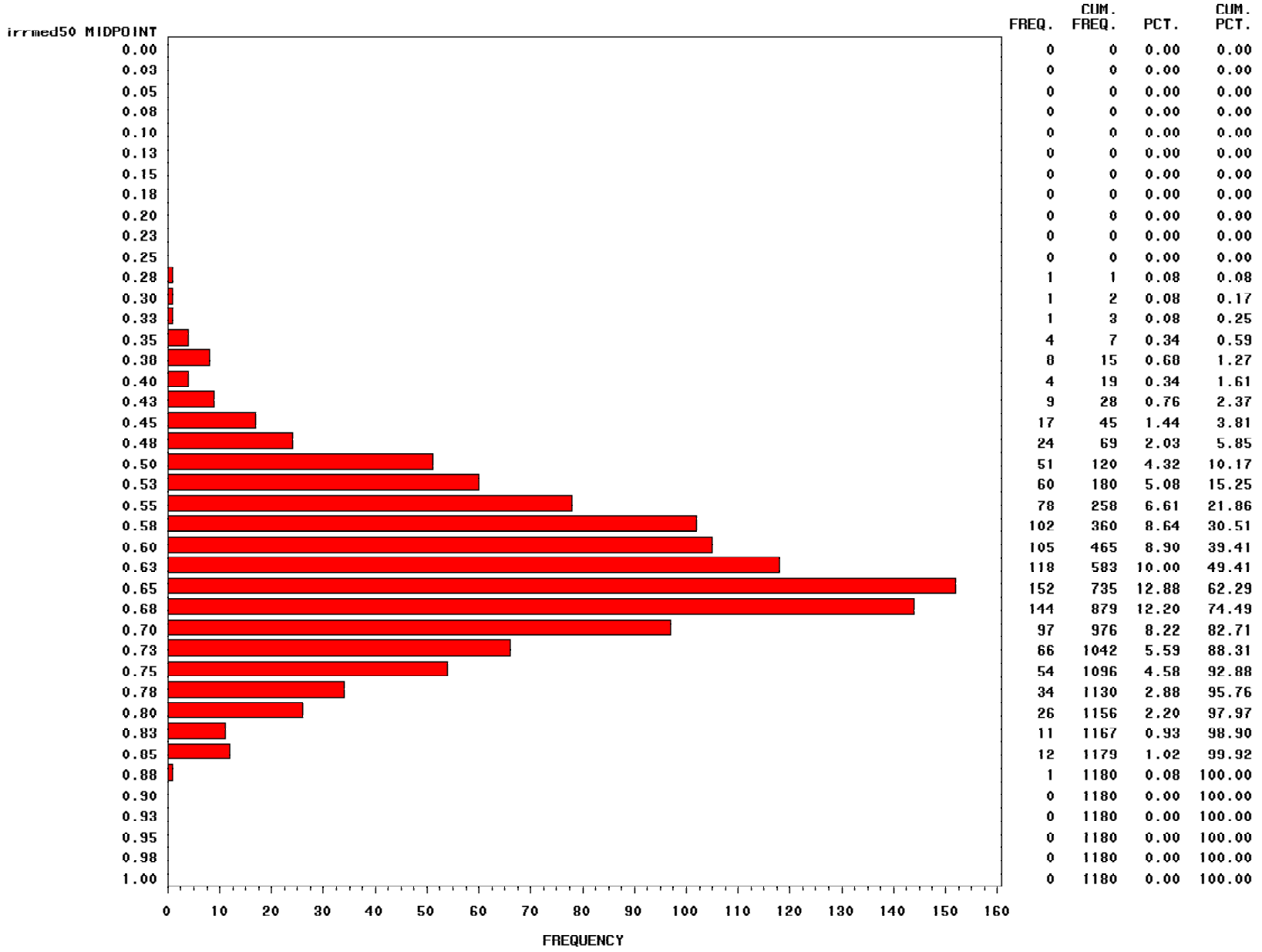
Inspecting equipment, structures, or materials to identify the cause of errors or other problems or defects.

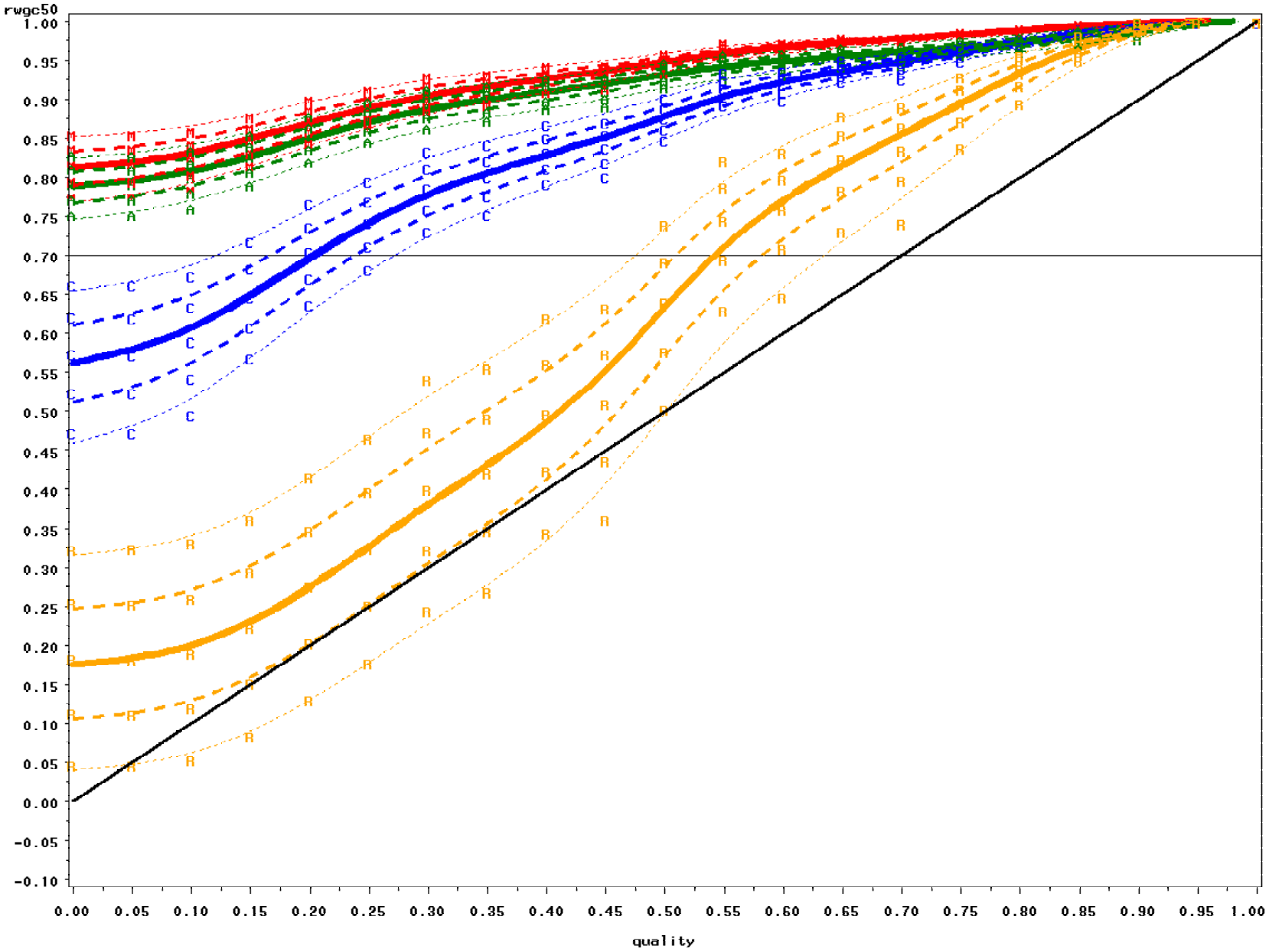
A. How <u>important</u> is INSPECTING EQUIPMENT, STRUCTURES, OR MATERIALS to the performance of *your current job*?

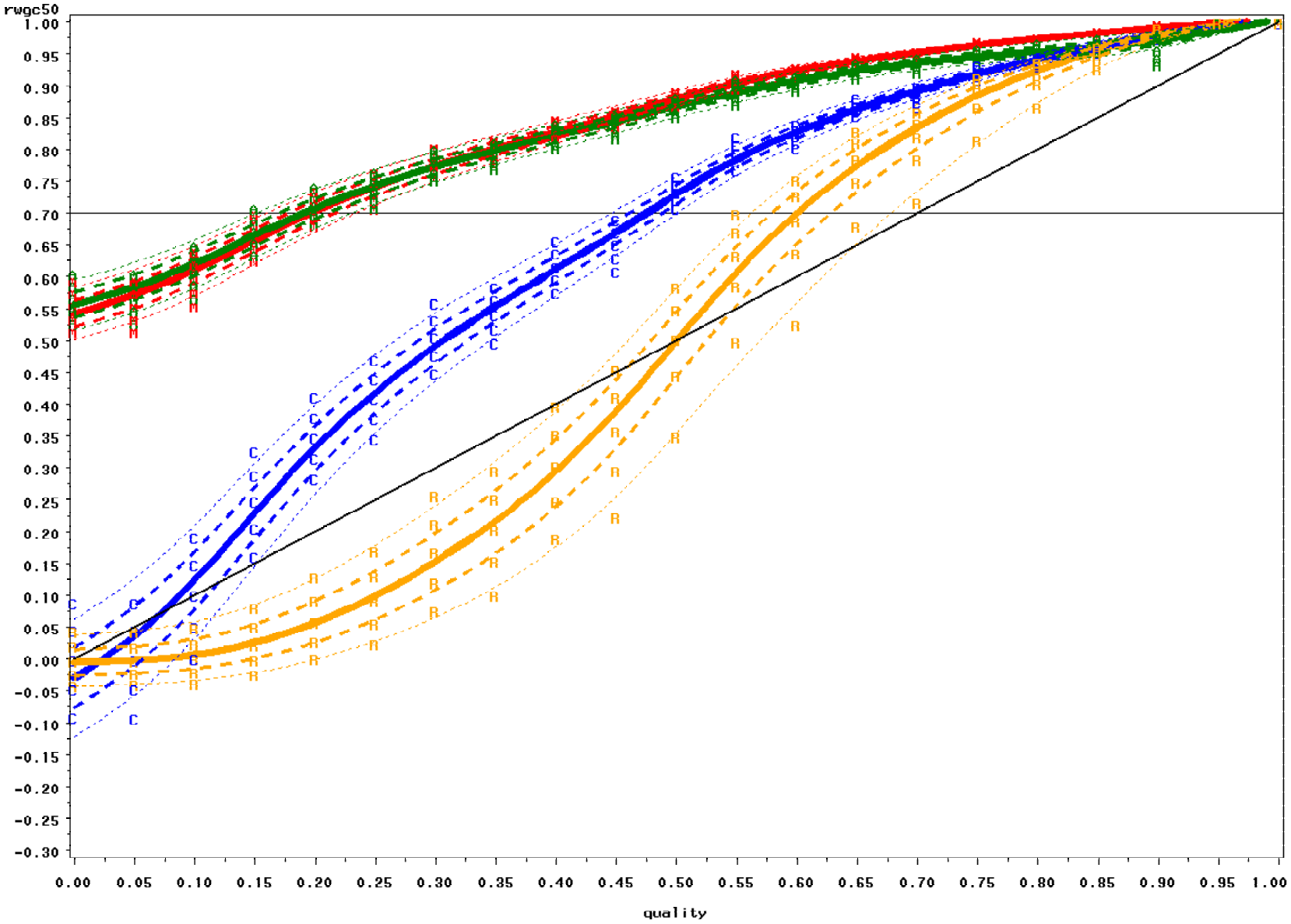
* If you marked Not Important, skip LEVEL below and go on to the next activity.

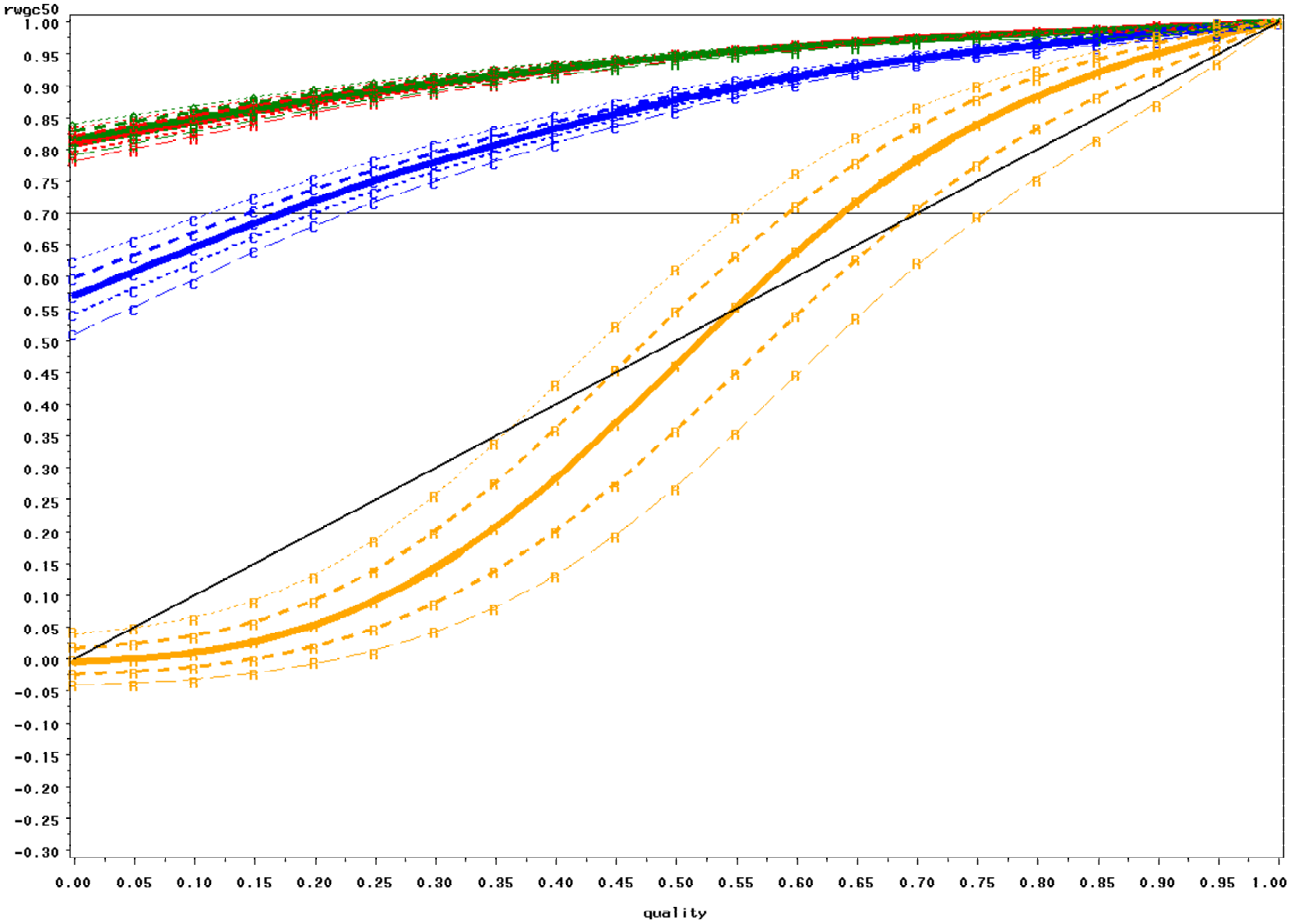
B. What <u>level</u> of INSPECTING EQUIPMENT, STRUCTURES, OR MATERIALS is needed to perform *your current job*?


meanawg
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
meanrwgC


irrmed50 MIDPOINT
FREQ.
CUM. FREQ.
PCT.
CUM. PCT.
0.00 0 0 0.00 0.00
0.03 0 0 0.00 0.00
0.05 0 0 0.00 0.00
0.08 0 0 0.00 0.00
0.10 0 0 0.00 0.00
0.13 0 0 0.00 0.00
0.15 0 0 0.00 0.00
0.18 0 0 0.00 0.00
0.20 0 0 0.00 0.00
0.23 0 0 0.00 0.00
0.25 0 0 0.00 0.00
0.28 1 1 0.08 0.08
0.30 1 2 0.08 0.17
0.33 1 3 0.08 0.25
0.35 4 7 0.34 0.59
0.38 8 15 0.68 1.27
0.40 4 19 0.34 1.61
0.43 9 28 0.76 2.37
0.45 17 45 1.44 3.81
0.48 24 69 2.03 5.85
0.50 51 120 4.32 10.17
0.53 60 180 5.08 15.25
0.55 78 258 6.61 21.86
0.58 102 360 8.64 30.51
0.60 105 465 8.90 39.41
0.63 118 583 10.00 49.41
0.65 152 735 12.88 62.29
0.68 144 879 12.20 74.49
0.70 97 976 8.22 82.71
0.73 66 1042 5.59 88.31
0.75 54 1096 4.58 92.88
0.78 34 1130 2.88 95.76
0.80 26 1156 2.20 97.97
0.83 11 1167 0.93 98.90
0.85 12 1179 1.02 99.92
0.88 1 1180 0.08 100.00
0.90 0 1180 0.00 100.00
0.93 0 1180 0.00 100.00
0.95 0 1180 0.00 100.00
0.98 0 1180 0.00 100.00
1.00 0 1180 0.00 100.00
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
FREQUENCY


rwgc50
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
quality


rwgc50
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
quality


rwgc50
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
quality