CAHRS
Center for Advanced Human Resource Studies

CAHRS / Cornell University
187 Ives Hall
Ithaca, NY 14853-3901 USA
Tel. 607 255-9358
www.ilr.cornell.edu/CAHRS/

WORKING PAPER SERIES

Measurement Error in Research on Human Resource Decisions and Firm Performance: How Much Error is There and How Does its Influence Effect Size Estimates?

Barry Gerhart
Patrick M. Wright
Gary C. McMahan
Scott A. Snell

Working Paper 98-30

ilr School of Industrial and Labor Relations
CORNELL UNIVERSITY
Advancing the World of Work

Measurement Error in Research on Human Resource Decisions and Firm Performance: How Much Error Is There and How Does its Influence Effect Size Estimates?

BARRY GERHART
Owen Graduate School of Management
Vanderbilt University
401 21st Avenue South
Nashville, TN 37203
Telephone: 615-322-3665
Fax: 615-343-7177
gerhartb@ctrvax.vanderbilt.edu

PATRICK M. WRIGHT
Center for Advanced Human Resource Studies
393 Ives Hall
Cornell University
Ithaca, NY 14853-3901
pmw6@cornell.edu

GARY C. MCMAHAN
Department of Management
University of Texas-Arlington
Arlington, TX 76019

SCOTT A. SNELL
Department of Management and Organizations
Pennsylvania State University
417 Beam B.A.B.
University Park, PA 16802
ssnell@psu.edu

http://www.ilr.cornell.edu/cahrs

Abstract

Recent empirical research finds that the relationship between human resource (HR) decisions and firm performance is significant in both statistical and practical terms. However, the typical research design in this area relies upon on a single respondent to validly assess firm-wide HR practices. To date, no study has adequately addressed the reliability of such measures, a basic requirement of construct validity. Previous efforts have either defined reliability so narrowly as to miss a major source of measurement error (raters) or have estimated the unreliability due to raters using incorrect methods. In both cases, the result is upwardly biased estimates of reliability. We estimate reliabilities using intraclass correlation and generalizability coefficients. Our reliability estimates suggest substantial measurement error in the types of HR effectiveness and HR practice measures typically used to predict firm performance. We discuss how this degree of measurement influences research and policy implications.

Measurement Error in Research on Human Resource Decisions and Firm Performance: How Much Error Is There and How Does its Influence Effect Size Estimates?

In recent years, we have seen a good deal of attention focused on better understanding how firms' human resources (HR) decisions influence their financial performance. Conceptual work continues to emphasize the traditional view that HR decisions have an impact through building human capital (skills and abilities) and enhancing motivation toward a firm's objectives. However, it also suggests that HR may differ from other types of resources (e.g., technology) because of the greater difficulty in imitating successful HR systems and because of HR's unique role in organizational learning (Barney, 1991; Becker & Gerhart, 1996; Lado & Wilson, 1994; Snell, Youndt & Wright, 1996; Wright & McMahan, 1992).

Early empirical studies found links between firm performance and individual HR policies, such as compensation (Gerhart & Milkovich, 1990; Gomez-Mejia, 1992) and employee selection (Terpstra & Rozell, 1993), thus supporting the emerging attention to the importance of HR decisions in understanding firm performance. More recent research has established a link between a broader array of HR policies and business performance (Huselid, 1995; Huselid, Jackson & Schuler, 1997; Delery & Doty, 1996). Further, the effect sizes in these studies have been substantial in practical terms. For example, Gerhart's (1997) review found that in three studies just cited, a measure of accounting profits, return on assets, was about 20 percent higher in firms having HR practices one standard deviation above the mean on dimensions such as HR effectiveness and what have become known as high performance work practices: pay for performance, participation in decisions, investment in training, and so forth.

The empirical work on linkages between HR decisions and firm performance is widely regarded as necessary and important because of its potential policy relevance (Becker & Gerhart, 1996; Ichniowski, Kochan, Levine, Olson, & Strauss, 1996). And, the work that has been done has been exemplary in several respects, including its focus on measuring and quantifying effects in dollar terms, the use of relatively large samples of organizations, and, in some cases, the examination of contingency factors.

However, this line of empirical research is susceptible to the same types of problems generic to any new line of inquiry. One such problem is construct validity. Schwab (1980) has convincingly demonstrated that the construct validity of key measures must be explicitly addressed early in a research program. Otherwise, we run the risk of building a set of substantive findings whose validity may later prove to be open to question. The first step in examining construct validity is developing a theory-based definition that guides researchers in

developing appropriate measures. However, as Becker and Gerhart (1996) have shown, there is a great deal of inconsistency across studies regarding the specific HR practices that are included in studies of HR and business performance. Becker and Gerhart also note that even when the same type of HR practice is included in different studies, its measurement may differ. For example, one approach to measuring pay for performance is to ask what percentage of employees are covered by a pay for performance plan. An alternative approach is to ask what is the average ratio of variable payments to base salary. As of yet, there is no agreement on these types of construct definition issues.

The second step in studying construct validity is collecting empirical evidence. A necessary step in building a case for the validity of a measure is demonstrating an acceptable level of reliability. Random measurement error (i.e., unreliability) in measures of HR practices or HR effectiveness leads to a downward bias in the parameter estimates (e.g., regression coefficients) for these variables in an equation having firm performance as the dependent variable. To correct an unstandardized regression coefficient for error in the independent variable, one divides the regression coefficient by the reliability estimate (McNemar, 1969, p. 173; Hunter & Schmidt, 1977, p. 1056). (Note that the magnitude of this correction is greater than the magnitude of the correction for attenuation in a correlation because the latter divides by the square root of the reliability.) For example, with a reliability of .80, one would have to multiply the observed regression coefficient of firm performance on HR by 1.25. Such a correction is useful, but may not change the policy and research implications. However, reliabilities of .50 and .20 would imply correction factors of 2 and 5, respectively, which could well force us to reconsider policy and research implications.

Unfortunately, we do not yet have the reliability estimates necessary for drawing inferences about construct validity and for correcting regression coefficients for measurement error. One reason is that researchers have relied almost exclusively on internal consistency indices of reliability (e.g., Spearman-Brown, Cronbach's coefficient alpha), which incorporate only the portion of measurement error that is due to item sampling. However, when observers are asked to describe organizational properties, "the observer becomes a potentially important source of error variance" (Schwab, p. 16). Yet, the majority of substantive research (e.g., Delery & Doty, 1996; Gomez-Mejia, 1992; Huselid, 1995; Huselid & Becker, 1996; Huselid et al., 1997; Terpstra & Rozell, 1993) on the link between HR and firm performance uses a single rater to describe HR practices or HR effectiveness for an entire organization. Huselid and Becker (1996) argue that an index of interrater reliability is "more appropriate" (p. 415) than the widely used internal consistency estimates in such designs.

A second reason for the lack of necessary reliability evidence is that researchers using multiple raters have often estimated James, Demaree, and Wolf's (1984) T_{wg} index, which assesses interrater agreement, not interrater reliability (Kozlowski & Hattrup, 1992). According to Schmidt and Hunter (1989), a reliability coefficient grounded in classical measurement theory requires estimates of both between and within targets (e.g., organizations) sources of variance. In contrast, T_{wg} focuses entirely on ratings variance within a single target. Thus, Schmidt and Hunter (1989) argued that T_{wg} cannot be interpreted as an index of interrater reliability. In an important sense, James et al. (1984) recognized this distinction from the start in that they cautioned researchers not to use T_{wg} to estimate reliability when more than one target was being rated. Instead, they stated that the intraclass correlation was the appropriate reliability index in such a case. They recommended that T_{wg} be used as an index for assessing within-target agreement of raters and as one criterion in deciding whether such ratings were similar enough to justify their aggregation. However, as Kozlowski and Hattrup (1992) suggest, the fact that James et al. initially referred to T_{wg} as index of reliability was "unfortunate" and "seems to have been the source of some confusion in the literature" (p. 161). James, Demaree and Wolf (1993) subsequently agreed with Kozlowski and Hattrup's recommendation that T_{wg} not be used as an index of reliability.

Although we have not seen T_{wg} used in empirical research on HR and firm performance, per se, it has, until recently been widely used in related areas, such as research on HR and business performance at the facility level, and in empirical research in the groups and strategy literatures. Although the use of T_{wg} or a similar type of index of agreement can be useful and informative, there is now a consensus that T_{wg} is not an appropriate index of reliability. We believe that reliabilities as estimated by appropriate indexes (e.g., the intraclass correlation) will typically be lower than the coefficients produced by the T_{wg} index, which are often .90 or higher. Therefore, one possible consequence of the "confusion" regarding T_{wg} described by Kozlowski and Hattrup (1992) is that these high T_{wg} agreement coefficients may have contributed to a false sense of confidence regarding the interrater reliability of HR measures in studies of firm performance.

The preferred method for estimating the interrater reliability of interval level scales is the intraclass correlation (James, 1982; Shrout & Fleiss, 1979). In a review of intraclass correlations for measures such as organization culture and other work and organization properties, Gerhart (1997) found that raters often exhibited significant unreliability. Based on 6 different studies (Dess & Robinson, 1984; James, 1982; Judge & Cable, 1997; Ostroff & Schmidt, 1993; Sutcliffe, 1994; Viswesvaran et al., 1996) intraclass correlations ranged from a

low of .12 for participation in decision making to a high of .55 for a measure of relative firm sales.[i]

There is little reason to believe that interrater reliability would be any better in the HR and firm performance literature. Respondents are asked to provide a single numerical rating that accurately describes each HR practice on a firm-wide basis, when, in actuality, these practices often vary significantly with respect to location, type of employee, or business unit. Were reliabilities for widely used HR measures to fall within the same range reported by Gerhart (1997), their construct validity would be suspect and effect sizes likely significantly biased. If, as described earlier, the uncorrected effect of a 1 standard deviation increase in HR is a 20 percent increase in firm performance, the corrected effect size using the middle of the reliability range (about .33) would be 60 percent, or 3 times as large as the uncorrected effect size. Clearly, this magnitude of measurement error would have a large impact on our substantive conclusions.

Consequently, the purpose of our study is to estimate interrater reliabilities of firm level ratings of HR practices and HR effectiveness, key variables in the recent literature on HR and firm performance. A second purpose is to demonstrate how two widely used indices, T_{wg} and internal consistency reliability, yield coefficients that overestimate actual reliabilities in such applications. A third purpose is to demonstrate how better-suited methods, such as intraclass correlations and generalizability analysis, can be used to obtain more accurate estimates of reliability.

Method

Sample

Our main sample is composed of 41 HR managers (mostly vice presidents and directors) from 12 firms. We only included firms where there were at least two respondents. The mean number of employees is 46,396 and the median number of employees is 41,800. The firms represent a variety of industries, including banking, energy, processed food, airlines, insurance, computers, food service, chemicals, and pharmaceuticals. Means for demographic variables in the HR managers sample were: years with company (17), years in present position (5), and age (47). In addition, we were able to collect data from 52 line managers (mostly vice presidents or general managers of strategic business units) for one of the measures, HR effectiveness, described below. Means for demographic variables in the line managers sample were: years with company (18), years in present position (3), and age (48).

Measures

Ten items pertaining to employment practices for managerial/professional and hourly employees appear in Table 1. Note that these items are very similar to those used by Huselid

(1995) and Huselid and Becker (1996) in their Employee Skills and Organizational Structures scale.[ii] In addition to similarity of content, the format of the items is also similar in that each item asks about the percentage of employees covered by each practice. The measures that were most consistently related to profitability in the Delery and Doty (1996) study, profit-sharing and results-oriented (versus behavior-based) performance measures, are also represented by our items, especially numbers 2 and 10.

A study by Huselid, Jackson, and Schuler (1997), focused on the capabilities of HR staff and satisfaction with the firm's HR effectiveness. Although we also focus on HR effectiveness in our study, we chose to use a set of items that we had developed in previous research (see Table 2). Without knowing the empirical convergence between our measure and the Huselid et al. measure, we can only draw limited conclusions about the likely reliability of the measure of HR effectiveness they used. However, we feel that the evaluative nature of these items provides an interesting point of comparison with the description-oriented items in Table 1.

Analyses

Interrater reliability was estimated using the ICC(1,k) version of the intraclass correlation defined by Shrout and Fleiss (1979, p. 423) where k is the number of raters. ICC(1,1) estimates the reliability of one rater in the case of a design where each organization is rated by a different set of raters (i.e., raters are nested within organizations). With k greater than one, ICC(1,k) estimates the reliability of an average based on k raters. For comparison purposes, we also report James et al.'s (1984) T_{wg} coefficient for each scale.

To estimate internal consistency reliability, we used both the Spearman-Brown formula and the ICC(2,k) version of the intraclass correlation (Shrout & Fleiss, 1979), which estimates the reliability where the facet of measurement (in this case, items) is crossed with the object of measurement (firms). In other words, all firms are rated on all items. Huselid and Becker (Huselid, 1995; Huselid & Becker, 1996) standardized items prior to forming scales, so Spearman-Brown, which works off of correlations (rather than variances and covariances, which Cronbach's coefficient alpha uses) makes sense. We also used ICC(2,k) to estimate internal consistency to demonstrate the flexibility of the intraclass correlation in estimating reliability.

Finally, we estimated generalizability (G) coefficients (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991), which have the advantage (over reliability coefficients based in classical test theory) of simultaneously recognizing multiple sources of measurement error. According to Shavelson and Webb (1991, p. 93), relative G coefficients, which we use, are "analogous to the reliability coefficient in classical theory, namely, true-score

variance divided by expected observed-score variance (i.e., an intraclass correlation coefficient)."

The first step in estimating G coefficients is to conduct a generalizability (G) study to estimate the variance component associated with each factor. Variance components are estimated using analysis of variance models. It is important to note that, as a general rule, the variance component for a factor is not equivalent to the estimated mean square for that factor because the expected value of the mean square is typically a combination of variance components. Thus, one must correctly define the expected mean square to accurately compute variance components (Shavelson & Webb, 1991). We relied on the maximum likelihood VARCOMP procedure in SAS to estimate variance components (Searle, 1987).

The second step is a decision (D) study, which uses the variance components estimated in the G study to decide what type of measurement design is necessary to achieve an acceptably high G coefficient. For example, in our case, we can determine how the G coefficient changes as we add items, raters, or both. In this sense, a D study is similar to the use of the Spearman-Brown prophecy formula to estimate how many items would need to be added to achieve a certain level of reliability. The difference is that a D study can compare the gains in a single coefficient that come from adding either more items, more raters or both.

In a design with random factors, the true-score variance (numerator) is defined by the variance component (VC) for the object of measurement, in our case, firms. The expected observed-score variance (denominator) is defined as the sum of VCs for terms where a facet of measurement (e.g., items, raters) is crossed with the object of measurement. The ratio of these two components is the G coefficient for one rater and one item. If, in the D study, one wishes to estimate the G coefficient expected when using more than one item or rater, we divide each VC in the denominator by either the number of items, the number of raters, or both as appropriate. In our design, where firm is the object of measurement, and items and raters are facets with raters nested within firms, the G coefficient is defined as (Shavelson & Webb, 1991, p. 90):

$$VC_{firms} / (VC_{rater\ x\ item,\ firm\ x\ rater\ x\ item,\ e} / n_{raters}n_{items} + VC_{firm\ x\ item} / n_{items} + VC_{rater,\ firm\ x\ rater} / n_{raters})$$

Results

Table 1 reports the ICC(1,1)s for each item. With the exception of the formal job analysis question (item 4) about managerial/professional employees, every ICC(1,1) is closer to zero than to one. The mean interrater reliability, as estimated via ICC(1,1) was .162 for managerial/professional employees and .204 for hourly employees. These results indicate that raters within firms are only marginally more consistent with one another than with raters from different firms.

Table 2 reports the ICC(1,1)s for each of the HR effectiveness items. The mean ICC(1,1) across items is .301, which is somewhat higher than the estimates reported in Table 1. At first blush, the fact that subjective evaluations of effectiveness show better interrater reliability than descriptive assessments of actual practice may be surprising. One explanation is that HR professionals may focus more on effectiveness in their day to day work and less on generating firm-wide estimates of the percentage of employees covered by various types of employment practices. There is probably more discussion and sharing of data and opinions around effectiveness than around the coverage-based measures included in Table 1.

Tables 3a and 3b add internal consistency estimates of reliability (using both ICC and Spearman-Brown methods) and combine these with interrater reliability estimates (obtained using ICCs) to obtain G coefficients, which capture both item and rater error sources of error simultaneously. For the analyses reported in Table 3a, we used only three of the items reported in Table 1 because we were unable to obtain G coefficients of greater than zero when we used all ten of the items in Table 1. We chose the three items (7. employment tests used for hiring, 8. hours of training provided to employees each year, 10. individual contingent pay) that seemed to have the strongest combination of internal consistency and interrater reliability. Therefore, we note that the G coefficients reported in Table 3a are higher than they would be otherwise.

However, to compare reliability estimates based on generalizability analysis with estimates based on more traditional methods, we felt it was important to have an example where the G coefficient was greater than zero. We also estimated the reliability that would be obtained by using 11 items (the number used by Huselid (1995) and Huselid and Becker, 1996) having the same level of average intercorrelation as the three we used.

TABLE 1

Intraclass Correlations (ICCs) for Human Resource Practice Items and Scale, HR Respondents

	Managerial/Professional			Hourly		
	ICC(1,1)	N_{Raters}	N_{Firms}	ICC(1,1)	N_{Raters}	N_{Firms}
1. Problem solving groups/quality circles	.032	29	8	.230	25	7
2. Group contingent pay (group bonuses, gainsharing)	.325	33	10	.301	29	9
3. Formal information sharing systems	.000	22	7	.000	17	6
4. Formal job analysis	.519	25	7	.229	20	6
5. Attitude surveys	.013	32	9	.266	29	8
6. Formal grievance procedure/open door	.031	30	9	.233	30	9
7. Employment tests used for hiring	.191	24	7	.178	25	7
8. Hours of training provided employees each year	.090	21	7	.000	17	6
9. Formal performance appraisal	.000	28	8	.426	24	7
10. Individual contingent pay	.421	28	8	.000	24	7
Mean	.162			.204		

Note: All items ask: what percentage of your workforce is covered by or experience each of the following HR practices?

TABLE 2

Intraclass Correlations (ICCs) for Human Resource Effectiveness Items and Scale, HR Respondents

		ICC(1,1)	N_{Raters}	N_{Firms}
1.	The HR department is performing its job the way I would like it to be performed	.496	41	12
2.	The department is very responsive to meeting customer (front line managers and employees) needs	.688	40	12
3.	The department provides me with useful and timely information regarding HR issues	.149	40	12
4.	This department has helped to enhance the firm's competitive position	.482	40	12
5.	This department provides value-added contributions to the firm's bottom line	.447	40	12
6.	This department contributes to building and/or maintaining the firm's core competence	.276	40	12
7.	This department contributes to building the firm's human capital (employees/managers) as a source of competitive advantage	.204	40	12
8.	The policies, practices, and procedures coming from the HR department help front-line business partners in their jobs	.000	40	12
9.	The HR department has developed a well coordinated set of policies, practices, and procedures	.052	40	12
10.	The HR policies, practices, and procedures help support the firm's business plan	.211	40	12
Mean		.301	40	12
Scale		.466	40	12

Response format: 1 = Not at all to 7 = To a great extent

TABLE 3A
Reliability Estimates for a Three-item HR Practices Scale

Reliability Estimate Given Different Source(s) of Measurement Error:	Managerial/ Professional	Hourly
Sampling of Items		
ICC(2, k = 3 items)	.668	.810
Spearman-Brown (3 items)	.639	.753
Sampling of Raters		
ICC(1, k = 1 rater)	.353	.407
Sampling of Items and Raters		
Generalizability Coefficient (1 rater, 3 items)	.271	.276
Generalizability Coefficient (2 raters, 3 items)	.422	.416
Generalizability Coefficient (1 rater, 11 items)	.363	.325
Generalizability Coefficient (3 raters, 11 items)	.627	.579

Notes:
- Each of the three HR practices items was standardized prior to being summed to form the scale. The three items are: employment tests used for hiring, individual contingent pay, and annual hours of training for employees.
- The sample sizes used in calculating the Spearman-Brown reliabilities were 22 raters in the managerial and professional sample and 20 raters in the hourly sample.
- For the generalizability and ICC analyses, we used only 2 raters from each firm to achieve a balanced design, which made estimation of the necessary mean squares and variance components more straightforward. For the managerial/professional sample, there were 7 firms and 2 raters at each firm. For the hourly sample, there were 6 firms and 2 raters at each firm.

TABLE 3B
Reliability Estimates for a Ten-item Human Resource Effectiveness Scale

Reliability Estimate Given Different Source(s) of Measurement Error:	
Sampling of Items	
ICC(2, k = 10 items)	.891
Spearman-Brown (10 items)	.883
Sampling of Raters	
ICC(1, k = 1 rater)	.524
Sampling of Items and Raters	
Generalizability Coefficient (1 rater, 10 items)	.475
Generalizability Coefficient (1 rater, 20 items)	.503
Generalizability Coefficient (2 raters, 10 items	.637
Generalizability Coefficient (3 raters, 10 items)	.718

Notes:
* The sample size for calculating the Spearman-Brown reliability was 42 raters.
* For the generalizability analyses and ICC analyses, we used 2 raters from each firm to achieve a balanced design, which made estimation of the necessary mean squares and variance components more straightforward. There were 12 firms.

The most important finding from Table 3a is that internal consistency estimates of reliability seriously overestimate overall reliability because they incorporate only item sampling (not rater sampling) as a source of measurement error. For example, in describing HR practices for hourly employees, the Spearman-Brown estimate of internal consistency is .753. However, the interrater reliability for the three item scale as estimated by ICC(1,1) is only .407. The G coefficient, which recognizes sampling of both raters and items as error sources is still lower at .276. In other words, the internal consistency estimate is .753/.276 = 2.7 times greater than the G coefficient.

One of the valuable features of generalizability analysis is that one can examine the most efficient way to achieve higher levels of reliability. Because rater differences are the major source of measurement error in this case, we obtain the largest increases in the G coefficient by adding raters rather than by adding more items (See Table 3a). But, even with 3 raters and 11 items, the G coefficients remain less than Nunnally's (1978) recommended minimum level of .70.

The analyses using the HR effectiveness items did not require us to select a subset of items. Using all 10 items, we obtained an internal consistency (Spearman-Brown) reliability estimate of .883 (Table 3b). Interrater reliability, as estimated by ICC(1,1) was .524. The estimated G coefficient using 1 rater and 10 items was .475. Again, the biggest increases in reliability come from adding raters rather than items. With 3 raters and 10 items, the G coefficient increases to .716, exceeding the suggested .70 minimum.

Thus far, we have limited ourselves to examining reliability among HR managers. However, it is quite possible for there to be strong reliability among HR managers, but low convergence with another key group, line managers. Data on the HR effectiveness scale were available from both HR managers and line managers. The correlation between firm level mean responses of line managers and HR managers was .523 ($t = 1.73$, $n = 10$). However, to obtain the expected correlation between a single randomly selected line respondent and a single randomly selected HR respondent, we need to attenuate the correlation between mean responses.[iii] Using a formula from Nunnally and Bernstein (1994, p. 257), we estimated the attenuated correlation to be .257. This is the expected correlation between an HR manager and line manager responding to the same 10 items and describing the same firm. Thus, although the estimated reliability (based on the G coefficient) of the HR effectiveness scale using HR respondents was .475, the estimate is considerably lower, .257, if one defines reliability as the correlation between individual line manager and HR manager responses.

Table 4 reports ICC(1,1)s for common HR benchmarks. In some cases, the ICC(1,1)s are significantly higher than those obtained above for the HR practices and HR effectiveness items and scales. For example, HR managers show substantial reliability in describing the employees/HR staff and payroll/revenues ratios. These ratios are widely used benchmarks of effectiveness in many firms and are probably estimated and communicated on a regular basis. On the other hand, less time and effort may go into the measurement and reporting of the practices shown in Table 1.

Given the above reasoning, we were surprised at the low interrater reliability in measuring the HR Budget/revenues ratio. Upon closer inspection of the data, we discovered that there was actually a great deal of agreement within firms. In fact, the ratio was, in almost all cases, estimated by respondents to be .00. But, even though there was high agreement within firms, there was little or no variance across firms, resulting in a low reliability.[iv]

TABLE 4
Interrater Reliability (Intraclass Correlations) for HR Benchmarks

Respondents		ICC(1,1)	N_{Raters}	N_{Firms}
Employees/HR Staff	HR Managers	.832	25	9
Payroll/Revenues	HR Managers	.710	20	8
HR Budget/Revenues	HR Managers	.009	19	7

TABLE 5
Intraclass Correlation, ICC(1,1) Versus T_{wg}

	T_{wg}	ICC(1,1)	Items
Employment Tests Used for Hiring, Managerial/Professional Employees	.765	.191[a]	1
Hours of Training Provided Employees Each Year, Managerial/Professional Employees	.722	.090[a]	1
Individual Contingent Pay, Managerial/Professional Employees	.880	.421[a]	1
HR Practices Scale (using above three items), Managerial/Professional Employees	.906	.353[b]	3
HR Effectiveness Scale, Respondents: HR managers	.976	.524[c]	10
HR Effectiveness Scale, Respondents: Line Managers	.921	.143	10
Mean	**.862**	**.287**	

[a]From Table 1 [b]From Table 3a [c]From Table 3b

Finally, Table 5 demonstrates that there are remarkably different measurement error implications depending on whether one uses ICC(1,1) or James et al. (1984) T_{wg} index. Across the six measures shown in Table 5, the mean for the T_{wg} index is .862, whereas the mean for ICC(1,1) is .287. Because of the confusion regarding its interpretation, T_{wg} has been often been interpreted as an index of interrater reliability in organizational research. However, the recent consensus is that T_{wg} cannot be interpreted as an index of reliability (James, Demaree, & Wolf, 1993; Kozlowski and Hattrup, 1992; Schmidt & Hunter, 1989) and our results demonstrate that the T_{wg} index greatly overestimates interrater reliability.

Discussion

The field of strategic HR management has increasingly focused on the impact of HR decisions on measures of firm performance. Although significant progress has been made in understanding the links between HR practices and firm performance (Becker & Gerhart, 1996), there has been insufficient attention paid to construct validity issues such as reliability in this literature. The typical research design, which uses a single rater to assess HR and other organization-level constructs, introduces a source of measurement error, raters, that has been largely ignored to this point. Yet, our findings show that this oversight matters and needs to be addressed if we wish to have confidence in substantive findings.

Past strategic HR research has typically taken a narrow approach to reliability, focusing almost exclusively on the amount of error variance due to item sampling, which is estimated via internal consistency reliability coefficients such as Cronbach's alpha and Spearman-Brown. However, our findings illustrate how even internally consistent scales can be extremely unreliable measures of organization-level constructs. The generalizability coefficient for a three-item HR practices scale (for hourly employees) constructed to <u>maximize</u> reliability was only .276, even though the internal consistency for that same scale was .810. Similarly, the generalizability coefficient for a ten-item scale of HR effectiveness was .475, whereas the internal consistency reliability estimate was .891. Thus, our findings suggest that focusing exclusively on error variance due to items and ignoring error variance due to raters is likely to cause one to seriously overestimate reliability.

In other cases, researchers have recognized the need to account for error variance due to raters, but have used the T_{wg} index to do so. Although T_{wg} may be useful for some purposes, the recently emerging consensus is that T_{wg} should not be interpreted as an index of interrater reliability (James, Demaree, & Wolf, 1993; Kozlowski and Hattrup, 1992; Schmidt & Hunter, 1989). Consistent with this view, we found that the magnitude of the T_{wg} index was roughly three times as large as the magnitude of the intraclass correlation, the recommended index of

interrater reliability. Thus, our empirical findings reinforce these recent conceptual arguments that T_{wg} should not be used to assess interrater reliability.

Now, consider the implication of our measurement reliability findings for the substantive empirical literature, which finds that a 1 standard deviation increase in the sort of HR variables shown in Tables 1 and 2 is associated with roughly a 20 percent increase in firm performance. Our reliability estimates for the HR practices scale were .276 for hourly employees and .271 for managerial/professional employees. Using the higher of the two implies that a correction factor of 1/.276 = 3.6 needs to be applied to the uncorrected relationship where single-rater designs are used. With the correction, we would expect that a 1 standard deviation increase in HR practices of the sort described in Table 1 would be associated with a 72 percent improvement in firm performance. Moreover, because .276 is likely a high estimate of reliability (because we screened items to maximize internal consistency and interrater reliability), it is quite possible that actual reliabilities would be lower and corrections thus greater using different data or assumptions.

Are effect sizes of this magnitude plausible? To some, such corrected effect sizes may be so implausibly large that they will dismiss them. To others, the corrected effect sizes will only reinforce the view that HR decisions are a crucial source of firm performance and competitive advantage. Our data cannot tell us which view is correct, but our data do allow us to demonstrate that measurement issues need to be more thoroughly addressed before we can hope to provide the answer.

Limitations of the Study

One response to our findings is to argue that some raters may be more accurate than others and that our reliability estimates have been pulled down by giving equal weight to all respondents from an organization. For example, it is possible that the top HR person is more accurate than other respondents in our study. Unfortunately, we were unable to identify the positions of our respondents because of the anonymous nature of our survey. However, our findings demonstrate that there is fairly strong interrater reliability for measures of HR effectiveness, the employees/HR staff ratio, and the payroll/revenues ratio. Thus, to make the differential accuracy argument, one would have to make the case that HR respondents who are not in the top role are highly accurate sometimes (assuming for the sake of argument that agreement with the top HR person denotes accuracy) and quite out of touch and inaccurate at other times (i.e., in describing HR practices).

A second limitation has to do with the modest sample size, both in terms of number of organizations and number of raters. As the sample size decreases, variability in parameter

estimates (e.g., reliability coefficients) increases across samples. However, it seems unlikely that the sample size could account for the differential pattern of results within our sample, which is that internal consistency reliabilities are consistently high, whereas interrater reliabilities are consistently lower. That recognizing multiple sources of error would lead to lower reliability estimates also seems unlikely to be influenced by sample size. In addition, we believe our general finding, that single rater designs introduce substantial error in measuring organization-level HR properties, is likely to be replicated in other samples because, as discussed earlier, we know that raters have been found to be an important source of measurement error in related literatures (Gerhart, 1997). This fact has received less attention than it should, perhaps because of past confusion in the literature regarding both how interrater reliability should be assessed (Kozlowski & Hattrup, 1992) and how to recognize multiple sources (e.g., items, raters) of measurement error in one estimate. We hope that our study helps clarify these issues.

Implications for Future Research

The fact that existing HR measures may be subject to significant measurement error problems has important implications for future research. These implications concern ways of reducing the magnitude and impact of measurement error, the role of systematic measurement error, and what our findings mean for other areas of organizational research.

First, and most clearly, our findings indicate the need to use more than one respondent to describe organizational properties and to report interrater reliabilities. Researchers may also be able to achieve improved reliability by developing a greater consensus regarding both how HR practices should be measured (e.g., percentage of employees covered by a practice versus a rating of the importance of the practice) and which HR practices should be consistently measured across studies (Becker & Gerhart, 1996; Wright & Sherman, 1997). Another recommendation is to consider correcting parameter estimates (e.g., regression coefficients) for the attenuating effects of measurement error using LISREL or other such approaches.

Second, although our focus here was on random measurement error, construct validity also requires evidence that systematic error is within acceptable levels. Unlike random error, systematic measurement error can cause regression estimates to be biased upward. One such scenario would occur when HR assessments are endogenous to firms' financial performance (Gerhart, 1997). Indirect evidence of this possibility comes from a recent study of Fortune reputation ratings by Brown and Perry (1994), which concluded that there is a "financial performance halo" that causes a "perceptual distortion" (p. 1349) when respondents assess HR-related aspects of corporate reputation. Thus, even with acceptable reliability, assessments of

HR-related variables may be susceptible to such systematic errors. Future research needs to examine this possibility.

Finally, we suggest that our findings are relevant to the broader strategy literature where there has also been a heavy reliance on single respondents to measure firm properties and virtually no evidence regarding the interrater reliability of such measures (Starbuck & Mezias, 1996). We hope that our findings will encourage strategy researchers to re-examine the construct validity of their measures and determine whether their substantive findings stand up under such scrutiny.

Conclusion

Despite the importance of construct validity in drawing accurate inferences about the magnitude and causality of relationships between variables of substantive interest, we too often neglect even basic construct validity issues such as the accurate estimation of measurement reliabilities. As a consequence, "our knowledge of substantive relationships is not as great as is often believed, and (more speculatively), not as great as would be true if the idea of construct validity received more attention" (Schwab, 1980, p. 4). Our findings demonstrate that this concern is very relevant to the literature on the link between HR and firm performance, where an important component of measurement error, rater differences, has been largely ignored. Before strategic HR researchers can credibly argue for the strong positive impact of HR on firm performance, we must do a better job of gauging the impact of measurement error, both random and systematic, on our findings. We suspect that other areas of organization research (e.g., business strategy) need to recognize and deal with a similar challenge.

References

Barney, J. 1991. Firm resources and sustained competitive advantage. Journal of Management, 17, 99-120.

Becker, B. & Gerhart, B. 1996. The impact of human resource management on organizational performance: Progress and prospects. Academy of Management Journal, 39, 779-801.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. 1972. The dependability of behavioral measurements: Theory of generalizability of scores and profiles. New York: John Wiley.

Delery, J.E. & Doty, H.D. 1996. Modes of theorizing in strategic human resource management: Tests of universalistic, contingency, and configurational performance predictions. Academy of Management Journal, 39, 802-835.

Dess, G.G. & Robinson, R.B. Jr. 1984. Measuring organizational performance in the absence of objective measures: The case of the privately-held firm and conglomerate business unit. Strategic Management Journal, 5, 265-273.

Gerhart, B. 1997. Human Resource Management and Firm Performance: Measurement Issues and Their Effect on Causal and Policy Inferences. Paper presented at the Cornell Center for Advanced Human Resource Studies Conference on Human Resource Strategy, October 1997, Ithaca, NY.

Gerhart, B. & Milkovich, G.T. 1990. Organizational differences in managerial compensation and financial performance. Academy of Management Journal, 33, 663-691.

Gomez-Mejia, L.R. 1992. Structure and process of diversification, compensation strategy, and firm performance. Strategic Management Journal, 13, 381-397.

Hunter, J.E. & Schmidt, F.L. 1977. A critical analysis of the statistical and ethical implications of various definitions of test fairness. Psychological Bulletin, 83, 1053-1071.

Huselid, M.A. 1995. The impact of human resource management practices on turnover, productivity, and corporate financial performance. Academy of Management Journal, 38, 635-672.

Huselid, M.A. & Becker, B.E. 1996. Methodological issues in cross-sectional and panel estimates of the human resource-firm performance link. Industrial Relations, 35, 400-422.

Huselid, M.A., Jackson, S.E., & Schuler, R.S. 1997. Technical and strategic human resource management effectiveness as determinants of firm performance. Academy of Management Journal, 40, 171-188.

Ichniowski, C., Kochan, T.A., Levine, D., Olson, C., & Strauss, G. 1996. What works at work: Overview and assessment. Industrial Relations, 35, 299-333.

James, L.R. 1982. Aggregation bias in estimates of perceptual agreement. Journal of Applied Psychology, 67, 219-229.

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

Judge, T.A. & Cable, D.M. 1997. Applicant personality, organizational culture, and organization attraction. Personnel Psychology, 50, 359-394.

Hunter, J.E. & Schmidt, F.L. 1977. A critical analysis of the statistical and ethical implications of various definitions of test fairness. Psychological Bulletin, 83, 1053-1071.

James, L.R., Demaree, R.G., & Wolf, G. 1993. T_{wg}: An assessment of within-group interrater agreement. Journal of Applied Psychology, 78, 306-309.

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.

Lado, A.A., & Wilson, M.C. 1994. Human resource systems and sustained competitive advantage; A competency-based perspective. Academy of Management Review, 19, 699-727.

McNemar, Q. 1969. Psychological statistics. New York: John Wiley & Sons, 4th edition.

Nunnally, J.C. 1978. Psychometric theory. New York: McGraw-Hill, 2nd edition.

Nunnally, J.C. & Bernstein, I.H. 1994. Psychometric theory. New York: McGraw-Hill, 3rd edition.

Ostroff, C. & Schmitt, N. 1993. Configurations of organizational effectiveness and efficiency. Academy of Management Journal, 36, 1345-1361.

Schwab, D.P. 1980. Construct validity in organizational behavior. Research in Organizational Behavior, 2, 3-43.

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.

Searle, S.R. 1987. Linear models for unbalanced data. New York: Wiley.

Shavelson, R.J. & Webb, N.M. 1991. Generalizability theory: A primer. Newbury Park, CA: Sage.

Shrout, P.E. & Fleiss, J.L. 1979. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Snell, S.A., Youndt, M.A., & Wright, P.M. 1996. Establishing a framework for research in strategic human resource management: Merging resource theory and organizational learning. Research in Personnel and Human Resources Management, 14, 61-90.

Starbuck, W.H. & Mezias, J.M. 1996. Opening Pandora's box: Studying the accuracy of managers' perceptions. Journal of Organizational Behavior, 17, 99-117.

Sutcliffe, K.M. 1994. What executives notice: Accurate perceptions in top management teams. Academy of Management Journal, 37, 1360-.

Terpstra, D.E. & Rozell, E.J. 1993. The relationship of staffing practices to organizational level measures of performance. Personnel Psychology, 46, 27-48.

Viswesvaran, C., Ones, D.S., & Schmidt, F.L. 1996. Comparative analysis of the reliability of job performance ratings. Journal of Applied Psychology, 81, 557-574.

Wright, P.M., & McMahan, G.C. 1992. Theoretical perspectives for strategic human resource management. Journal of Management, 18, 295-320.

Wright, P.M. & Sherman, W.S. 1997. Failing to find fit in strategic human resource management: Theoretical and empirical problems. Paper presented at the Cornell Center for Advanced Human Resource Studies Conference on Human Resource Strategy, October 1997, Ithaca, NY.

Notes

i. In a number of cases, Gerhart found it necessary to compute intraclass correlations for a single rater using reported intraclass correlations that pertained to the mean of multiple raters. Therefore, the range of intraclass correlations differs from the range reported in the studies reviewed, which sometimes focused on mean rather than individual ratings.

ii. Differences from the Huselid (1995) and Huselid and Becker (1996) studies are as follows. Our item 2 does not ask explicitly about profit sharing, per se. We also do not have an item that pertains to the type of promotion rules used. Finally, item 10 asks the percentage of employees covered by individual contingent pay whereas the corresponding item from Huselid and Becker's work asks the percentage of employees whose performance appraisals are used to determine their compensation.

iii. Table 3b indicates that the generalizability coefficient with 2 raters and 10 items was .637. With 1 rater and 10 items it was .475. We were unable to estimate the generalizability coefficient using line item responses because of problems obtaining variance components. Therefore, we used the intraclass correlation (based on 52 respondents, 11 firms), which captures the main source of error variance (rater differences). ICC(1,k) was .441, whereas ICC(1,1) was .143.

iv. This finding led us to revisit the reliabilities reported in Table 1. We found that, with one exception, the low reliabilities reflected substantial disagreements within firms. The exception is for the item asking about coverage of managerial and professional employees by formal performance appraisal systems. All but 3 of 28 raters responded that 100 percent of such employees were covered.