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 Resources and Firm Performance: Additional Data and Suggestions for Future Research

Patrick M. Wright
Timothy M. Gardner
Lisa M. Moynihan
Hyeon Jeong Park
Barry Gerhart
John Delery

Working Paper 00 – 21

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

Measurement Error in Research on Human Resources and Firm Performance:

Additional Data and Suggestions for Future Research

Patrick M. Wright
Department of Human Resource Studies
School of Industrial and Labor Relations
Cornell University
Ithaca, NY 14853-3901

Timothy M. Gardner
Department of Human Resource Studies
School of Industrial and Labor Relations
Cornell University
Ithaca, NY 14853-3901

Lisa M. Moynihan
Department of Human Resource Studies
School of Industrial and Labor Relations
Cornell University
Ithaca, NY 14853-3901

Hyeon Jeong Park
Department of Human Resource Studies
School of Industrial and Labor Relations
Cornell University
Ithaca, NY 14853-3901

Barry Gerhart
School of Business
University of Wisconsin
Madison, WI 53706-1323

John Delery
College of Business Administration
University of Arkansas
Fayetteville, AR 72701

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

Abstract

Gerhart and colleagues and Huselid and Becker recently debated the presence and implications of measurement error in measures of human resource practices. This paper presents data from three more studies, one of large organizations from different industries at the corporate level, one from commercial banks, and the other of autonomous business units at the level of the job. Results of all three studies provide additional evidence that single respondent measures of HR practices contain large amounts of measurement error. Implications for future research are discussed.

As organizations seek to develop sources of competitive advantage, researchers and practitioners have looked to firms' human resources. Recent research by Huselid (1995), MacDuffie (1995), Delery and Doty (1996), and others has demonstrated significant relationships between human resource (HR) practices and business performance. This line of research has estimated that a one standard deviation increase in the use of "progressive" or "high performance" work practices can result in up to a 20% increase in firm performance (Becker & Gerhart, 1996; Gerhart, 1999).

Gerhart, Wright, McMahan and Snell (in press) examined the extent to which measurement error might exist in the measures of HR practices that have been used in past research. They presented data from 14 large organizations demonstrating very low interrater reliabilities, with ICC (1,1) estimates for individual practices averaging .16 for managerial jobs and .20 for hourly jobs. In addition, they estimated the interrater reliability for a 16 item HR practice scale at only .21. Finally, a 6 item HR practice scale constructed artificially to maximize reliability resulted in an interrater reliability of .61, but still the generalizability coefficient (taking into account error due to items as well as error due to raters) reached only .42.

Huselid and Becker (in press) suggested that these results did not reflect what might exist with past research on the HR – firm performance relationship for a number of reasons. First and foremost, they argued that the size of the organizations in the Gerhart Wright, et al study (over 40,000 employees on average) is quite different from most of this research (e.g., the Huselid study sample averaged just over 4,000 employees per firm). Second, they argued that in these large organizations, the diversification (both the large number of independent business units and the global geographic spread) would have made it difficult for the Senior VP of HR to accurately describe the practices that exist across the whole corporation. Third, they took issue with the wording of the items, noting differences between those used by Gerhart Wright, et al and their more recent research. Finally, they argued that simply adding raters who were not as knowledgeable about the HR practices (e.g., a VP of Compensation or a Director of a HR within a business unit) as the Senior VP of HR would decrease, rather than increase reliability.

Gerhart, Wright and McMahan (in press) acknowledged that all of these issues might have resulted in marginally lower estimates, but suggested that even in the best case scenarios, one would probably expect to observe significant error due to raters. Indeed, they presented further evidence of low interrater reliability based on a sample of refineries, which, given their relatively small size and the standardization of HR practices within refineries, might have been expected to yield high interrater reliabilities. However, Gerhart et al. (b) called for research that would help document the factors (e.g., size, unit of analysis) influencing interrater reliability. Our

purpose in this paper is to respond to this this call for research that helps us understand how design factors influence the level of measurement error in measures of HR practices. We present three studies, each based on different samples and designs. The first study essentially replicates the Gerhart, Wright, et al. (in press) study examining interrater reliability among respondents at a small sample of large corporations responding to practices across a number of managerial/professional/technical jobs. It is expected that like the previous study, this should result in a lower bound estimate of the reliability of such measures. The second study examines interrater reliability between HR directors and a job incumbent in three job categories at a sample of commercial banks when practices are assessed for specific jobs. Finally, the third study attempts to assess the upper bound estimate of interrater reliability through examining measures of HR practices for six specific job categories in 33 single location business units with an average of 500 employees per location. These autonomous business units are all part of the same corporation, in the same industry, serving similar customer markets using similar technologies.

Measurement Error and Its Potential Impact on Effect Sizes

As noted by Gerhart, Wright, McMahan, and Snell (in press), one of the first steps in construct validation requires assessing the measurement error that exists in the proposed measure of a construct (Schwab, 1980). Measurement error can come from a number of sources, the most common of which are items, time, and raters. One assesses the amount of measurement error due to item sampling through internal consistency estimates such as Cronbach's alpha. Estimates of the amount of measurement error due to time are usually assessed through test-retest correlations. Finally, one can assess the amount of measurement error due to raters through computing interrater reliability indices.

Research on HR and firm performance generally has emphasized assessing measurement error due to items. Virtually all research demonstrating relationships between scales of HR variables and firm performance (e.g., Delery & Doty, 1996; Huselid, 1995; Huselid, Jackson, & Schuler, 1996) report internal consistency estimates of the HR scales, usually finding estimates above .60. These estimates seldom reach the .70 level suggested as a minimum by Nunally and Bernstein (1984).

Seldom has any research examined the amount of measurement error due to time. This largely stems from a lack of any longitudinal research efforts in the area of HR practices. Although Huselid has gathered data from the same firms in four different data collection efforts with two-year intervals between, by and large, no test-retest correlations have been reported. One explanation is the difficulty in interpreting these correlations as reliabilities, because it is

impossible to assess whether any deviation from one data collection to the next is due to error or true changes in HR practices. (In the one instance where Huselid and Becker (1996) did report a test-retest correlation, for a six month interval, using the the question, "the percentage of the workforce that was unionized," the observed correlation was .70.).Reporting reliability estimates that incorporate only one of the several potential sources of error leads to higher reliability estimates than would be obtained if these multiple error sources were recognized simultaneously using generalizability theory. Gerhart, Wright et al. (in press) used generalizability analysis to demonstrate that even when moderate error exists due to items (Spearman-Brown reliability of .709) and to raters (ICC of .612), the total amount of error in a measure still can be quite high (generalizability coefficient of .418).

Thus, research examining the HR – firm performance relationship has predominantly only assessed error due to items, and some might argue even looking at this error source is inappropriate. Internal consistency estimates assume that the items are intended be to sampled from a common domain, yet HR practices may not represent an underlying factor or construct. In such cases, HR practices are summed as an index, rather than a scale. In addition, assessing error due to time can be problematic in terms of identifying which variance is due to error versus which is due to actual changes in HR variables. Researchers have largely ignored error due to raters, which is particularly problematic if one discounts the appropriateness of assessing error due to items and recognizes the problems with assessing error due to time. These factors lead to the conclusion that getting good estimates of the amount of error due to raters is a critical issue with regard to measurement error in HR – firm performance research.

Assessing measurement error is important in any research field because of the impact that it has on observed relationships. Random error typically attenuates an observed relationship such that it will underestimate the true relationship between two variables. This is why Gerhart, Wright et al. (in press) noted that if all of the observed error in their study was random, then it would imply that the true estimated impact of HR practices would be such that a one standard deviation increase in HR practices would result in between 46 and 96 percent increases in firm performance. However, if the measurement error were systematic, then the observed effect size might actually be overestimating the true effect size. Thus, while we cannot estimate how much measurement error is random versus systematic, we can safely conclude that observed effect sizes are profoundly influenced by the amount of measurement error that exists. This points to the need to further estimate how much measurement error can exist in measures of HR practices.

Study 1

Method

As part of a larger study on the extent to which respondents' implicit performance theories might influence their reports of HR practices (available in Gardner, Wright, & Gerhart, 1999), data were gathered from multiple line and HR executives at 16 large corporations. Using the membership of the Center for Advanced Human Resource Studies at Cornell University, we solicited participation from the company sponsor representatives. For each company that agreed to participate we requested that at least 2 Senior HR and 2 Senior Line executives complete a survey regarding the HR practices that existed in their firm. Surveys were sent to the company contact who was asked to distribute them to the appropriate people.

Respondents were instructed to complete the survey with regard to the HR practices that existed in their corporation. They returned their individual surveys to the researchers in postage paid pre addressed envelopes. All responses were kept confidential.

Measures

Previous results from Gerhart et al. (in press) indicated strong correlations between reported HR practices for managerial and hourly jobs. Because of this, and in an effort to limit the number of items subjects would respond to, we focused on the former category. Respondents reported the percentage of managerial/professional/technical employees that were covered by a set of HR Practice items similar to those used by Huselid (1995), Becker and Huselid (1998), and Gerhart, Wright, et. al (in press). The items appear in Table 1.

Results

We computed intra-class correlations, ICC(1,1) and ICC(1,k) (Shrout & Fleiss, 1979) on each of the HR practice items and on the total scale. Bliese (2000) notes the different interpretations of ICC (1,1) as being either the proportion of variance explained by group membership or as the degree of reliability associated with a single respondent assessment of the group mean. Our use of ICC(1,1) was as an estimate of the reliability of a single respondent's report of an item, i.e., using a single respondent to estimate the group mean. ICC(1,k) estimates the reliability of using the average of k respondents' ratings to measure the item. In this case, there were an average of 2.81 respondents per item. The results of these analyses can be seen in Table 1.

As can be seen in Table 1, the ICC(1,1)'s ranged from .03 to .99 with a mean of .42 for the traditional HR practice items. We also constructed a scale summing these items which resulted in an ICC(1,1) of 26. The ICC(1,k)'s ranged from .09 to .99 with a mean of 60. The ICC(1,k) for the scale was .52.

Table 1. Interrater Reliability Estimates for Objective HR Items in Study 1.

Item	ICC(1,1)	ICC(1,k)
1. ...has their merit increase or other incentive pay determined by a performance appraisal?	.18	.38
2. ...receives formal performance appraisals?	.49	.72
3. ...is promoted based primarily on merit (as opposed to seniority?)	.76	.90
4. ...has any part of their compensation determined by a skill-based compensation plan?	.50	.73
5. ...is eligible for bonuses based on individual performance or company-wide productivity or profitability?	.19	.39
6. ...is regularly administered attitude/satisfaction surveys?	.99	.997
7. ...is administered an aptitude, skill, or work sample test prior to employment?	.55	.76
8. ...has access to a formal grievance procedure/complaint resolution system?	.17	.37
9. ...receives more than 40 hours of formal training per year on a regular basis?	.48	.69
10. ...receives sensitive information on the company's operating performance (costs, quality, etc.)	.03	.09
11. What proportion of non-entry level jobs have been filled from within over the past 5 years?	.55	.76
12. If the market rate for total compensation (Base+Bonus+Benefits) is considered to be the 50th percentile, what is your firm's target percentile for total compensation?	.69	.88
13. What proportional change in total compensation could a low performer normally expect as a result of a performance review?	.30	.55
14. What proportional change in total compensation could a high performer normally expect as a result of a performance review?	.15	.34
15. For the five positions that your firm hires most frequently, how many qualified applicants do you have per position (on average)?	.21	.45
HR Scale	.26	.52
Average for Items	.42	.60

Discussion

The results of this study demonstrate higher interrater reliability than that observed by Gerhart et al. (in press). The average item ICC(1,1) of .42 is higher than the .16 found previously. These results lend support for cautious optimism regarding the potential to find reasonably reliable measures of HR practices. On the other hand, the ICC(1,1) for the scale is .26, which implies a substantial correction factor (of four) for any observed regression coefficient using this scale as an independent variable in a firm performance equation.

It is important to understand why it might be that the reliabilities are higher here. First, by focusing attention on only one (albeit broad) category of jobs, respondents' information

processing requirements may have been lower. Second, by allowing the contact person to distribute the survey, it is possible that s/he selected a more likeminded individual to complete the second survey. Third, in looking at our data, we discovered that on the very high reliability items (#3, promotions based on merit and #6, regularly administered attitude/satisfaction survey), the significant between-company variance that is necessary to obtain a high reliability came in the form of all companies reporting very high values, except for one company, which reported very low values. If that one company had not been in the sample, reliability would have been much lower. Fourth, similar to Huselid and Becker's (in press) experience, we received a few phone calls from respondents indicating that they were engaged in significant data gathering to ensure accuracy of responses.

However, one should note that although these results exceeded those observed previously, they still fail to meet the .70 value considered to be a minimum for measurement (Nunnally & Bernstein, 1994). In addition, the analyses only focused on the error due to raters and ignored error due to items (i.e., internal consistency). So these values, while higher than those previously reported by Gerhart et al. (in press), still indicate that significant amounts of error exist in measures of HR practices. Indeed, as noted, they imply a large correction factor.

It should be recognized that this study in large part was simply a replication of the Gerhart, Wright et al. (in press) study. By using large, potentially diversified corporations and asking respondents to report practices across a variety of jobs, we run the risk of observing lower bound estimates of reliability. The nature of the sample sets up a situation where one would expect to find the lowest level of interrater reliability. What is also needed and called for by Huselid and Becker (in press) is a study which could provide an upper bound estimate for the reliability of HR practice measures. Such a study would have to control for as many extraneous sources of error as possible such as industry, size, diversification, etc. Studies 2 and 3 seek to accomplish this.

Study 2

Sample

A sample of 1050 commercial banks was drawn from the population of all U.S.-based banks with 350 banks being drawn from each of the following categories: assets greater than $25 million but less than or equal to $100 million, total assets greater than $100 million but less than or equal to $300 million, and total assets greater than $300 million [See Delery and Doty (1996)]. Data were gathered in the early 1990s by sending surveys to the HR director at each bank. HR directors and job incumbents in three jobs (personal trust officer, loan officer, and teller) completed the survey regarding the HR practices that existed for the three jobs. Thus, in

each case where matching data is available, one pair of respondents exists: the HR director and the job incumbent in the focal job.

Procedure

The HR director at each bank was sent a packet containing the survey materials. (S)he was asked to complete a separate survey regarding HR practices in each of the three jobs mentioned above. Job descriptions were enclosed to ensure that all respondents were clear on the exact jobs to be described. They were then asked to provide names of people in each of the three jobs. The HR director distributed surveys to the job incumbents. All surveys were returned in self-addressed stamped envelopes.

Respondents were asked to indicate their level of agreement/disagreement with statements regarding 28 HR practices using a likert-type scale. These items and their reliabilities appear in Table 2.

Results

Table 2. Interrater Reliability Results for Study 2.

			ICC (1,1)	ICC (1,k)
Internal Career Opportunities (scale)			.07	.13
	1.	Individuals in this job have clear career paths within the organization.	.06	.11
R	2.	Individuals in this job have very little future within this organization.	.15	.26
	3.	Employees' career aspirations within the company are known by their immediate supervisors.	.01	.02
	4.	Employees in this job who desire promotion have more than one potential position they could be promoted to.	.26	.41
Training (scale)			.29	.45
	1.	Extensive training programs are provided for individuals in this job.	.31	.46
	2.	Employees in this job will normally go through training programs every few years.	.16	.28
	3.	There are formal training programs to teach new hires the skills they need to perform their jobs.	.26	.41
	4.	Formal training programs are offered to employees in order to increase their promotability in this organization.	.20	.33
Results-oriented Appraisal (scale)			.18	.30
	1.	Performance is more often measured with objective quantifiable results.	.13	.23

	2.	Performance appraisals are based on objective, quantifiable results.	.15	.27
Employment Security (scale)			.17	.28
	1.	Employees in this job can expect to stay in the organization for as long as they wish.	.17	.29
	2.	It is very difficult to dismiss an employee in this job.	.16	.28
	3.	Job security is almost guaranteed to employees in this job.	.12	.21
	4.	If the bank were facing economic problems, employees in this job would be the last to get cut.	.00	.00
Participation (scale)			.21	.34
	1.	Employees in this Job are allowed to make many decisions.	.21	.34
	2.	Employees in this Job are often asked by their superior to participate in decisions.	.20	.33
	3.	Employees are provided the opportunity to suggest improvements in the way things are done.	.12	.21
	4.	Superiors keep open communications with employees in this Job.	.19	.31
Job Descriptions (scale)			.06	.11
	1.	The duties of this Job are clearly defined.	.09	.16
	2.	This Job has an up-to-date Job description.	.18	.31
	3.	The Job description for this Job contains all of the duties performed by individual employees.	.01	.02
R	4.	The actual Job duties are shaped more by the employee than by a specific Job description.	.11	.20
Profit-sharing				
	1.	Individuals in this Job receive bonuses based on the profit of the organization.	.47	.64
		Average for items	.16	.26
		Average for subscales	.21	.32

Note: An R next to an item indicates it was reverse scored.

The results indicate few differences from Gehart, Wright et al. (in press). The average ICC(1,1) across items was .16 and the average ICC(1,k) across items was .26. Rather than use the entire scale, subscale ICC's were reported based on the subscales used by Delery and Doty (1996). The average subscale ICC(1,1) was .21, and the average subscale ICC(1,k) was .32.

Discussion

We expected this study to provide higher reliability estimates than those reported by Gerhart, Wright et al. (in press) or in our Study 1, in part, because it focused on one industry. And, rather than having respondents assess HR practices across allmanagerial/professional/technical jobs, they were asked only about three very specific jobs. In addition, although some larger banks were included, the average size of the firms was much smaller than Gerhart, Wright, et al's (in press) or study 1 above.

However, one could argue that HR managers and employees have different perspectives, and these differences would work against finding interrater reliability. Thus, in Study 3, we sought to examine interrater reliability among job incumbents as well as between job incumbents and HR respondents.

Study 3

Sample

In an effort to control for as many extraneous sources of variance as possible and to provide best-case estimates of interrater reliability, we examined reliability in 190 job groups within 33 business units that were not only small in size, but also existed within a single corporation. The corporation is in the food service industry, and is organized to create an entrepreneurial spirit in each of its 62 businesses. To do this, businesses are limited in size to approximately 700 employees and $800 million in sales. When any individual business grows larger than this, it is broken into two separate businesses. Consequently, on average, the businesses have approximately 500 employees and approximately $500 million in sales with limited variance around these numbers. Thus, the business units in this sample are of a size significantly smaller than the average size in Huselid's (1995) sample.

In addition, because the businesses are all within the same corporation, they all use similar basic processes and similar basic technologies. However, they are managed with a principle called "earned autonomy." This provides that as long as a business is meeting its performance goals, Presidents are allowed almost complete autonomy in how they manage the business, particularly with regard to people. Only benefits and safety programs are mandated at corporate headquarters; the rest of the HR practices vary freely across businesses.

Finally, each of the businesses has 6 basic job categories: sales, warehouse, merchandising, delivery drivers, administrative staff, and supervisors. Corporate HR contacts assured us that these job groups represent potential differences in HR practices (i.e., within businesses, different practices exist for each grouping, but the same practices should exist within each grouping within a business).

Procedure

The data were collected as part of an employee attitude survey process instituted at this corporation. The VP of HR hoped to develop empirical evidence for tracing the impact of HR on business unit performance. Thus, in addition to the actual attitudinal items on the survey, we were allowed to collect information on the presence of a number of HR practices, both from the employees themselves and the senior HR Director within each business.

Employee surveys were developed by the researchers in conjunction with the top corporate HR staff. In addition, the research team developed a separate survey of HR practices for each job group to be completed by the HR Director. The core items were identical across the two surveys with a few additional items appearing on the HR Director survey (items that the HR staff believed might be too sensitive to present to employees). Corporate HR marketed the survey to the 62 business units and 33 agreed to participate (business unit response rate of 53%).

Business unit HR Directors were instructed to randomly select 20% or more of the employees in each job group from each of the three shifts. Each shift, employees met in groups on company time. HR Directors explained the purpose of the meeting, the survey process, and a timeline for results. The HR person distributed the surveys to employees, gave them time to complete them, and had the employees place the surveys in one large sealable envelope per meeting. The HR person then sent the unopened envelopes directly to the researchers. The response rate for employees in these meetings was 99.5%, with a total of 3,445 employees responding, yielding an average of 104 raters per business unit and 17.75employees per job category. The survey covered 19.6% of the total number of employees in the 33 participating business units.

HR Directors were also instructed to complete and return a survey of HR practices directly to the researchers. Thirty of the 33 HR Directors returned surveys for a response rate of 91%.

With exceptions noted on Table 3, response options for survey items for both the employees and HR managers were Yes, No, and I don't know. "Yes" responses were coded as 1 and "No" and "I don't know" responses were coded as zero for both groups. See Gardner, Moynihan, Park, and Wright (2000) for detailed justification for this procedure. To calculate ICC's we used the raw 1/0 indicators for each item for all employees in each job groups. To determine the existance of a practice in each of the job groups from the perspective of the employees the percentage of employees within the job group indicating "Yes" was calculated.

Results

In order to examine the amount of measurement error due to raters, we performed two sets of analyses. First, ICC's were computed on the data gathered from the employees to provide an estimate of the reliability of the measures from similarly situated individuals (i.e., all in the same jobs).

Second, Huselid and Becker (in press) and Gerhart et al. (in press) both distinguish between HR policies (the HR practices that are supposed to be taking place) and HR practices (those that are actually being done in the organization). They both agree that the actual practices are those that most likely impact employees, but they tend to disagree regarding which is actually being assessed when HR respondents are used. Thus, we computed correlations between measures gathered from employees in a job (without a doubt the HR practices) and those gathered from the HR Director (who was asked to respond regarding practices but might respond based on the policies that should exist rather than the practices that do exist). Table 3 reports all three analyses.

As can be seen in Table 3, the average ICC(1,1) for the set of HR practice items was .16 with a range of .05 to .28; the average ICC(1,k), where k = 17.75, was .71 with a range of .52 to .92. For the scale, the ICC(1,1) was .29 and the ICC(1,k) was .88. Surprisingly, the ICC(1,1) values barely differ from those observed by Gerhart, Wright et al. (in press). The ICC(1,k) values indicate that using an average based on multiple (17.75 in this case) raters provides a reasonably reliable measure of HR practices.

Table 3. Interrater Reliability for HR Items in Study 3.

Item	ICC(1,1)	ICC(1,k)	r_{pb} ee-HR
Applicants undergo structured interviews (job related questions, same questions asked of all applicants, rating scales) before being hired.	.08	.63	.11
Applicants for this job take formal tests (paper and pencil or work sample) before being hired.	.16	.78	.38**
On average how many hours of formal training do employees in this job receive each year?[b]	.20	.82	.28**
Employees in this job regularly (at least once a year) receive a formal evaluation of their performance.	.28	.88	.38**
Pay raises for employees in this job are based on job performance.	.26	.87	.42*
Employees in this job have the opportunity to earn individual bonuses (or commissions) for productivity, performance, or other individual performance outcomes.	.37	.92	.64**
Qualified employees have the opportunity to be promoted to positions of greater pay and/or responsibility within the company.	.11	.70	.24**
Employees in this job have a reasonable and fair complaint process.	.05	.52	.06
Item	ICC(1,1)	ICC(1,k)	r_{pb} ee-HR
Employees in this job are involved in formal participation processes such as quality improvement groups, problem solving groups, roundtable discussions, or suggestion systems.	.14	.75	.24**
Employees in this job communicate with people in other departments to solve problems and meet deadlines.	.13	.75	.36**
How often do employees in this job receive formal company communication regarding:[c]			
a. Company goals (objectives, actions, etc)?	.17	.79	.48**
b. Operating performance (productivity, quality, customer satisfaction, etc.)?	.17	.79	.34**
c. Financial Performance (profitability, stock price, etc.)?	.13	.73	.34**
d. Competitive performance (market share, competitor strategies, etc.)?	.11	.70	.23**
HR Index	.29	.88	.62[e]**
Average of Items	.16	.71	.30

*p < .05
** p < .01
[a] With the exception of those marked, the response option for these questions was “Yes, No, I don’t know.”
[b] Response option was “Hours _____” Dichotomized 1 = (>15 hrs) 0 = (< 15 hrs)
[c] Response options for these questions were: “Never, Annually, Quarterly, Monthly, Weekly, Daily.” Dichotomized 1 = Quarterly or more frequent; 0 = Never or Annually
d N varies from 178 to 186 job groups depending on missing data.
[e]Pearson's product moment correlation

In comparing the employees' responses with those of the HR Directors, point biserial correlations [correlations between a binary variable (data from the HR manager) and continuous variable (percentage of employees indicating the practice exists)] were computed for all job groups. The average r_{pb} across all items was .30 with a range of .06 to .64. The pearson correlation between the summed scales was .62.

Discussion

This study demonstrates that even when practices are measured within jobs, within sites, and within businesses (i.e., when virtually no true variance should exist in the HR policies and only minor differences should exist in practices), single respondent measures fall far short of acceptable reliability standards. One would be hard pressed to conceive of a situation where higher interrater reliability should exist. The units were small, single business, and single location. Compared to other research on HR practices, very little variance existed in size, technology, and products. The jobs were clearly defined. The respondents were the most knowledgeable (both incumbents and HR directors).

Under even these ideal circumstances, the average single-item interrater reliability failed to reach .20 and the scale failed to reach .30. In addition, significant disagreement existed between employees in the job and the HR Director with responsibility for that business with correlations averaging .24.

Discussion

Gerhart, Wright et al. (in press) reported low levels of reliability for single respondent measures of HR practices in research on the HR – firm performance relationship. Huselid and Becker (in press) raised legitimate concerns regarding the generalizability of those results, and this paper sought to explore the extent to which those results might generalize to other research studies in this area. This paper has presented data from three different studies shedding light on the extent to which measurement error due to raters might exist in research on HR practices at the firm level. The results generally confirm Gerhart, Wright et al's findings regarding the low reliabilities of single-respondent HR measures.

In Study 1, we expected to observe reliability levels consistent with those reported by Gerhart et al. (in press, a), who used senior HR executives in large corporations. Although we observed higher reliabilities here, we continued to observe considerable amounts of measurement error.

Study 2 provides evidence on the convergence between ratings provided by HR Directors and job incumbents. We found lower convergence than in Study 1, despite the

presence of design factors (e.g., single industry, only three specific jobs rated) that one would ordinarily expect to contribute to higher, not lower, reliabilities. In addition, Study 3 also demonstrated very little convergence between the HR manager closest to the actual shop floor HR practices and those employees impacted by those practices. It is unclear if these results, together, imply that HR managers are accurately reporting the HR policies and employees simply are not seeing those policies operationalized into their respective practices, or that HR respondents are simply unaware of the actual practices that exist. However, it certainly implies that HR managers may not be an ideal source for assessing the actual HR *practices* that exist.

Most surprising perhaps were the results of Study 3, which we viewed as almost a best case scenario for obtaining high reliabilities. By focusing on HR practices within jobs and within locations that were highly similar and relatively small (around 500 employees per location), we expected to find considerably higher reliability. However, this did not prove to be the case. As such, the Study 3 results are consistent with those obtained in Studies 1 and 2 and those obtained by Gerhart, Wright et al. (in press, a, b). In 3 of the 4 studies that report interrater reliabilities by item, we observe ICC(1,1)s in the range of .15-.17. Although interrater reliabilities of items were higher in the fourth study, they still fell far short of acceptable levels. In the case of scale interrater reliabilities, levels in the range of .25 to .30 were the general rule (although higher in the refineries studied by Wright et al, in press, b). Thus, our findings combined with those of our earlier studies would seem to demonstrate quite clearly that our concerns about low the interrater reliability of HR practices in single-respondent designs is not specific to any one sample, contrary to Huselid and Becker's (in press) hypothesis). Rather, the finding of low interrater reliability generalized across samples that vary according to unit size, industry diversity, and whether few or many jobs are the focus of measurement.

Interestingly, the values obtained across these studies are quite consistent with those seen in much of the groups or multi-level research (See Gerhart, 1999). Thus, our finding of low interrater reliability should perhaps not be surprising. What is different, however, is that these other literatures use a design (multiple raters) that significantly reduces the reliability problem, whereas in the strategic HRM literature, single respondent studies continue to predominate.

Implications for Research

First, consistent with the suggestions of Gerhart, Wright et al. (in press) these results indicate the need to exercise caution in interpreting effect sizes obtained in past substantive research on the relationship between HR and firm performance. Our findings strongly suggest that significant random error exists in single-respondent measures of HR practices. Second, although we did not address the possibility of systematic error (which would typically result in

existing effect sizes being upwardly biased, we believe that our findings should raise broader measurement concerns like this. To date, only one study (Gardner et al. 1999) has attempted to determine if systematic error *can* exist, but that study did not seek to determine if it *does* exist.

Second, our findings certainly call for considerable methodological attention to be focused on ways of reducing the amount of measurement error. The most obvious way of reducing this error is through increasing the number of raters. Gerhart, Wright et al (in press) described how generalizability studies demonstrate a greater payoff from adding raters than from adding items. Study 3 described above provided empirical justification for this suggestion. Although the reliability of using any individual rater (ICC1,1) was inadequate, the reliability from using multiple raters (ICC1,k) achieved reasonable levels. However, practical considerations may preclude researchers from obtaining an average of 17+ raters per job in many cases..

Another way to increase reliability would be to develop better measures of HR practices. Huselid and Becker (in press) noted that they have consistently upgraded their measures with every wave of data collection, a very commendable effort. They have made the items more specific to reduce error in interpretation by raters and more accurately represent practices known to be effective (e.g., validated employment tests rather than simply tests). Certainly the more specifically worded the item, the greater the expected reliability. However, the items used in Study 3 above were both objective and clearly written, yet still unreliable when based on a single respondent.

Another avenue for increasing reliability might be through exploring different rating scales. Currently no consensus exists as to the proper rating scale. For example, Huselid and his colleagues understandably lean toward an objective scale, using ratings of the percentage of employees covered by a given practice. Study 3, above, asked respondents to indicate whether the practice was in existence or not. However, others have used more subjective rating scales. Delery and Doty (1996; study 2 above) used a likert type scale measuring the extent of agreement with statements regarding the use of HR practices. Snell and his colleagues as well as Wright, McMahan, Sherman, McCormick (1999) generally asked subjects to respond regarding the extent to which practices were used. At this point one cannot say with any certainty that the empirical data supports the use of one format over another. However, the converse is also true: the data do not support the idea that "objectifying" the rating eliminates, or even reduces, measurement error.

Additionally, attention should focus on ensuring that the most knowledgeable rater(s) are used. Huselid and Becker (in press) correctly argue that adding poorly informed raters

increases neither reliability nor validity. They suggest that in their experience, the recipient of the survey often sends it to multiple people, each of whom possesses specialized knowledge regarding the functional practices that exist in the organization (e.g., the compensation specialist completes the compensation practice items, etc.). Certainly instructions should clearly indicate who should complete the whole, or parts, of the survey to ensure that naï ve raters are not introducing error into the measures. In addition, researchers could include accuracy checks on the survey, asking who completed each section with their title, and/or asking for the respondent's confidence in rating items.

Relatedly, little consensus exists regarding whether this vein of research should attempt to assess the HR *policies* (the practices that the organization seeks to have implemented) or the HR *practices* (the actual practices that employees are subjected to in their work roles). Huselid and Becker (2000) and Gerhart et al. (2000) both seem to agree that the practices provide the real mechanisms that should impact employee behavior and skills, and thus, firm performance. However, if one assumes that employees are the best sources of this information, then Studies 2 and 3 seem to indicate that HR managers do not provide an accurate source of this information. Certainly, the choice of HR policies or HR practices should not be taken lightly in the design of future research.

Also, researchers must attend to the information processing requirements of survey completion to aim the measures at a level that will permit respondents to feasibly provide reliable and valid information. For example, surveys that require respondents to describe practices that exist across multiple job groups, businesses, industries, and geographic locations most likely far exceed any individual's information processing capabilities. Surveys that more narrowly focus on a job group (e.g., top executives) , a single business, or a single location should present respondents with information processing requirements more within their capabilities. Note, however, that this implies different respondents. If one focuses on top executives, then the individual charged with leadership development might be more appropriate than the Senior VP-HR. If focused on single business, then the Director or VP of HR in that business might be a better respondent than the corporate VP-HR. If focused on a single site, then the manager/director of HR for the site might be the most appropriate respondent.

While this study has focused entirely on the reliability of survey methodology, one also might consider alternative data collection methods. For example, while MacDuffie (1995) used surveys, he also followed up the survey at a sample of sites to verify the accuracy of the survey responses by studying the archival records. Additionally, Welbourne and Andrews (1996)

coded prospectuses from firms going through initial public offerings. These researchers have demonstrated that the field should not solely limit itself to survey designs.

Finally, rather than letting the performance measures drive the design, more attention should be focused on letting the sample or design drive the performance measure. As Rogers and Wright (1999) noted, it appears that much of the current HR – firm performance literature starts with corporate performance measures because they are publicly available. Then HR practices are measured at the corporate level because that is the level of the dependent variable. Little attention is paid to the logic, feasibility, and desirability of measuring HR practices at the corporate level given the variety and complexity of HR practices that exist within corporations of any reasonable size. In addition, while profitability seems to be the ultimate criterion to which we seek to tie HR practices, corporate profitability measures may mask profound variance in profitability of business units, mirroring the variance in HR practices.

Research conducted by Huselid and his colleagues has provided a valuable start in understanding how managing human resource might affect firm performance. Rather than 10-15 more such studies with the same limitations, the field would achieve greater contributions from studies at different levels of analysis where practices are more uniform and performance measures less distal from the effects of these practices.

Conclusion

This paper presented three studies demonstrating that measures of HR practices from single respondents contain unacceptably high levels of measurement error. This error exists regardless of the size or complexity of the organization. Such high levels of measurement error make interpretations of observed effect sizes difficult.

Our results challenge researchers interested in strategic HRM to pay considerably more attention to designing studies that minimize measurement error. Although gathering data from multiple respondents best achieves this goal, we have also suggested other design changes that we hope will also contribute to future research that helps us draw more confident conclusions regarding how HR practices impact firm performance.

References

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

Bliese, P. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (eds.), Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions. San Francisco, CA: Jossey-Bass. Pp. 349-381.

Chadwick, C. 2000. Empirical insights on the origins of synergies in strategic human resource systems. Paper presented at the 2000 Academy of Management Meeting, Toronto, CANADA

Delery, J. 1998. Issues of Fit in strategic human resource management: Implications for research. Human Resource Management Review, 8(3): 289-310.

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

Gardner, T., Moynihan, L., Park, H., Wright, P. 2000. Unlocking the black box: Examining the processes through which human resource practices impact business performance. Paper presented at the 2000 Academy of Management Meeting, Toronto, ON, Canada.

Gardner, T., Wright, P., & Gerhart, B. 1999. The HR-firm performance relationship: Can it be in the eye of the beholder? Paper presented at the 1999 Academy of Management Meeting, Chicago, IL.

Gerhart, B. (1999). Human resource management and firm performance: Measurement issues and their effect on causal and policy inferences. Research in Personnel and Human Resources Management, Supplement 4, 31-51.Gerhart, B., Wright, P., & McMahan, G. (in press, b). Measurement error and estimates of the HR- firm performance relationship: Further evidence and analysis. Personnel Psychology, 53:

Gerhart, B., Wright, P., McMahan, G., & Snell, S. (in press). Measurement error in research on human resources and firm performance: How much error is there and how does it influence effect size estimates? Personnel Psychology, 53:

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

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

Huselid, M. & Becker, B in press. Comment on Measurement error in research on human resources and firm performance: How much error is there and how does it influence effect size estimates? Personnel Psychology, 53:

MacDuffie, J. 1995. Human resource bundles and manufacturing performance: Organizational logic and flexible production systems in the world auto industry. Industrial and Labor Relations Review. 48: 197-221.

Nunnally, J. & Bernstein, I. 1994. Psychometric Theory. New York: McGraw-Hill, Inc.

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.Welbourne, T. & Andrews, A. 1996. Predicting the performance of initial public offerings: Should human resource management be in the equation? Academy of Management Journal, 39: 891-919.

Wright, P., McCormick, B., Sherman, S. & McMahan, G., 1999. The role of human resource practices in petro-chemical refinery performance. International Journal of Human Resource Management, 10: 551-571.