# Making Stronger Causal Inferences: Accounting for Selection Bias in Associations Between High Performance Work Systems, Leadership, and Employee and Customer Satisfaction

Joseph A. Schmidt

University of Saskatchewan

Dionne M. Pohler

University of Toronto



Author Note

Joseph A. Schmidt, Edwards School of Business, University of Saskatchewan; Dionne M. Pohler, Centre for Industrial Relations and Human Resources, University of Toronto.

Correspondence concerning this article should be addressed to Joseph Schmidt, Edwards School of Business, University of Saskatchewan, 25 Campus Drive, Saskatoon, SK, Canada S7N 5A7.

This research was funded by the Social Sciences and Humanities Research Council of Canada (#430-2014-00383).

# Making Stronger Causal Inferences: Accounting for Selection Bias in Associations Between High Performance Work Systems, Leadership, and Employee and Customer Satisfaction

## Abstract

We develop competing hypotheses about the relationship between high performance work systems (HPWS) with employee and customer satisfaction. Drawing on eight years of employee and customer survey data from a financial services firm, we employed a recently developed empirical technique — covariate balanced propensity score (CBPS) weighting — to examine if the proposed relationships between HPWS and satisfaction outcomes can be explained by reverse causality, selection effects, or commonly omitted variables such as leadership behavior. The results provide support for leader behaviors as a primary driver of customer satisfaction, rather than HPWS, and also suggest that the problem of reverse causality requires additional attention in future HR systems research. Model comparisons suggest that the estimates and conclusions vary across CBPS, meta-analytic, cross-sectional, and time-lagged models (with and without a lagged dependent variable as a control). We highlight the theoretical and methodological implications of the findings for HR systems research.

**Keywords:** human resource management, leadership, employee satisfaction, customer satisfaction, endogeneity, longitudinal, covariate balanced propensity score

Scholars began to report robust relationships between HR systems and firm performance outcomes in the 1990s (e.g., Arthur, 1994; Huselid, 1995; MacDuffie, 1995), which led to a surge of empirical research in strategic HR management over the next two decades (Paauwe, 2009). Much of the research suggests meaningful associations between HR systems and a variety of employee and organizational outcomes (see Jackson, Schuler, & Jiang, 2014; Subramony, 2009); a comprehensive meta-analysis showed that "high performance" HR management systems influenced firm financial outcomes via human capital, employee motivation, voluntary turnover, and operational performance (Jiang, Lepak, Hu, & Baer, 2012).

Despite the large body of literature supporting the efficacy of high performance work systems (HPWS), some have questioned the methods employed in strategic HR studies. For instance, Wright and colleagues (2005) highlighted that many HR-performance studies up to that point were either cross-sectional or even post-predictive (i.e., HR practices were measured after performance), demonstrating the potential for reverse causality. Most research designs now ensure that measures are collected in the appropriate temporal sequence; however, past performance is still often found to be correlated with future HR practices, making it difficult to ascertain causal direction (Paauwe, 2009). Moreover, while HR scholars usually attempt to mitigate bias associated with common method variance and measurement error in research designs, we argue that reverse causality, omitted variables, and selection effects often go unchecked, or even unmentioned, in most studies.

Concerns about bias – elsewhere referred to as endogeneity – in non-experimental research are certainly not new (Cook & Campbell, 1979; Rubin, 1974). More recently, Antonakis and his colleagues (2010, 2014) levied critiques against the social science literature, arguing that researchers often make causal claims – or tacitly assume causality – without addressing major threats to causal inference. It is understandable that many strategic HR studies do not meet the conditions required to infer causality. Random assignment is next to impossible or difficult to implement in most field settings, naturally occurring experiments are rare, and lab studies are criticized for lacking external validity. It is also challenging to obtain agreement from organizations and participants to collect longitudinal data, and the extensive investment of resources required to collect data over time can be prohibitive. Moreover, if empirical results are consistent with established theory and meta-analytic techniques can account for the limitations in primary research to produce correlational estimates that approach the “true” population effects, should we be concerned about drawing causal inferences from any single study?

While there should always be a place in the scientific process for correlational designs and descriptive research to highlight new phenomena, develop insights, and challenge theoretical assumptions, we argue that strategic HR and applied psychology scholars should pay more attention to causal inference for two reasons. First, the Open Science Collaboration (2015) has raised concerns about the reproducibility of findings in psychological science. The Collaboration suggested that some of the practices that may lead to reproducibility problems in experimental studies include “selective reporting, selective analysis, and insufficient specifications of the conditions necessary or sufficient to obtain the results” (p. aac4716-1) as well as publication bias. Combining these practices with additional sources of bias present in observational research suggests that the reproducibility problem may be exacerbated in studies that rely on non-experimental data. Others have also called the reproducibility of meta-analyses into question, arguing that measurement error and flexible inclusion criteria can alter meta-analytic findings (Lakens, Hilgard, & Staaks, 2016). Meta-analyses also suffer from a “garbage in, garbage out” problem if appropriate methods are not applied (Rosenthal & DiMatteo, 2001). If causal inferences cannot be confidently drawn about the estimates in primary studies, meta-analytic estimates are not causal either. As a field matures, subjecting its theories and associated causal assumptions to more rigorous empirical tests becomes essential for scientific advancement.

Second, it is difficult to develop robust theory in the absence of rigorous empirical designs. In a recent critique of the role of theory in organizational sciences, Cucina and McDaniel (2016) argue that a hypothesis must have strong empirical support that is both methodologically rigorous and well-replicated before it can be considered a theory. This approach to theory development is consistent with approaches in other scientific disciplines and requires reasonably strong causal inference to ensure that the theory has a “high probability of being correct” (Cucina & McDaniel, 2016, p. 1117). As Sutton and Staw (1995) state, “Theory emphasizes the nature of *causal* relationships” (italics added, p. 378); thus, research designs that allow for stronger causal inference are necessary to develop good theory.

Our study makes some important contributions in this area. First, we follow in the tradition established by others (e.g., Antonakis et al., 2010; Wright et al., 2005) to highlight important yet often overlooked sources of bias that applied researchers should consider when designing non-experimental studies. Second, we highlight an econometric technique – covariate balanced propensity score (CBPS) weighting – that can be useful for reducing the impact of selection effects and omitted variable bias in non-experimental research designs. Propensity scoring has rarely been adopted in management research to date (Connelly, Sackett, & Waters, 2013), partly because of the minor influence economics has had on empirical techniques in HR research, and partly because statistical packages required to apply propensity scoring to continuous treatment variables have only been developed recently (e.g., Fong, Hazlett, & Imai, 2014; Fong, Ratkovic, Hazlett, & Imai, 2015). Third, our access to a longitudinal dataset (over eight years) allows for the comparison of estimates from the propensity scoring approach to estimates using techniques that have been previously employed in organizational research, such as pooled cross-sectional analysis and lagged dependent variable analysis. Finally, we answer calls to test conflicting theories and competing hypotheses through the development of more rigorous research designs (Cucina & McDaniel, 2016). We are thus able to shed light on the debate over whether it is the HR system or something else that explains the associations between HR systems with employee and organizational outcomes. The approach we adopt provides insight into which theoretical perspective is best explained by the data and contributes to future theory development in the strategic HR field.

More specifically, we develop and test hypotheses about the effects of HPWS on employee and customer satisfaction, and also develop and test additional hypotheses to determine if reverse causality, selection effects, and/or commonly omitted variables, such as leader behaviors, explain the associations between these constructs. We further compare the standardized regression coefficients and relative weights from the covariate balanced propensity score models to meta-analytic, cross-sectional, and lagged regression models to determine the extent to which selection bias may influence the results of designs that are more common in the strategic HR literature.

We note that the choice of study variables is partly opportunistic; we exploit secondary data compiled from eight years of employee and customer surveys in our focal firm. However, investigating HPWS, employee satisfaction, and customer satisfaction as our primary variables of interest is also justifiable. Researchers are paying closer attention to the proximity or "distance" between predictors and outcomes. For instance, Paauwe (2009) proposes that we need "performance indicators that are far more proximal in terms of what HR practices can actually affect, such as changes, for example, in employee attitudes" (p. 135). Financial indicators (e.g., profitability and market value) are often affected by numerous external and internal factors, and have weak associations with HR systems compared to measures of employee attitudes and customer satisfaction (Jiang et al., 2012). Employee satisfaction – an overall feeling of positive affect toward one's job and employer – has been shown to at least partially result from processes controlled by the organization, including the HR system (Messersmith et al., 2011; Takeuchi, Chen, & Lepak, 2009; Wu & Chaturvedi, 2009). Customer satisfaction is influenced by employees' service performance, which has been shown to be affected by high performance work systems and individual and collective employee attitudes and subsequent behaviors (Chuang & Liao, 2010; Liao et al., 2009). Moreover, one of the key roles of an HR system is to provide a context to ensure employees have the ability, motivation, and opportunity to contribute to the organization's goals (e.g., Jiang et al., 2012), which, in our study's focal organization, requires attracting and retaining customers.

**Causal Effects of HR Systems on Employee and Customer Satisfaction**

Much of the strategic HR research to date has focused on so-called high-performance work systems. While definitions vary, these systems normally refer to specific combinations of HR practices – including "flexible job assignments, rigorous and selective staffing, extensive training and development, developmental and merit-based performance appraisal, competitive compensation, and extensive benefits" (Takeuchi, Lepak, Wang, & Takeuchi, 2007, p. 1069). HPWS are designed to foster greater employee commitment and motivation, leading to higher individual and organizational performance.

Different theories have been offered to explain the "black box" of causal processes linking HPWS to performance outcomes. For instance, scholars proposed that employee attitudes and behaviors are important mediating variables between HR systems and firm-level outcomes (Purcell & Kinnie, 2007). One common perspective suggests that firms will perform well when HR systems provide employees with the ability, motivation, and opportunity to demonstrate desired behaviors (Barrick, Thurgood, Smith, & Courtright, 2015; Blumberg & Pringle, 1982; Kehoe & Wright, 2013). Another influential theory proposes that well-designed HR systems can create “strong” organizational cultures or climates that influence employee attitudes and behaviors (Bowen & Ostroff, 2004), and some research has supported this perspective (Askoy & Bayazit, 2014; Katou, Budhwar, & Patel, 2014).

Finally, applications of social exchange theory (Blau, 1964) suggest that employees who receive HR practices focused on investments in their personal welfare and career development will form a relational psychological contract with the organization that engenders positive organizational attitudes (Rousseau, 1995). Employees will develop a felt obligation to reciprocate organizational support (i.e., Eisenberger, Armeli, Rexwinkel, Lynch, & Rhoades, 2001) and are more likely to remain loyal to the organization (Batt & Colvin, 2011; Shaw, Dineen, Fang, & Vellella, 2009), which is indicative of overall employee commitment and satisfaction with their employers.

Consistent with the foregoing theory and empirical research, a comprehensive meta-analytic structural equation model showed that the standardized path coefficients between high performance work systems with employee attitudes and operational outcomes (including customer satisfaction) were .62 and .34, respectively (Jiang et al., 2012). We also conducted our own meta-analysis specific to our particular study variables using metaBUS to derive more relevant effect size estimates for our study (see the Appendix for a more detailed explanation of the methods and the meta-analytic correlation matrix).[1] The meta-analytic correlation between HPWS and positive job affect (which included employee job satisfaction) was .12, and the correlation between HPWS and customer satisfaction was .27. Our meta-analytic effect sizes differ from those reported in Jiang et al. (2012) because we only focused on variables directly relevant to this study, and drew upon a more limited number of samples. We thus propose the following hypothesis based on the aforementioned HR systems theory and meta-analytic estimates of HPWS effect sizes on employee attitudes and operational outcomes:

*Hypothesis 1:* Perceptions of HPWS have small to moderate positive effects on (a) employee satisfaction and (b) customer satisfaction.

**Alternative Explanations for the Relationship between HR Systems and Outcomes**

The foregoing body of theoretical and empirical work has been critical in elevating the status of HR as a field; however, limited conclusions can be drawn from studies that do not adequately address major sources of bias. While it is true that non-experimental research designs do not all suffer from the same threats to validity, bias can be a more serious problem than some researchers acknowledge. As Antonakis et al. (2010) point out: “If *x* is endogenous the coefficient of *x* simply has no meaning. The true effect could be higher, lower, or a completely different sign” (p. 1088).

One of the major problems plaguing HR systems research is that there are many possible alternative explanations (i.e., reverse causality, omitted variables, selection effects) for the observed relationship between HR systems and employee and organizational outcomes. Many of these alternative explanations are often not measured or accounted for in non-experimental

[1] metaBUS (www.metabus.org) is an ongoing project to curate the findings in the applied psychology and management literatures (see Bosco, Steel, Oswald, Uggerslev, & Field, 2015; Bosco, Uggerslev, & Steel, 2017). It contains a search engine that provides meta-analytic estimates and links to the original research.

research designs, and thus many studies do not meet the required conditions for inferring causality: “(1) *x* must precede *y* temporally; (2) *x* must be reliably correlated with *y* (beyond chance); and (3) the relation between *x* and *y* must not be explained by other causes” (Antonakis et al., 2010, p. 1087).

Since the publication of Wright et al. (2005), HR research designs usually ensure that measures are collected in the appropriate temporal sequence (Paauwe, 2009) and positive empirical associations have been reliably established between HPWS and various outcomes (Jiang et al., 2012). With regard to the third condition of causal inference, HR researchers have made strides in addressing specific threats to validity – in particular, non-independence of observations and clustering (with multi-level data), measurement error, and common-method variance (e.g., Liao, Toya, Lepak, & Hong, 2009). Yet, the problem of omitted variable bias, and its special case of omitted selection, is more intractable and often goes unaddressed in organizational research (Connelly et al., 2013).

Experimental designs where participants are randomly assigned to control and treatment conditions generally meet the third condition of causal inference. Random assignment creates control and treatment groups composed of individuals with roughly the same distribution of observed and unobserved characteristics. This allows for a direct test of the counterfactual argument (i.e., what would have happened if an individual in the treatment group was assigned to the control group or vice versa?). Randomly controlled experiments are therefore the gold standard for making causal inferences (Cook & Campbell, 1979). When only non-experimental data is available, however, other approaches are required.

When participants are not randomly assigned, the observed and/or unobserved variables that cause their selection to the treatment or control conditions (or the value/level of continuous treatment variables) may also cause the observed outcome. For example, it is difficult to determine the causal effect of a training program on job performance if researchers are unable to randomly assign employees to training and control conditions. If only highly conscientious employees choose (or are chosen by their managers) to attend the training program, it is possible that the employees' conscientiousness, and not the training, influenced their subsequent job performance. Gelman and Hill (2007) discuss how this type of research design allows us to make *predictive inferences* – conscientious employees who receive training are likely to perform well in the future. However, it does not allow us to make causal inferences about the effect of training on employee performance because we are unable to test counterfactual arguments. In other words, it may not be possible to determine if the training would have improved the performance of less conscientious employees, or if highly conscientious employees would have performed poorly if they did not receive training.

**Reverse Causality in HR Systems Research.** As previously highlighted, in a seminal publication on the relationship between HR practices and firm performance, Wright and colleagues (2005) demonstrated that the positive association observed in studies of HR and firm performance may be due to reverse causality. While most research designs now ensure that measures are collected in the appropriate temporal sequence, past performance is still often found to be correlated with future HR practices (Paauwe, 2009).

Shin and Konrad (2014) argue that HR theories often treat organizations as closed systems where the causal arrow from HR systems to both employee- and firm-level outcomes only points in one direction; however, high firm performance produces slack resources that allow for greater subsequent investments in HR systems. General systems theory suggests that feedback loops create bidirectional associations between performance and HR system adoption and implementation. For instance, performance may influence leader perceptions about the efficacy of HR systems and their decisions about whether or not continued implementation of the systems is warranted. Empirical research is suggestive of reciprocal associations between HR systems and productivity over time (Shin & Konrad, 2014), and between customer satisfaction and employee job satisfaction (Zablah, Carlson, Donovan, Maxham, & Brown, 2016).

Implicit performance theories may provide another explanation for reverse causation. Gardner and Wright (2009) posited that research respondents have implicit theories about HR systems that “may bias the recall of information in a way consistent with the theory the researcher is trying to test” (p. 58). In an experiment, they found that both executives and MBA students relied on company performance information to respond to survey items about HR practices. Employees who have positive attitudes about working for the organization, and/or perceive that the organization is performing well, may engage in post hoc rationalization and attribute these outcomes to the HR systems in the organization.

Some research does not support reverse causality explanations, showing that HR system implementation precedes changes in employee attitudes and organizational outcomes (Koys, 2001); however, another stream of research provides stronger support for reverse causality. For instance, Schneider, Hanges, Smith, and Salvaggio (2003) found that previous financial performance was a stronger predictor of firm-level job satisfaction rather than the other way around. Ryan, Schmidt, and Johnson (1996) found that customer satisfaction predicted subsequent job satisfaction. Finally, Guest, Michie, Conway, and Sheehan (2003) unexpectedly found that previous year profitability was associated with HR system implementation, but not vice versa. HR studies that do not account for reverse causality in their study designs may be ignoring a major source of endogeneity. We thus test the following hypothesis:

*Hypothesis 2:* (a) Employee satisfaction and (b) customer satisfaction have a reverse causal effect on employee perceptions of HPWS.

**Leader Behavior as an Omitted Variable in HR Systems Research.** Scholars have criticized HR research for ignoring the influence of omitted variables, not least of which is the role of leadership (e.g., Wright et al., 2005). Indeed, in one study of the relationship between high performance work systems and customer service, Chuang and Liao (2010) acknowledged that leadership was a major form of omitted variable bias in their research design (p. 182).

Empirical research on senior leadership teams and front-line managers is consistent with the idea that leaders' attitudes and behaviors may be an important omitted variable in the strategic HR literature. Research has shown that top management teams were more likely to implement high performance and innovative human resource practices when they believed in the efficacy of HR systems and valued employee welfare (Arthur, Herdman, & Yang, 2016; Osterman, 1994). A field experiment showed that bank managers in historically high performing branches were more likely to implement HR practices after receiving training than managers in lower performing branches (Krackhardt, McKenaa, Porter, & Steers, 1981). While these studies do not provide concrete evidence for leadership-as-third-variable explanations, they do suggest that effective leaders may be more likely to adopt and implement high performance HR systems. When considered together with the extensive body of research about the effects of leader behaviors on employee attitudes, motivation, and group or organizational performance (see the meta-analysis by Judge & Piccolo, 2004), there is a possibility that leadership is causing both the adoption and implementation of high performance HR systems, as well as affecting employee attitudes and operational performance. We thus test the following hypotheses:

*Hypothesis 3:* Leader behavior is an omitted variable responsible for the observed relationships between perceptions of HPWS with (a) employee satisfaction and (b) customer satisfaction.

*Hypothesis 4:* Leader behavior is an omitted variable responsible for the observed reverse relationships between (a) employee satisfaction and (b) customer satisfaction with perceptions of HPWS.

**Addressing Selection Bias in Non-Experimental Data**

Since all potential outcomes cannot be observed for the same individual or organization, the researcher is required to construct or select an appropriate counterfactual or control group (see Rubin, 1974; 2005). We adopt a procedure – covariate balanced propensity score weighting – that helps to overcome some of the limitations of non-experimental data and traditional approaches to estimating propensity scores. The covariate balanced propensity score procedure combines the two purposes of propensity score analysis by estimating the likelihood of treatment based on observed variables and optimizing covariate balance (Friedman, 2012). Through the application of a set of covariate balancing weights, this allows for the construction of a counterfactual or comparison group by "matching" the treated individuals (e.g., those who received greater HR system investments) with individuals that did not receive the treatment, but are otherwise similar to the treated individuals on observed variables that affect their selection into the treatment. While propensity scoring approaches assume that selection is based on observed variables, which is a strong assumption, it may also improve the balance of unobserved variables between the groups, particularly if the lagged dependent variable is also used to estimate the propensity score (Angrist & Pischke, 2009). At minimum, this approach forces researchers to consider design issues separately from model estimation by explicitly considering potential omitted variables and selection effects.

It is important to note that simply controlling for selection variables in regression models may not eliminate the problems with selection bias. Even if the appropriate controls are included in the model, the sample may not be balanced across the outcomes on the selection variable(s), leading to an extrapolation problem due to lack of common support (i.e., lack of overlap in the covariate distribution between the levels of treatment: Angrist & Pischke, 2009). In our previous training example, if mostly conscientious employees received the training, controlling for conscientiousness would not allow us to say much about how training would have affected outcomes for less conscientious employees. With regard to the current research, it is plausible that positions with more experienced employees or with greater managerial responsibility were likely to have received higher HR system investments, making it difficult to determine if high HR system investments increased satisfaction outcomes in positions with less experienced employees or lower managerial responsibility. It may also be possible for selection variables to influence response tendencies and artificially inflate associations between survey constructs, which further highlights the value of applying this method to observational survey data. As outlined below, we entered a number of potential selection variables in the covariate balancing propensity score procedure, including measurement period, tenure with the organization, gender, management responsibility, and the lagged dependent variable. Leader behavior was only included as a selection variable in the tests of Hypotheses 3 and 4.

## Methods

We obtained longitudinal data from a large financial services organization in Canada. The firm has over 2,200 employees and a variety of locations throughout the country and it has operated at a high level of sustained profitability for a number of years. Most of the data was derived from eight waves of an employee survey that was administered during the same month each year from 2006 to 2013. The total response rates were quite high in each year of data collection, ranging from 80% (2011 and 2012) to 91% (2006). The high response rates reduce concerns about non-response (one form of selection bias). The customer satisfaction outcomes were collected from separate surveys of the firm's customers over the same time period. We removed all senior leader responses from our analyses because senior leaders likely referred to themselves or their peers when responding to questions about leadership. The archival data was anonymous to both the researchers and the employer, making it exempt from review by the Institutional Ethics Committee.

**Level of Analysis and Sample Sizes**

The level of analysis is an important methodological consideration in our research design. The focal independent variables in this study were derived from surveying employees about their attitudes and characteristics of the work environment. Such employee perceptions can be viewed as the psychological climate, defined as "perceptions that assess the meaning of work environments to individuals" (James et al., 2008, p. 8), and are likely influenced by both individual and situational characteristics. The effects of individual differences on psychological climate perceptions creates some degree of within-group (or between person) variance in climate perceptions; however, common environmental influences should create a reasonable degree of convergence in climate perceptions such that it can be defined as a group-level construct (see Schneider, Ehrhart, & Macey, 2013). We therefore operationalized HR systems, leader behaviors, and employee satisfaction as collective psychological climate constructs. Based on information obtained from the company, employees working in similar positions in the same business unit are subject to the same HR systems and supervisors. We thus applied the direct consensus composition model (Chan, 1998) and aggregated individual responses to the level of positions within business units after checking agreement indices. Variance did exist across the HR systems at this level and is consistent with other research that suggests firms differentiate their HR systems across occupation groups within the same organization (Lepak & Snell, 2002; Schmidt, Pohler, & Willness, 2018). The five position types included in the final analysis were administrative, front-line employee, professional, front-line supervisor, and manager.

We applied different assumptions to the customer satisfaction measure. While customer service outcomes are more proximally related to HR than financial performance indicators (e.g., profit), there are still numerous factors that could drive customer satisfaction in the financial services industry, only one of which is the internal organizational climate and direct interaction with front-line employees. Therefore, we applied additive composition assumptions (Chan, 1998) to aggregate individual customer responses to the business-unit level. That is, we do not make assumptions about the distribution or agreement in satisfaction ratings among a business unit's customers, only that higher mean levels are indicative of greater customer satisfaction.

The final sample for models that included employee satisfaction as the dependent variable consisted of aggregate data from an average of 3.98 employees in an average of 257.88 positions per year (positions are nested within a yearly average of 78.87 business units). There was some year-to-year fluctuation in the position- and business unit-level samples sizes because the organization engaged in some minor restructuring during the data collection period, which resulted in both the addition and elimination of positions and business units. Senior HR personnel who had been with the organization during the entire measurement period provided us with the necessary information to match the positions and business units across years.

Customer satisfaction data was only available at the business-unit level for a subsample of front-line employees and units with direct customer interaction. We report the results from an exit survey (a survey distributed after customers decided to leave the organization), which were then matched to the employee survey responses of the front-line employees who had direct contact with the customers. The exit survey data was based on an average of 25.50 business units per year provided by an average of 19.28 customers per business unit per year. Again, there was some fluctuation in the number of business units due to restructuring and some smaller business units did not lose any customers in some years. An average of 10.79 front-line employees provided data from each business unit per year.

**Measures**

With the exception of customer satisfaction, all other measures were included in a proprietary employee survey administered by an external consulting firm. The employee survey measures have undergone rigorous psychometric assessment with responses from employees in more than 7,000 organizations worldwide. The consulting firm provided us with the raw data that contained anonymized responses for each individual employee, so we could independently assess the psychometric properties of the survey measures. Unless otherwise specified, respondents answered the questions on a 6-point Likert scale (1 = strongly disagree to 6 = strongly agree).

**HR System (HPWS).** Employee perceptions of the HR system were measured with three items about compensation, three items about job design, and three items about training and development opportunities. These constructs are consistent with the ability-motivation-opportunity framework of strategic HR management and measures of high performance work systems (e.g., Jiang et al., 2012). Effective training and development practices provide employees with the skills and abilities to perform well, compensation practices motivate employees to meet performance objectives, and well-designed jobs provide employees with opportunities to apply their skills and motivation. An example of a compensation item is, "My performance has a significant impact on my pay." "The tools and resources I have allow me to be as productive as possible," is a sample job design item. A sample training and development item includes, "There are sufficient opportunities in this organization for me to improve my skills in my current role."

**Leader Behavior.** Seven items measured respondents' perceptions about the leadership behaviors of their immediate supervisor. Sample items included "My manager inspires me to do my best work every day," and "My manager effectively deals with poor performance in our team."

**Employee Satisfaction.** Employee satisfaction was measured with five items, two of which included, "I truly enjoy my day-to-day work tasks" and "I feel like I 'fit in' well here." Because we were unable to rely on measures from well-validated scales of employee satisfaction, we collected data in a separate construct validation study to determine the validity of our measure. We report the results of this study in Appendix A of the supplemental online materials.[2]

**Customer Satisfaction**. As previously mentioned, customer satisfaction ratings were taken from an exit survey administered to former customers after they decided to leave the organization. Customers provided various reasons for leaving the organization: 18.6% of respondents chose to leave due to interest rates, service fees, or lack of relevant loan products; 6.6% cited poor customer service, inflexibility, or feeling unappreciated; 46.5% stated that they no longer needed financing, were retiring, or were consolidating loans elsewhere; the remaining 28.3% did not respond to the question or chose the "other" category.

[2] We thank an anonymous reviewer and the editor for suggesting this approach to provide additional validity evidence for our measure.

The survey consisted of five questions and each question was rated on a different five-point scale. Two example questions are, "How satisfied are you with the overall level of service you received from X?" (rated from 1 = not at all satisfied to 5 = completely satisfied) and "Compared with other financial institutions, how would you rate X in terms of overall value you receive?" (rated from 1 = much less value than other institutions to 5 = much more value than other institutions). Per the confirmatory factor results reported in Appendix B in the supplemental online materials, we used three of the five questions to create the scale scores.

**Selection and Control Variables.** A number of variables were used to create the propensity score weights described later and also subsequently entered as controls in the regression analyses. We controlled for the time period (i.e., survey year) to account for time varying trends in unmeasured factors that influenced the survey responses. Demographic diversity, including gender, has been linked to relationship conflict and satisfaction with teams (Thatcher & Patel, 2012), and may also be an important form of omitted selection bias in our particular firm; thus, respondent gender aggregated (averaged) to the position by business-unit level was also entered as a control. Employee tenure with the organization was entered as a control and also represents a type of selection effect – longer serving employees likely received greater HPWS investments and may have chosen to remain because they are satisfied with the organization. The years of service variable (tenure) was measured on an ordinal scale (1 = < 1 year, 2 = 1-2 years, 3 = 3-5 years, 4 = 6-10 years, 5 = 11-15 years, 6 = 16-20 years, 7 = 21-25 years, 8 = > 25 years); therefore, all eight of the categories aggregated to the position level were entered as controls. A dummy variable for people management responsibility (0 = does not manage others, 1 = manages others) was entered as a control for the analyses only for employee satisfaction. Managers likely have different perspectives on HR systems and effective leader behaviors compared to front-line workers, which may have influenced their survey responses. Management responsibility was not entered as a control in analyses with the customer satisfaction data because only front-line workers, who did not have management responsibilities, provided the employee data for these analyses. Finally, the lagged dependent variable was also included as a selection and control variable, which helps balance the sample on unobserved variables that may have affected selection into the treatment (or level of the continuous independent variable) and also accounts for time-varying trends in the dependent variable (Angrist & Pischke, 2009).

**Factor Structure and Data Aggregation**

The multilevel reliability and agreement indices for the direct consensus composition constructs (HPWS, leader behavior, and employee satisfaction) are reported in Table 1. We report both the multilevel omegas and alphas (Geldhof, Preacher, & Zyphur, 2014) for comparison. Although Geldhof et al. found that the omega values provided the least biased estimates in most situations, they suggested that the alpha statistic may be more appropriate when there are lower ICC values and fewer than 15 participants per cluster, which characterizes some time periods of our data. With one exception,[3] the estimates exceeded .70 for every construct and time period at both levels of analysis. The ICC(1), ICC(2), and $r_{wg}$ statistics are also reported in Table 1. The ICC(1) values were statistically significant and within acceptable ranges for the variables across most years of the survey, indicating that there was between-group variance in the survey constructs. The ICC(2) values were somewhat lower than desirable, which is primarily a function of the small group sample sizes. The mean $r_{wg}$ values were .70 or greater

[3] The between-level omega for employee satisfaction at time three was .48. This may have been due to the relatively lower ICC(1) value that year and some model convergence problems (i.e., there was a negative between-level item variance). The alpha values and agreement indices suggest that between-level reliability was adequate that year.

for all constructs across all measurement periods, suggesting a reasonably high level of rater agreement. Of note, the ICC values for perceptions of HPWS were relatively low in year 6 (2011), which suggests more within-group variance that year. Controlling for measurement period in the propensity score matching and regression models helps account for some of the time-specific organizational factors that may have influenced response patterns and within-group agreement on these constructs. We also conducted multilevel confirmatory factor analysis with the employee survey constructs and the results generally fit the data well. The results are all reported in Appendix B of the supplemental online materials.

## Analysis

**Covariate Balanced Propensity Score Weighting**. Standard propensity scoring techniques are generally applied when the predictor is a dichotomous variable consisting of treatment and control conditions (e.g., received training or not). Logistic regression is conducted on the predictor variable to create a propensity score, which is defined as a "conditional probability that expresses how likely a participant is to be assigned or to select the treatment condition given certain baseline characteristics" (Thoemmes & Kim, 2011, p. 92) – the baseline characteristics being the selection variables. Individual cases in the treatment and nontreatment conditions are matched on the basis of the propensity score to create distributions of the baseline characteristics that are balanced in each condition. Distributional balance is required to test counterfactual arguments (Gelman & Hill, 2007).

Continuous survey variables are often split into binary "treatment" and "nontreatment" conditions to apply traditional propensity scoring procedures. This approach requires an arbitrary judgment about how to dichotomize the variable (i.e., where to make the split), which results in a loss of information and raises some concerns about model misspecification. Fong et al. (2014)

also show that covariates balanced on a dichotomized variable are not necessarily balanced on the original continuous treatment variable. Therefore, Fong et al. (2014) extended Imai and Ratkovik's (2014) propensity scoring methodology and developed a procedure that can be applied to continuous treatment variables. This covariate balanced propensity score procedure assigns weights to each observation, which optimizes covariate balance and minimizes correlations between the treatment and covariates. The weighting scheme is subsequently applied in regression models to estimate the effects.

We conducted this analysis with the CBPS package (version 0.10) developed by Fong et al. (2015) in R 3.1.5. The first step in the procedure is to estimate the propensity scores and probability weights for the cases. As described previously, measurement period (i.e., year), gender, years of service, people management responsibility, and the lagged dependent variable likely affected how people responded to survey items and/or influenced the selection of cases into higher or lower levels of the treatment (i.e., level of HPWS). Therefore, these variables were entered as the selection variables to derive the propensity score. People management responsibility was not used as a selection variable in the analyses involving customer satisfaction because these analyses only involved front-line employees. As displayed in the boxplots in Figure 1, the mean and variance of the correlations between the covariates and the HPWS variable decreased substantially after the propensity score balancing procedure was applied, indicating that the procedure improved covariate balance in the sample. Although not displayed, the weighting procedure produced a very similar pattern of effects for leadership and satisfaction, which were the treatment variables in certain models.[4]

[4] Researchers often "trim" their data based on the propensity score to ensure greater common support/overlap in the covariate distribution between the levels of treatment (Angrist & Pischke, 2009). To the authors' knowledge, standard guidelines for trimming using the CPBS procedure have not been developed to date, and thus we decided not to trim our data.

**Lagged Dependent Variable Regression with the Covariate Balanced Propensity Score (CBPS)**. After the CBPS weights were assigned to the cases in the data file, we conducted lagged dependent variable regression analyses to estimate hypothesized effects. Lagged dependent variable regression (Angrist & Pischke, 2009) controls for time-varying, pre-measurement variable trends. To conduct this analysis, information from all waves of data collection were entered into a vertical panel data file. The independent (*x*) and dependent (*y*) variables were time-lagged to ensure temporal precedence, and all regression models controlled for time, the lagged dependent variable of interest and relevant covariates. The lagged dependent variable regression with CBPS is expressed as:

$$Y_{t+1} = a + bX_{t1\ldots k} + bX_{Yt1\ldots k} + bX_{CBPSt1\ldots k} + \cdots + bX_{mt1\ldots k} + e$$

where *t* = time and *CBPS* = covariance balanced propensity score treatment variable. We also calculated the cluster-adjusted standard errors in Mplus 7.3 to account for the nested nature of the data (i.e., non-independence of positions within business units).[5]

After cases were removed due to the loss of one year to accommodate the lagged dependent control and selection variable, the employee satisfaction regression models based on the employee survey data contained 1,607 observations, and the customer satisfaction models from the surveys of employees and exiting customers consisted of 163 observations. As previously mentioned, the sample size was smaller for the customer satisfaction outcome because not all business units had direct interactions with customers.

[5] Inconsistent inference is another common threat to validity and occurs when researchers do not adjust their standard errors to account for non-independence of observations (Antonakis et al., 2010). Propensity score weighting can bias standard error estimates. Although the CBPS package adjusts standard error estimates in weighted regression with non-clustered data, we are unaware of an adjustment procedure for cluster-robust standard errors or multilevel regression models.

**Relative Weight Analysis.** Comparing standardized coefficients to determine the relative importance of variables in a regression model is problematic if predictor variables are correlated (Tonindandel & LeBreton, 2011). Therefore, we conducted relative weight analysis as an alternative effect size estimate using the R code developed by Tonidandel and LeBreton (2015), which addresses collinearity among predictors by orthogonally transforming the variables and provides information about the proportional contribution of each covariate to the total model variance. We report the relative weight (RW), which is interpreted as the variance in the dependent variable ($R^2$) accounted for by the independent variable.

**Comparative Analyses.** Following our main analyses, we also conducted pooled cross-sectional analysis using OLS regression as well as non-CBPS time-lagged dependent variable regression (both with and without the lagged dependent variable as a control) to compare the results and effect sizes from our approach to those that are more commonly employed in the strategic HR literature. We also conducted regression analysis with the meta-analytic correlation matrix reported in the Appendix. Results across these models are reported in the same tables to facilitate comparison.

## Results

The descriptive statistics and correlations among the study variables are reported in Table 2. Tables 3 and 4 show the results of the hypotheses tests, and Appendix C of the supplemental online materials shows the results of additional analyses with the leader behavior variable. While we use the lagged dependent variable CBPS model results to test the hypotheses, comparison of effect sizes between different models is also highly informative given the purpose of our study, and we highlight comparisons between models in the next section.

Hypothesis 1 states that perceptions of HPWS cause (a) employee and (b) customer satisfaction. The results displayed in Table 3 indicate that Hypothesis 1 was not supported as the regression coefficients from the CBPS models were not statistically significant. The relative weights indicate that HPWS also accounted for very little variance in the employee and customer satisfaction models.

Turning to the other explanations, Hypothesis 2 states that (a) employee and (b) customer satisfaction have a reverse causal effect on perceptions of HPWS. The CBPS models in Table 3 show that the effect of employee satisfaction on HPWS was not statistically significant, but the effect of customer satisfaction on HPWS was significant ($\beta$ = .14, *SE* = .06, *p* = .016, *RW* = .02). Thus, Hypothesis 2b received support.

Hypotheses 3 and 4 address omitted variable explanations. Hypothesis 3 states that leader behavior is an omitted variable explaining associations between HPWS with (a) employee and (b) customer satisfaction. As shown in Table 4, HPWS was not related to employee or customer satisfaction when leader behavior was entered as a control and selection variable; thus, Hypothesis 3 was not supported. Hypothesis 4 was also unsupported as the reverse causal effect of employee satisfaction on HPWS was still not statistically significant when leadership behavior was entered in the model. Similarly, the reverse effect of customer satisfaction on HPWS remained significant when leader behavior was entered as a selection variable ($\beta$ = .21, *SE* = .10, *p* = .033, *RW* = .02). Leader behavior was not an omitted variable that explained this effect.

**Leader Behavior Results and Model Comparisons**

While not the primary focus of our study, leader behavior appeared to have a causal effect on customer satisfaction (see Online Appendix L: $\beta$ = .14, *SE* = .07, *p* = .046, *RW* = .02). Given the weaknesses associated with strict reliance on significance testing, it is also informative to compare the range of effect sizes produced by the different models. Figure 2 compares the relative weights produced by the CBPS models to the range of relative weights from the metaBUS, cross-sectional, and time-lagged models for both the employee and customer satisfaction variables. In most cases, the CBPS models produced relative weights that were near the bottom of the range of weights produced by other models, and in some cases the differences appeared to be substantial. For instance, the CBPS model indicated that perceptions of HPWS accounted for one percent of the variance in employee satisfaction; yet, the cross-sectional model suggested that HPWS accounts for 41% of the variance. Similarly, the relative weight for the reverse causal effect of employee satisfaction on HPWS was .01 in the CBPS model, while it was as high as .42 in the cross-sectional model. The differences in relative weights between the customer satisfaction models were not as extreme and, in some cases, the weights produced by the CBPS models fell well within the range of weights from the other models. Overall, the results from the CBPS models were most similar, though not identical, to the results using the time-lagged models that included the lagged dependent variable as a control.

## Discussion

We have presented a complex series of results; however, there appear to be two effects for which we can make reasonable causal inferences with this data: (1) leader behavior predicted customer satisfaction, and (2) customer satisfaction predicted HPWS. In other words, our results suggest leader behavior, rather than HPWS, is the primary driver of customer satisfaction, and that perceptions of HPWS actually result from the reverse causal relationship with customer satisfaction.

Our data ultimately allows us to conduct a constructive replication of previous research in two ways: 1) we replicate previous designs (e.g., cross-sectional, time-lagged) used in other studies, and generally find a pattern of results that is consistent with prior research; and 2) we replicate across different designs using the same data (e.g., cross-sectional, time-lagged, CBPS), and we find different results with the CBPS design. This suggests that selection bias could be a problem in previous research that has linked HPWS with employee and organizational outcomes.

Many theoretical advances have been made in our understanding of how HR systems and leader behaviors may contribute to employee attitudes and organizational performance. Empirical studies are increasingly addressing concerns associated with temporal measurement of variables (Wright et al., 2005), common-method variance, and construct measurement and development (e.g., Gerhart, Wright, McMahon, & Snell, 2000). However, causal inference remains somewhat elusive in much observational HR/OB research (Antonakis et al., 2010; Paauwe, 2009), reducing scholars' ability to develop and test theory, or offer managers prescriptive advice. If managers have scarce resources to dedicate to improving employee attitudes and ultimately enhancing organizational performance, it is often not clear if they should invest in HPWS, in leadership development, in both, or in neither. We propose that our study and results make three important theoretical and methodological contributions to the literature, with associated implications for future research and practice.

The first primary contribution is methodological and highlights that selection bias may play a particularly important role in understanding previous associations between HPWS and employee and organizational outcomes. Antonakis et al. (2010) summarize seven categories of major threats to causal inference including: omitted variables, omitted selection, simultaneity (including reverse causality), measurement error, common-method variance, inconsistent inference, and model misspecification. HR systems research is now at the point where scholars who design their studies pay detailed attention to issues of timing, measurement, and collection of data from multiple sources (Paauwe, 2009). Notwithstanding these welcome advances to empirical studies in this area, many non-experimental research designs still ignore some of the most common and problematic threats to validity, specifically, omitted variables, selection bias, and reverse causality/simultaneity (Antonakis et al., 2010). Consequently, an important contribution of our study is to highlight how these sources of endogeneity continue to pose threats to causal inference in HR systems and leadership research, and hinder theoretical and empirical developments in these fields.

In short, endogeneity matters, and researchers must take care to account for its various sources in their research designs. While there are some ways to account for selection effects in cross-sectional designs (i.e., instrumental variable analysis; see Angrist and Pischke, 2009), it is very difficult to find a justifiable instrumental variable in survey-based research. Moreover, lagged outcome variables cannot be included as selection or control variables when there is only one data collection point. Cross-sectional research designs can produce misleading results and should rarely be relied upon for deductive hypothesis testing, unless it is clear that the independent variable has immediate effects on the outcome (e.g., when studying discrete emotional states).

Another noteworthy finding was that the CBPS analysis produced different results than many of the other designs, and the effect sizes were sometimes larger and sometimes smaller (or near zero). This serves as an important illustration that bias can result in effect sizes that are qualitatively different from those obtained using more rigorous causal designs. Selection biases increase the risk of committing either Type I *or* Type II errors. For example, Connelly et al. (2013) applied propensity score matching to demonstrate how effects can change direction after accounting for selection bias. In a sample of SAT test-takers, they found that the effect of the treatment (test coaching vs. no test coaching) on SAT scores changed from negative to positive after pre-test scores were included in the propensity score-matching model. Students who scored poorly on the pre-test were more likely to seek out test coaching and were also more likely to receive lower scores on the post-test than others, even after receiving the treatment. The negative treatment effect disappeared when the treatment and control groups were matched on the pre-test scores. In our study, a similar phenomenon may have occurred for the effect of HPWS on employee satisfaction, which became small and nonsignificant in the CBPS model. Employee satisfaction in previous years may have affected HPWS investments in subsequent years due to unobserved managerial decisions, communication about HPWS, and leadership behavior toward employees (indeed, note that employee satisfaction had a reverse causal effect on leader behavior, see Online Appendix M). Models that did not include prior satisfaction as both a control and selection variable may have thus overstated the relationship between HPWS and employee satisfaction. For instance, while the model that controlled for lagged satisfaction also became non-significant, it still overstated the relative weight compared to the CBPS model.

The second important contribution has both methodological and theoretical implications. Given our results suggest that HPWS is more likely to be caused by customer satisfaction than the other way around, from a theoretical perspective, this suggest that HPWS are likely implemented due to organizational slack, or as an employee reward for good operational performance, rather than in an attempt to improve performance. Indeed, qualitative interviews we conducted with leaders in this company suggest that employees are partially compensated based on customer satisfaction outcomes. Good customer satisfaction outcomes may have resulted in higher subsequent compensation, or it may also be due to employees' implicit performance theories (i.e., employees perceive the organization to be performing well, and attribute this to HPWS). This result also suggests that HR systems may not be a strong motivation-enhancing mechanism for employees and may be better viewed as either a benefit or a reward; future research should thus seek to investigate whether HPWS is better viewed as an incentive or a reward/benefit. From a methodological perspective, if researchers propose to test the effects of HR or leadership on operational outcomes, at minimum they must ensure they measure and control for the effect of prior performance and/or employee attitudes about the organization's performance. CBPS or matching designs that include prior performance as a selection variable improve upon lagged designs by accounting for possible selection bias.

The final major contribution is primarily theoretical. HPWS did not have robust effects on either employee or customer satisfaction, at least in the company we investigated. Managers and HR professionals may care about HPWS insofar as it serves as a benefit that attracts employees, or for ethical/humanistic reasons, but investments in leadership appear to be more important for improving customer satisfaction. Moreover, it is unlikely that leaders enhanced customer satisfaction solely by improving employee attitudes given that leadership had a relatively weaker effect on employee satisfaction in the CBPS model (see Online Appendix K: $\beta$ = .09, $SE$ = .05, $p$ = .066, $RW$ = .01). Indeed, prior research suggests there may be only weak to moderate associations between job satisfaction and job performance (Iaffaldano & Muchinsky, 1985; Judge, Thoresen, Bono, & Patton, 2001), and effective leaders may be better able to motivate employees through goal setting, setting clear role expectations, and providing ongoing feedback on job performance, which may not necessarily enhance employee attitudes (e.g., Latham & Yukl, 1976). Moreover, the effectiveness of HR practices may be dependent on leaders implementing them properly, highlighting the value of continuing to study how line manager communication and implementation of HR systems affects employee and organizational outcomes (e.g., Den Hartog, Boon, Verburg, & Croon, 2013; Sikora, Ferris, & Van Iddekinge, 2015). We did not measure all possible mediating mechanisms or outcome variables in our study and we do not advocate for organizations to halt HPWS implementation based on these results. Rather, the results suggest that research testing the causal effects of HPWS and leadership on employee motivation and goal commitment may be more fruitful than studying their impact on employee job attitudes.

If employee satisfaction is not affected by HPWS (or leadership), what is its cause? One possible factor could be the employee's disposition. Some research is consistent with this assertion showing that genetic influences account for substantial variance in job satisfaction irrespective of situational influences (e.g., Hahn, Gottschling, Konig, Spinath, 2016). Another could be our inability to account for recruitment and selection practices in the HPWS measure. If organizations develop HR practices designed to attract and select employees with positive affect, the HR system may still have a causal effect on employee attitudes that we were unable to identify. Future research should specifically seek to include recruitment and selection measures in HR studies of this nature and also measure other employee attitudes that were not investigated in this study.

**Limitations**

While we attempted to account for a number of sources of endogeneity, concerns remain and we would be remiss if we did not advocate caution when interpreting these findings. Our study is still limited to using non-experimental field data that is subject to a number of measurement issues, as well as other threats to validity, which we outline in detail below.

The anonymous nature of the survey made it impossible to match individual responses over time or to specific customers. There were theoretical reasons to aggregate the constructs to a higher level of analysis and we demonstrated that the factor structure of the constructs was similar across levels. The aggregation process has been shown to reduce measurement error because individual response errors tend to average out (Bliese, 1998). Nevertheless, it is possible that there were different processes operating at each level of analysis and it may not be realistic to generalize these results to individuals.

There are remaining questions about the construct validity of the survey measures given that they had not been previously validated in the academic literature. Although the confirmatory factor analyses and supplemental validity study for the employee satisfaction measure provided some construct validity evidence, there are still concerns about the inclusiveness and representativeness of the survey items. As mentioned previously, the HPWS measure did not include any items about recruitment and selection, and the employee satisfaction measure may have contained items that captured perceptions of person-job or person-group fit. Given the robust associations between these measures in many of the cross-sectional and lagged DV analyses, we doubt that construct validity problems were the source of null effects in the CBPS models. However, future research should seek to address the potential deficiencies in these measures and continue to ensure that endogeneity problems are not exacerbated by measurement issues.

Given that employees provided data for many of the study constructs, we likely were unable to completely eliminate concerns about common method variance. However, a number of features of the initial research design reduced its potential impact: individual responses were aggregated to the group level minimizing the effect of individual response tendencies; there was a one-year time lag between independent and dependent variable measurement; and the customer satisfaction data was collected from a different source. These features of the research design and the data indicate that common method bias likely did not have undue influence on the final results.

Customers who responded on the exit surveys were by definition leaving the organization, and this population should not be generalized to ongoing customers with the firm. For instance, the organization's exiting customers may have been more willing to provide honest feedback, which reduces measurement error in the dependent variable, but they may have also been less motivated to respond, increasing non-response bias. Some research on employee exit interviews is consistent with the former assertion showing that employee responses about their reasons for quitting appeared to be more forthcoming in surveys that were administered after employees had officially left the organization as compared to exit interviews that occurred during their last week on the job (see Giacolone, Knouse, & Monttagliani, 1997). Conversely, other research on consumers and member-based organizations suggests that customers who choose to stay with an organization are more likely to express their dissatisfaction to improve the organization than those that exit (Hirschman, 1970). Future research should focus on the nature of the population providing the outcome measures and how this might impact the findings.

We were able to test for some omitted variable and selection problems in strategic HR systems research, not least of which is accounting for the potential omitted selection effects of leadership, organizational tenure, gender, and managerial responsibility. We were, however, unable to measure other potential omitted variables (e.g., organizational culture), though given that our study is in one firm, these variables may be more likely to be held constant across units. Related, it is possible that the organization implemented major changes to the HR system or leadership development programs prior to the measurement period, which caused discontinuous improvements in employee attitudes or customer satisfaction. The initial implementation may have spillover effects that are not adequately captured using lagged dependent variables in the construction of our propensity scores, and may underemphasize the role played by the HR system. That is, the analysis may simply capture a steady state following a discontinuous change that occurred prior to the measurement period. While our data spans a relatively long time frame (eight years), and we know based on interviews with company managers that some changes occurred in the firm's HR systems during our time period, this explanation cannot be completely ruled out.

Using covariate balanced propensity scores theoretically allowed us to construct a more appropriate counterfactual group; however, it is important to note that this procedure is not a panacea to the omitted selection problem. The covariate balanced propensity score approach requires an assumption that selection is based on observable characteristics, and many decisions are still required about the specifics, such as the variables that should be included in the balancing equations and the choice of matching algorithms (Calidendo & Kopeinig, 2005).[6] While CBPS has been shown to outperform alternative propensity score/matching approaches in some situations (Wyss et al., 2014), high quality data and detailed knowledge of the institutional context are required to be able to understand and/or measure the selection process and variables, which we had access to in this study, but which may reduce the appropriateness of this method for adoption in other non-experimental studies, particularly when they are conducted across firms. Nevertheless, determining the appropriateness of different propensity score balancing/matching approaches should, at minimum, require researchers to consider how to construct an appropriate counterfactual, leading to more rigorous study designs overall.

[6] There are also analytic strategies that can be adopted when selection is based on unobservable characteristics (see Blundell & Costa Dias, 2002).

## Conclusion

To adequately develop and test theory using non-experimental data and designs in strategic HR systems research, scholars should endeavor to apply methods and techniques widely used in other disciplines and fields to allow for stronger causal inference. Recently, controversy has risen over the role and development of theory in management research (Ashkanasy, 2016; Cortina, 2016), with some suggesting that we are becoming more concerned about the “entertainment value of theories than we are with their scientific rigor or real-world value” (Mathieu, 2016, p. 1132). Others have similarly argued that extensive and rigorous empirical testing is required before hypotheses should be labelled “theory” (Cucina & McDaniel, 2016). Theory-building and rigorous empirical designs/methods, however, are two sides of the same coin; they are mutually constitutive and both necessary for accuracy and causal inference. By paying attention to both theory about HR systems and requirements for causal inference, scholars can identify threats to validity, adopt designs and empirical techniques established to address many of these problems, and by doing so, develop theories that better explain the real world.

**References**

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91,* 340-345.

Aksoy, E., & Bayazit, M. (2014). The relationships between MBO system strength and goal-climate quality and strength. *Human Resource Management, 53,* 505–525.

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21,* 1086–1120.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D.V. Day (Ed.), *The Oxford Handbook of Leadership and Organizations* (pp. 93-117). New York: Oxford University Press.

Arthur, J. B. (1994). Effects of human resource systems on manufacturing performance and turnover. *Academy of Management Journal, 37*, 670–687.

Arthur, J. B., Herdman, A. O., & Yang, J. (2016). How Top Management HR Beliefs and Values Affect High-Performance Work System Adoption and Implementation Effectiveness. *Human Resource Management, 55*, 413–435.

Ashkanasy, N. M. (2016). Why we need theory in the organization sciences. *Journal of Organizational Behavior, 37,* 1125-1131.

Barrick, M. R., Thurgood, G. R., Smith, T. A., & Courtright, S. H. (2015). Collective organizational engagement: Linking motivational antecedents, strategic implementation, and firm performance. *Academy of Management Journal*, *58,* 111–135.

Batt, R., & Colvin, A. J. (2011). An employment systems approach to turnover: Human resources practices, quits, dismissals, and performance. *Academy of Management Journal, 54*, 696–717.

Blau, P. M. (1964). *Exchange and power in social life*. New York: Wiley.

Bliese, P. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods*, *1*, 355-373.

Blumberg, M., & Pringle, C. D. (1982). The missing opportunity in organizational research: Some implications for a theory of work performance. *Academy of Management Review, 7*, 560–569.

Blundell, R., & Costa Dias, M. (2002). Alternative approaches to evaluation in empirical microeconomics. *Portuguese Economic Journal*, *1*, 91–115.

Bosco, F. A., Steel, P., Oswald, F. L., Uggerslev, K. L., & Field, J. G. (2015). Cloud-based meta-analysis to bridge science and practice: Welcome to metaBUS. *Personnel Assessment and Decisions*, *1*, 3-17.

Bosco, F. A., Uggerslev, K. L., & Steel, P. (2017). metaBUS as a vehicle for facilitating meta-analysis. *Human Resource Management Review*, *27*, 237-254.

Bowen, D. E., & Ostroff, C. (2004). Understanding HRM-firm performance linkages: The role of the "strength" of the HRM system. *Academy of Management Review*, *29*, 203–221.

Caliendo, M., & Kopeinig, S. (2005). *Some practical guidance for the implementation of propensity score matching.* IZA Discussion Paper No. 1588: Institute for the Study of Labor.

Cammann, C., Fichman, M., Jenkins, D., & Klesh, J. (1979). The Michigan organizational assessment questionnaire. *Unpublished manuscript, University of Michigan, Ann Arbor*.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*, 234-246.

Chuang, C., & Liao, H. (2010). Strategic human resource management in service context: Taking care of business by taking care of employees and customers. *Personnel Psychology*, *63*, 153-196.

Connelly, B. S., Sackett, P. R., & Waters, S. D. (2013). Balancing treatment and control groups in quasi-experiments: An introduction to propensity scoring. *Personnel Psychology*, *66*, 407–442.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Cortina, J. M. (2016). Defining and operationalizing theory. *Journal of Organizational Behavior, 37,* 1142-1149.

Cucina, J. M., & McDaniel, M. A. (2016). Pseudotheory proliferation is damaging the organizational sciences. *Journal of Organizational Behavior, 37,* 1116-1125.

Den Hartog, D. N., Boon, C., Verburg, R. M., & Croon, M. A. (2013). HRM, Communication, Satisfaction, and Perceived Performance: A Cross-Level Test. *Journal of Management*, *39*, 1637–1665.

Eisenberger, R., Armeli, S., Rexwinkel, B., Lynch, P. D., & Rhoades, L. (2001). Reciprocation of perceived organizational support. *Journal of Applied Psychology*, *86*, 42-51.

Fong, C., Hazlett, C., & Imai, K. (2014). Covariate balancing propensity score for general treatment regimes. *Princeton Manuscript*. Retrieved from: http://imai.princeton.edu/research/files/CBGPS.pdf.

Fong, C., Ratkovic, M., Hazlett, C., & Imai, K. (2015). Package "CBPS*." Comprehensive R Archive Network.* Retrieved from: ftp://203.199.213.69/cran/web/packages/CBPS/CBPS.pdf, March 10, 2016.

Friedman, J. (2012). Tools of the trade: The covariate balanced propensity score. Retrieved from: http://blogs.worldbank.org/impactevaluations/tools-of-the-trade-the-covariate-balanced-propensity-score, February 26, 2018.

Gardner, T. M., & Wright, P. M. (2009). Implicit human resource management theory: a potential threat to the internal validity of human resource practice measures. *International Journal of Human Resource Management*, *20*, 57–74.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*, 72-91.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Gerhart, B., Wright, P. M., McMahan, G. C., & Snell, S. A. (2000). 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*, 803-834.

Giacalone, R. A., Knouse, S. B., & Montagliani, A. (1997). Motivation for and prevention of honest responding in exit interviews and surveys. *Journal of Psychology*, *131*, 438-448.

Guest, D. E., Michie, J., Conway, N., & Sheehan, M. (2003). Human resource management and corporate performance in the UK. *British Journal of Industrial Relations*, *41*, 291–314.

Hahn, E., Gottschling, J., König, C. J., & Spinath, F. M. (2016). The heritability of job satisfaction reconsidered: Only unique environmental influences beyond personality. *Journal of Business and Psychology*, *31*, 217-231.

Hirschman, A. (1970). *Exit, voice, and loyalty*. Cambridge, MA: Harvard University Press.

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.

Iaffaldano, M. T., & Muchinsky, P. M. (1985). Job satisfaction and job performance: A meta-analysis. *Psychological Bulletin*, *97*, 251-273.

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B* (Statistical Methodology) *76*, 243–263.

Jackson, S. E., Schuler, R. S., & Jiang, K. (2014). An aspirational framework for strategic human resource management. *Academy of Management Annals*, *8*, 1–56.

James, L. R., Choi, C. C., Ko, C.-H. E., McNeil, P. K., Minton, M. K., Wright, M. A., & Kim, K. (2008). Organizational and psychological climate: A review of theory and research. *European Journal of Work and Organizational Psychology*, *17*, 5–32.

Jiang, K., Lepak, D. P., Hu, J., & Baer, J. C. (2012). How Does Human Resource Management Influence Organizational Outcomes? A Meta-analytic Investigation of Mediating Mechanisms. *Academy of Management Journal*, *55*, 1264–1294.

Judge, T. A., & Piccolo, R. F. (2004). Transformational and transactional leadership: A meta-analytic test of their relative validity. *Journal of Applied Psychology*, *89*, 755-768.

Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological Bulletin*, *127*, 376-407.

Kanungo, R. N. (1982). Measurement of job and work involvement. *Journal of Applied Psychology*, *67*, 341-349.

Katou, A. A., Budhwar, P. S., & Patel, C. (2014). Content vs. process in the HRM-performance relationship: An empirical examination. *Human Resource Management*, *53*, 527–544.

Krackhardt, D., Mckenna, J., Porter, L. W., & Steers, R. M. (1981). Supervisory behavior and employee turnover: A field experiment. *Academy of Management Journal*, *24*, 249–259.

Kehoe, R. R., & Wright, P. M. (2013). The impact of high-performance human resource practices on employees' attitudes and behaviors. *Journal of Management*, *39*, 366–391.

Koys, D. J. (2001). The effects of employee satisfaction, organizational citizenship behavior, and turnover on organizational effectiveness: A unit-level, longitudinal study. *Personnel Psychology*, *54*, 101–114.

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: six practical recommendations. *BMC Psychology*, *4*, 24-34.

Latham, G. P., & Yukl, G. A. (1976). Effects of assigned and participative goal setting on performance and job satisfaction. *Journal of Applied Psychology*, *61*, 166-171.

Lepak, D. P., & Snell, S. A. (2002). Examining the human resource architecture: The relationships among human capital, employment, and human resource configurations. *Journal of Management*, *28*, 517–543.

Liao, H., Toya, K., Lepak, D. P., & Hong, Y. (2009). Do they see eye to eye? Management and employee perspectives of high-performance work systems and influence processes on service quality. *Journal of Applied Psychology*, *94*, 371-391.

Macduffie, J. P. (1995). Human Resource Bundles and Manufacturing Performance: Organizational Logic and Flexible Production Systems in the World Auto Industry. *Industrial & Labor Relations Review*, *48*, 197–221.

Mathieu, J. E. (2016). The problem with [in] management theory. *Journal of Organizational Behavior*, *37*,1132-1141.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Lawrence Erlbaum.

Messersmith, J. G., Patel, P. C., Lepak, D. P., & Gould-Williams, J. S. (2011). Unlocking the black box: exploring the link between high-performance work systems and performance. *Journal of Applied Psychology, 96*, 1105-1118.

Meyer, J. P., Allen, N. J., & Smith, C. A. (1993). Commitment to organizations and occupations: Extension and test of a three-component conceptualization. *Journal of Applied Psychology*, *78*, 538-551.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716-1 – aac4716-8.

Osterman, P. (1994). How common is workplace transformation and who adopts it? *Industrial & Labor Relations Review*, *47*, 173–188.

Paauwe, J. (2009). HRM and performance: Achievements, Methodological Issues and Prospects. *Journal of Management Studies, 46*, 129-142.

Purcell, J., & Kinnie, N. (2007). HRM and business performance. In P. Boxall, J. Purcell, & P. Wright (Eds.), *The Oxford Handbook of Human Resource Management* (pp. 533-551). New York: Oxford University Press.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*, 59–82.

Rousseau, D. (1995). *Psychological contracts in organizations: Understanding written and unwritten agreements*. Thousand Oaks, CA: Sage.

Rubin, D. (1974). Estimating the causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology, 66*, 688-701.

Rubin, D. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322–331.

Ryan, A. M., Schmidt, M. J., & Johnson, R. (1996). Attitudes and effectiveness: Examining relations at an organizational level. *Personnel Psychology*, *49*, 853–882.

Saks, A. M., & Ashforth, B. E. (2002). Is job search related to employment quality? It all depends on the fit. *Journal of Applied Psychology*, *87*, 646-654.

Schmidt, J. A., & Pohler, D. (2016). Accounting for endogeneity in associations between HR systems, leadership, and employee attitudes. *Academy of Management Proceedings*. doi: 10.5465/AMBPP.2016.237

Schmidt, J. A., Pohler, D., & Willness, C. R. (2018). Strategic HR system differentiation between jobs: The effects on firm performance and employee outcomes. *Human Resource Management*, *57*, 65-81.

Schneider, B., Ehrhart, M. G., & Macey, W. H. (2013). Organizational climate and culture. *Annual Review of Psychology*, *64*, 361–388.

Schneider, B., Hanges, P. J., Smith, D. B., & Salvaggio, A. N. (2003). Which comes first: employee attitudes or organizational financial and market performance? *Journal of Applied Psychology*, *88,* 836-851.

Shaw, J. D., Dineen, B. R., Fang, R., & Vellella, R. F. (2009). Employee-organization exchange relationships, HRM practices, and quit rates of good and poor performers. *Academy of Management Journal*, *52*, 1016–1033.

Shin, D., & Konrad, A. M. (2014). Causality between high-performance work systems and organizational performance. *Journal of Management*. Online advance publication. doi:10.1177/0149206314544746.

Shuck, B., Adelson, J. L., & Reio, T. G. (2016). The employee engagement scale: Initial evidence for construct validity and implications for theory and practice. *Human Resource Management*. doi: 10.1002/hrm.21811.

Sikora, D. M., Ferris, G. R., & Van Iddekinge, C. H. (2015). Line manager implementation perceptions as a mediator of relations between high-performance work practices and employee outcomes. *Journal of Applied Psychology*, *100*, 1908-1918.

Subramony, M. (2009). A meta-analytic investigation of the relationship between HRM bundles and firm performance. *Human Resource Management*, *48*, 745–768.

Sutton, R. I., & Staw, B. M. (1995). What theory is not. *Administrative Science Quarterly*, *40*, 371–384.

Takeuchi, R., Chen, G., & Lepak, D. P. (2009). Through the looking glass of a social system: Cross-level effects of high-performance work systems on employees' attitudes. *Personnel Psychology*, *62*, 1-29.

Takeuchi, R., Lepak, D. P., Wang, H., & Takeuchi, K. (2007). An empirical examination of the mechanisms mediating between high-performance work systems and the performance of Japanese organizations. *Journal of Applied Psychology*, *92*, 1069–1083.

Thatcher, S. M. B., & Patel, P. C. (2012). Group faultlines a review, integration, and guide to future research. *Journal of Management*, *38*, 969–1009.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*, 90–118.

Tonidandel, S., & LeBreton, J. M. (2011). Relative importance analysis: A useful supplement to regression analysis. *Journal of Business and Psychology*, *26*, 1–9.

Tonidandel, S., & LeBreton, J. (2015). RWA Web: A Free, Comprehensive, Web-Based, and User-Friendly Tool for Relative Weight Analyses. *Journal of Business and Psychology*, *30*, 207–216.

Wright, P. M., Gardner, T. M., Moynihan, L. M., & Allen, M. R. (2005). The relationship between HR practices and firm performance: Examining causal order. *Personnel Psychology*, *58*, 409–446.

Wu, P. C., & Chaturvedi, S. (2009). The role of procedural justice and power distance in the relationship between high performance work systems and employee attitudes: A multilevel perspective. *Journal of Management*, *35*, 1228-1247.

Wyss, R., Ellis, A. R., Brookhart, M. A., Girman, C. J., Jonsson Funk, M., LoCasale, R., & Stürmer, T. (2014). The role of prediction modeling in propensity score estimation: An evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *American Journal of Epidemiology*, *180*, 645-655.

Zablah, A. R., Carlson, B. D., Donavan, D. T., Maxham III, J. G., & Brown, T. J. (2016). A Cross-Lagged Test of the Association Between Customer Satisfaction and Employee Job Satisfaction in a Relational Context. *Journal of Applied Psychology*, *101*, 743-755.

## Appendix: Supplemental Meta-Analysis

We derived meta-analytic estimates of correlations between constructs in our study using metaBUS, which is an ongoing project to curate the millions of findings in the applied psychology and management literatures (Bosco et al., 2015; Bosco et al., 2017). MetaBUS contains a search engine that provides rapid meta-analytic estimates and links to the original research. We searched the metaBUS taxonomy for constructs that were similar to those in our study and used the following to derive the meta-analytic estimates (metaBUS taxonomic identification number follows in parentheses): tenure in organization (20301); sex/gender with female coded as higher (20545); management level (20314); high performance work systems (20528); leadership behavior (20201); positive job affect (12169); customer service (20163). The meta-analytic correlation matrix is show in Table A1.

**Table A1: Meta-Analytic Correlation Matrix**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1. Organizational Tenure | 1.00 | 414,962 (503) | 350,240 (130) | 29,398 (23) | 110,518 (269) | 306,316 (374) | 12,214 (2) |
| 2. Gender | -.05* | 1.00 | 258,390 (133) | 54,501 (26) | 190,055 (381) | 536,991 (494) | 20,493 (16) |
| 3. Management Position | .13* | -.08* | 1.00 | 3,007 (4) | 100,341 (152) | 132,073 (98) | -- |
| 4. High Performance Work Systems | -.01 | .02 | .10 | 1.00 | 2,043 (14) | 213,469 (53) | 1,081 (9) |
| 5. Leadership Behavior | .02* | .03* | .13* | .36* | 1.00 | 243,681 (517) | 11,081 (66) |
| 6. Positive Job Affect | .08* | .01* | .06* | .12* | .41* | 1.00 | 70,860 (37) |
| 7. Customer Service / Satisfaction | -.17* | .03* | -- | .27* | .07* | .01 | 1.00 |

*Notes.* The sample size weighted meta-analytic correlations appear below the diagonal. The total sample sizes and number of samples ($k$) appear above the diagonal.

$^{*}$ $p < .05$

**Table 1**

**Multilevel Alpha, ICC(1), ICC(2), and $r_{wg}$ at Each Time Period**

| *Construct* | *Time* | *N units* | *Avg N per unit* | *Within ω* | *Between ω* | *Within α* | *Between α* | *ICC(1)* | *ICC(2)* | $r_{wg}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| HPWS | 1 | 214 | 4.68 | .87 | .96 | .86 | .94 | .13** | .42 | .82 |
| | 2 | 250 | 4.00 | .88 | .94 | .88 | .93 | .12** | .35 | .83 |
| | 3 | 266 | 3.84 | .88 | .95 | .88 | .90 | .10** | .30 | .81 |
| | 4 | 272 | 3.82 | .87 | .96 | .87 | .85 | .05* | .17 | .79 |
| | 5 | 279 | 3.76 | .86 | .94 | .86 | .90 | .11** | .32 | .83 |
| | 6 | 271 | 3.79 | .88 | .92 | .88 | .80 | .03 | .10 | .78 |
| | 7 | 258 | 3.83 | .87 | .97 | .87 | .93 | .13** | .36 | .81 |
| | 8 | 253 | 4.09 | .85 | .96 | .85 | .90 | .13** | .38 | .73 |
| Leader Behavior | 1 | 214 | 4.68 | .93 | .98 | .93 | .97 | .11** | .37 | .73 |
| | 2 | 250 | 4.00 | .95 | .98 | .94 | .98 | .10** | .31 | .70 |
| | 3 | 266 | 3.84 | .94 | .98 | .94 | .96 | .10** | .30 | .74 |
| | 4 | 272 | 3.82 | .93 | .96 | .93 | .97 | .11** | .32 | .75 |
| | 5 | 279 | 3.76 | .94 | .95 | .93 | .94 | .08** | .25 | .77 |
| | 6 | 271 | 3.79 | .95 | .97 | .95 | .96 | .06** | .19 | .70 |
| | 7 | 258 | 3.83 | .95 | .98 | .94 | .96 | .08** | .25 | .74 |
| | 8 | 253 | 4.09 | .92 | .97 | .92 | .95 | .11** | .34 | .77 |
| Employee Satisfaction | 1 | 214 | 4.68 | .86 | .94 | .86 | .93 | .08** | .29 | .80 |
| | 2 | 250 | 4.00 | .86 | .96 | .86 | .95 | .10** | .31 | .79 |
| | 3 | 266 | 3.84 | .85 | .48 | .84 | .88 | .06** | .21 | .84 |
| | 4 | 272 | 3.82 | .83 | .97 | .83 | .95 | .10** | .30 | .84 |
| | 5 | 279 | 3.76 | .85 | .78 | .85 | .88 | .06** | .19 | .85 |
| | 6 | 271 | 3.79 | .86 | .91 | .86 | .87 | .05* | .17 | .84 |
| | 7 | 258 | 3.83 | .85 | .96 | .85 | .94 | .07* | .22 | .84 |
| | 8 | 253 | 4.09 | .82 | .93 | .81 | .92 | .09** | .29 | .81 |

Note. *ω* = omega, *α* = alpha. ** $p < .01$, * $p < .05$

**Table 2**

**Descriptive Statistics and Correlations among Study Variables**

| | M | SD | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Time | 4.00 | 1.92 | | | | | | | | | | | | | | | |
| 2. < 1yr service | 0.11 | 0.22 | -.14** | | | | | | | | | | | | | | |
| 3. 1-2 yr service | 0.08 | 0.19 | .03 | -.08** | | | | | | | | | | | | | |
| 4. 3-5 yr service | 0.28 | 0.32 | .01 | -.17** | -.14** | | | | | | | | | | | | |
| 5. 6-10 yr service | 0.22 | 0.31 | .06* | -.19** | -.15** | -.32** | | | | | | | | | | | |
| 6. 11-15 yr service | 0.12 | 0.24 | .02 | -.13** | -.10** | -.22** | -.19** | | | | | | | | | | |
| 7. 16-20 yr service | 0.07 | 0.19 | .00 | -.09** | -.08** | -.18** | -.15** | -.10** | | | | | | | | | |
| 8. 21-25 yr service | 0.05 | 0.18 | .01 | -.11** | -.08** | -.16** | -.15** | -.08** | -.06* | | | | | | | | |
| 9. > 25 yr service | 0.07 | 0.21 | -.05† | -.12** | -.10** | -.19** | -.17** | -.13** | -.07** | -.05† | | | | | | | |
| 10. Gender | 0.42 | 0.40 | -.01 | -.11** | -.03 | -.04 | .00 | -.06* | .00 | .10** | .18** | | | | | | |
| 11. People Manager | 0.31 | 0.45 | -.04 | -.21** | -.14** | -.07** | .12** | .03 | .01 | .08** | .18** | .15** | | | | | |
| 12. HPWS Lag | 4.91 | 0.59 | .04 | -.02 | -.01 | -.10** | .05† | -.04† | .02 | .07** | .10** | .09** | .40** | | | | |
| 13. HPWS | 4.88 | 0.63 | -.02 | -.08** | -.02 | -.04 | .04† | -.04 | .02 | .06* | .10** | .06* | .38** | .46** | | | |
| 14. Leader Behavior Lag | 5.17 | 0.65 | .09** | -.04 | -.05* | -.08** | .05* | -.01 | .05* | .04 | .09** | .06* | .34** | .62** | .35** | | |
| 15. Leader Behavior | 5.19 | 0.67 | .06* | -.07** | -.05† | -.03 | .04† | -.02 | .04† | .04† | .07** | .05† | .30** | .34** | .62** | .42** | |
| 16. Employee Satisfaction Lag | 5.27 | 0.50 | .12** | -.14** | -.07** | -.09** | .08** | .00 | .03 | .11** | .11** | .13** | .45** | .71** | .40** | .62** | .35** |
| 17. Employee Satisfaction | 5.28 | 0.53 | .01 | -.15** | -.05† | -.05 | .07** | -.03 | .01 | .10** | .11** | .12** | .40** | .38** | .74** | .31** | .63** |
| 18. Customer Satisfaction Lag | 3.91 | 0.28 | .35** | -.26** | .17* | -.09 | .13 | .01 | -.13 | .00 | .15† | .09 | -- | .22** | .03 | .24** | .03 |
| 19. Customer Satisfaction | 3.92 | 0.27 | .15† | -.05 | .10 | .01 | .10 | .05 | -.04 | -.02 | -.21** | -.02 | -- | .07 | .18* | .13† | .20** |

**Table 2 (cont.)**

| | 16 | 17 | 18 |
|---|---|---|---|
| 16. Employee Satisfaction Lag | | | |
| 17. Employee Satisfaction | .48** | | |
| 18. Customer Satisfaction Lag | .19* | .04 | |
| 19. Customer Satisfaction | .06 | .12 | .00 |

*Note.* The controls, selection variables, and lagged dependent variables are measured at time (t) and the dependent variables used as outcomes are measured at time (t+1). *N* observations = 163 for correlations involving the customer satisfaction variable, *N* observations = 1,607 for the remainder of the correlations.

$^{**}$ $p < .01$, $^{*}$ $p < .05$, $^{\dagger}$ $p < .10$

**Table 3**

**Directional and Reverse Causal Effects between HPWS, Employee Satisfaction, and Customer Satisfaction**

| | | | | | | Time-Lagged Models ($DV_{t+1}$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Directional Tests* | | | | | | | | | | | |
| | | *metaBUS* | | *Cross-sectional* | | *Without lagged DV control* | | *With lagged DV control* | | *CBPS* | |
| *Dependent Variable* | *Independent Variables* | *β* | *RW* | *β* | *RW* | *β* | *RW* | *β* | *RW* | *β* | *RW* |
| *Employee Satisfaction* | Employee Sat. Lag | -- | -- | -- | -- | -- | -- | .34 (.05)$^{**}$ | .12 | .17 (.10)$^{\dagger}$ | .04 |
| | HPWS | .11 (.01)$^{**}$ | .01 | .64 (.03)$^{**}$ | .41 | .26 (.04)$^{**}$ | .10 | .05 (.04) | .06 | -.08 (.11) | .01 |
| | *N* | 24,305 | | 1,607 | | 1,607 | | 1,607 | | 1,607 | |
| *Customer Satisfaction* | Cust. Satisfaction Lag | -- | -- | -- | -- | -- | -- | -.06 (.16) | .00 | -.06 (.15) | .00 |
| | HPWS | .27 (.01)$^{**}$ | .07 | .18 (.07)$^{**}$ | .04 | .09 (.09) | .01 | .11 (.09) | .01 | .09 (.08) | .01 |
| | *N* | 5,403 | | 163 | | 163 | | 163 | | 163 | |
| *Reverse Causal Tests* | | | | | | | | | | | |
| | | *metaBUS* | | *Cross-sectional* | | *Without lagged DV control* | | *With lagged DV control* | | *CBPS* | |
| *HPWS* | *Independent Variables* | *β* | *RW* | *β* | *RW* | *β* | *RW* | *β* | *RW* | *β* | *RW* |
| | HPWS Lag | -- | -- | -- | -- | -- | -- | .31 (.04)$^{**}$ | .11 | .23 (.08)$^{**}$ | .06 |
| | Employee Satisfaction | .11 (.01)$^{**}$ | .01 | .67 (.03)$^{**}$ | .42 | .29 (.04)$^{**}$ | .11 | .08 (.04)$^{*}$ | .06 | .06 (.07) | .01 |
| | *N* | 24,305 | | 1,607 | | 1,607 | | 1,607 | | 1,607 | |
| *HPWS* | HPWS Lag | -- | -- | -- | -- | -- | -- | .49 (.07)$^{**}$ | .20 | .49 (.07)$^{**}$ | .22 |
| | Customer Satisfaction | .28 (.01)$^{**}$ | .08 | .21 (.09)$^{*}$ | .04 | .14 (.09) | .01 | .04 (.08) | .01 | .14 (.06)$^{*}$ | .02 |
| | *N* | 5,403 | | 163 | | 163 | | 163 | | 163 | |

*Note*. CBPS = the Covariate Balanced Propensity Score weighted models, RW = Relative Weight. Standard errors are in parentheses. *N* for the metaBUS models is the harmonic mean of the sample sizes for each correlation in the meta-analytic correlation matrix (see the Appendix). Tenure with the organization, gender, and management status were included as control variables in the metaBUS models. Time, tenure with the organization, gender, and management status were included as controls in all employee satisfaction models. The customer satisfaction models included the same controls as the employee satisfaction models except for management status. The complete tables are in Online Appendices C through F. $^{**}$ $p < .01$, $^{*}$ $p < .05$, $^{\dagger}$ $p < .10$

**Table 4**

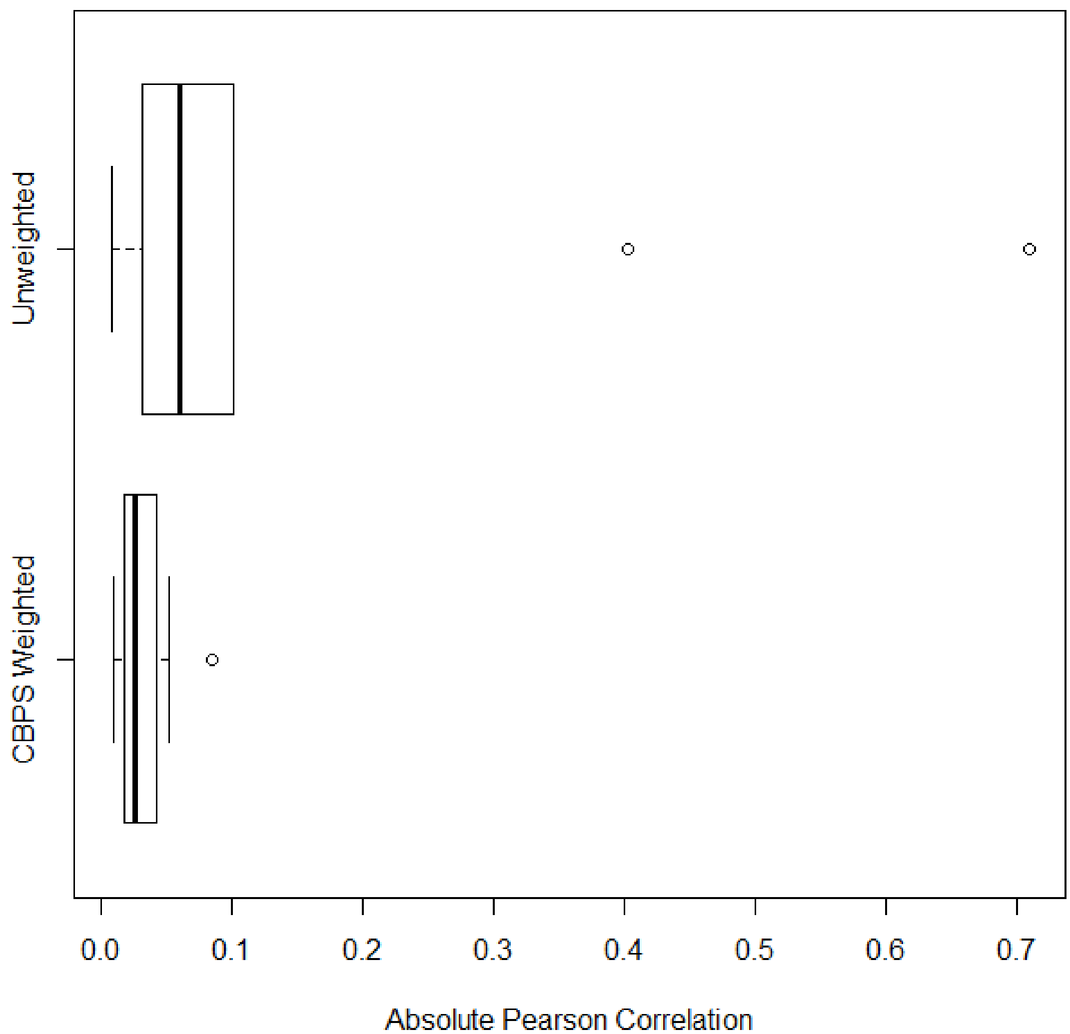
**Omitted Variable Analyses**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Directional Tests* | | | | | | | | | | | |
| | | | | | | *Time-Lagged Models ($DV_{t+1}$)* | | | | | |
| | | *metaBUS* | | *Cross-sectional* | | *Without lagged DV control* | | *With lagged DV control* | | *CBPS* | |
| *Dependent Variable* | *Independent Variables* | *β* | *RW* | *β* | *RW* | *β* | *RW* | *β* | *RW* | *β* | *RW* |
| *Employee Satisfaction* | Employee Sat. Lag | -- | -- | -- | -- | -- | -- | .34 (.05)$^{**}$ | .10 | .21 (.09)$^{*}$ | .04 |
| | Leader Behavior | .42 (.01)$^{**}$ | .16 | .26 (.04)$^{**}$ | .19 | .09 (.04)$^{*}$ | .04 | .00 (.05) | .03 | .22 (.13) | .05 |
| | HPWS | -.04 (.01)$^{**}$ | .01 | .49 (.04)$^{**}$ | .29 | .21 (.04)$^{**}$ | .07 | .05 (.04) | .05 | -.07 (.05) | .01 |
| | *N* | 24,305 | | 1,607 | | 1,607 | | 1,607 | | 1,607 | |
| *Customer Satisfaction* | Cust. Satisfaction Lag | -- | -- | -- | -- | -- | -- | -.07 (.15) | .00 | -.17 (.16) | .01 |
| | Leader Behavior | -.03 (.01)$^{\dagger}$ | .00 | .07 (.09) | .02 | .15 (.12) | .01 | .16 (.11) | .01 | .17 (.11) | .03 |
| | HPWS | .28 (.01)$^{**}$ | .07 | .14 (.10) | .03 | -.01 (.14) | .00 | .00 (.14) | .00 | .10 (.13) | .02 |
| | *N* | 5,403 | | 163 | | 163 | | 163 | | 163 | |
| *Reverse Causal Tests* | | | | | | | | | | | |
| | | *metaBUS* | | *Cross-sectional* | | *Without lagged DV control* | | *With lagged DV control* | | *CBPS* | |
| | *Independent Variables* | *β* | *RW* | *β* | *RW* | *β* | *RW* | *β* | *RW* | *β* | *RW* |
| *HPWS* | HPWS Lag | -- | -- | -- | -- | -- | -- | .29 (.04)$^{**}$ | .05 | .06 (.14) | .00 |
| | Leader Behavior | .37 (.01)$^{**}$ | .12 | .27 (.04)$^{**}$ | .19 | .14 (.03)$^{**}$ | .06 | .07 (.04)$^{\dagger}$ | .04 | .12 (.19) | .01 |
| | Employee Satisfaction | -.04 (.01)$^{**}$ | .01 | .52 (.04)$^{**}$ | .30 | .20 (.04)$^{**}$ | .08 | .06 (.04) | .09 | .22 (.16) | .06 |
| | *N* | 24,305 | | 1,607 | | 1,607 | | 1,607 | | 1,607 | |
| *HPWS* | HPWS Lag | -- | | -- | -- | -- | -- | .53 (.11)$^{**}$ | .16 | .43 (.13)$^{**}$ | .16 |
| | Leader Behavior | .34 (.01)$^{**}$ | .12 | .71 (.04)$^{**}$ | .46 | .31 (.07)$^{**}$ | .07 | -.06 (.11) | .04 | -.02 (.14) | .03 |
| | Customer Satisfaction | .25 (.01)$^{**}$ | .07 | .09 (.07) | .03 | .09 (.09) | .01 | .04 (.08) | .00 | .21 (.10)$^{*}$ | .02 |
| | *N* | 5,403 | | 163 | | 163 | | 163 | | 163 | |

*Note*. CBPS = the Covariate Balanced Propensity Score weighted models, RW = Relative Weight. Standard errors are reported in parentheses. *N* for the metaBUS models is the harmonic mean of the sample sizes for each correlation in the meta-analytic correlation matrix (see the Appendix). Tenure with the organization, gender, and management status were included as control variables in the metaBUS models. Time, tenure with the organization, and gender were included as controls in the remaining models. The complete tables are in Online Appendices G through J. $^{**}$ $p < .01$, $^{*}$ $p < .05$, $^{\dagger}$ $p < .10$

**Figure 1**

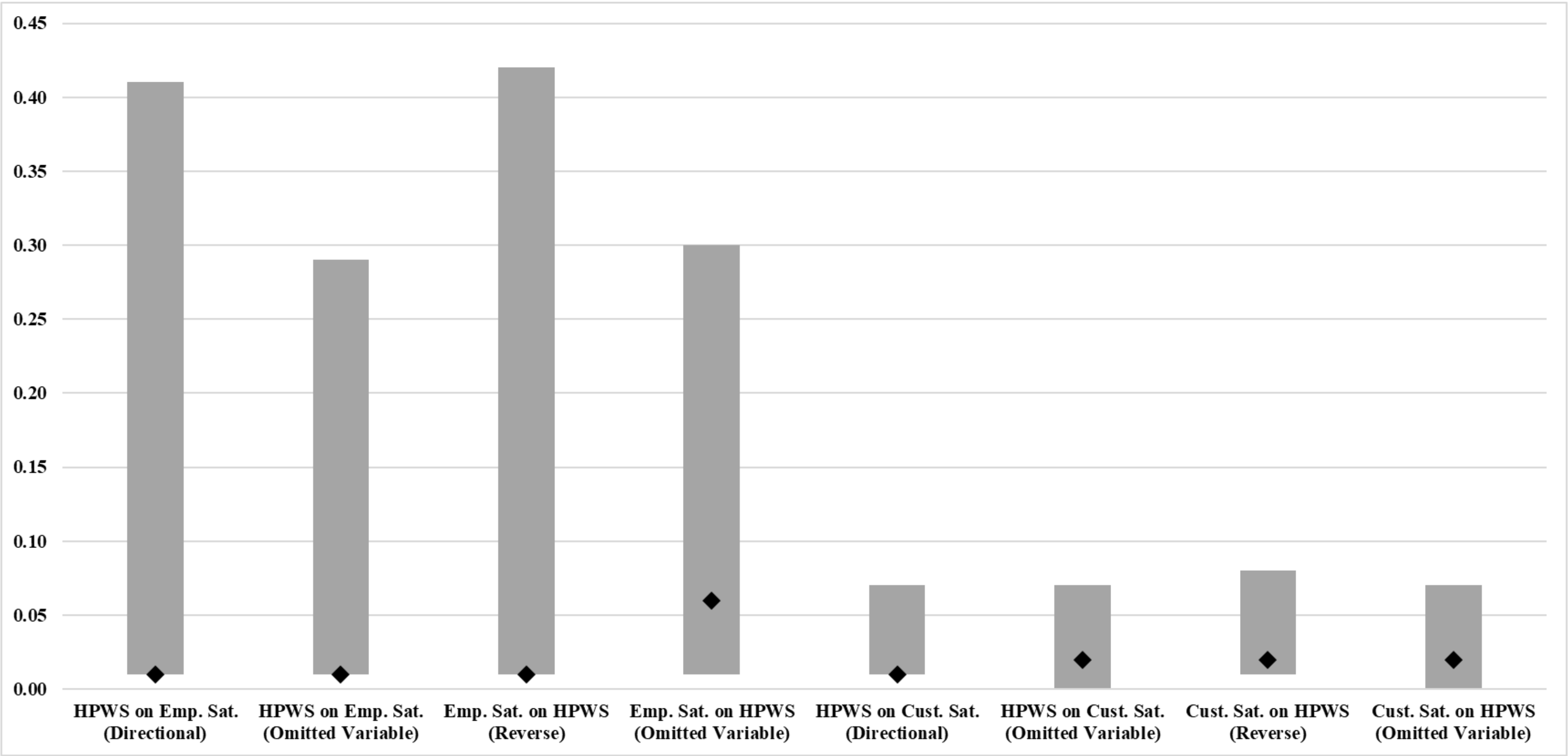
**Absolute Correlations between the Selection Variables and HPWS Before and After the CBPS Weighting Procedure**




*Note.* The open circles are covariate correlations outside percentile ranges represented by the boxplots.

**Figure 2**

**Range of Relative Weights for the Relationships Between HPWS, Employee Satisfaction, and Customer Satisfaction**




*Note*. The black diamonds are the relative weights from the CBPS models and the grey bars are the range of relative weights from the metaBUS, cross-sectional, and time-lagged (non-CBPS) models. Emp. Sat. = Employee Satisfaction and Cust. Sat. = Customer Satisfaction.