The romance of human resource management and business performance, and the case for big science
Toby D. Wall and Stephen J. Wood
*Human Relations* 2005 58: 429
DOI: 10.1177/0018726705055032

The online version of this article can be found at:
http://hum.sagepub.com/content/58/4/429
The romance of human resource management and business performance, and the case for big science

Toby D. Wall and Stephen J. Wood

ABSTRACT

It is often assumed that research over the last decade has established an effect of human resource management (HRM) practices on organizational performance. Our critical assessment of existing studies finds that, although collectively they have opened up a promising line of inquiry, their methodological limitations make such a conclusion premature. We argue that future progress depends on using stronger research methods and design that, in turn, will require large-scale long-term research at a level of magnitude that probably can only be achieved through partnerships between research, practitioner and government communities. We conclude that progress so far justifies investment in such big science.

KEYWORDS

company performance ● high performance work organization ● high performance work systems ● human resource management ● organizational performance

Picture the scene. A leading scholar specializing in human resource management (HRM) is called to court as an expert witness by an international company that has brought a case against a firm of consultants. The company has paid several hundred thousand pounds in consultancy fees, and invested many times that amount in its personnel function, to introduce the ‘performance-enhancing’ HRM practices recommended by the consultants. Three years on there is no evident return on their investment. The company is suing
the consultants on the grounds that they created misleading expectations of
the effect of HRM practices on performance. The expert witness is asked to
prepare a report addressing two questions: (i) whether it was reasonable of
the consultants to assume an effect of HRM practices on performance on the
basis of both researchers’ accounts of their own findings and how practi-
tioners have been encouraged to view them; and (ii) the extent to which the
evidence can, in fact, be interpreted as establishing such a causal link. In this
article we present the report that we would have submitted were we the
expert witness. Having adopted this format, we conclude by considering the
implications of our assessment for future research. First, however, we intro-
duce the core concepts.

Conceptual background

HRM is a term used to represent that part of an organization’s activities
concerned with the recruitment, development and management of its
employees (Wood & Wall, 2002). Within that domain, current interest is
focused on HRM systems emphasizing all or most of the following practices:
sophisticated selection methods, appraisal, training, teamwork, communica-
tions, empowerment, performance-related pay and employment security.
Collectively, these are deemed to contribute to the skill and knowledge base
within the organization, and to employees’ willingness to deploy their
learning to the benefit of the organization. For this reason, authors have used
labels such as ‘human capital-enhancing’ (Youndt et al., 1996), ‘high
commitment’ (Wood & de Menezes, 1998) or ‘high involvement’ (Guthrie,
2001; Vandenberg et al., 1999) to characterize the approach. They often
contrast this with traditional, Tayloristic or control approaches to manage-
ment, which emphasize low skill, limited employee discretion and a tight
division of labour (e.g. Arthur, 1994; Lawler, 1986; Walton, 1985).

The rationale for focusing on the high-commitment approach to
HRM lies in the assumption that the practices will enhance organizational
performance. Related to this is the view that the practices will be a source
of unique and sustainable advantage because, unlike other initiatives such as
the introduction of new technology, they result in skills and knowledge that
are largely organization specific and are therefore difficult to imitate
(e.g. Barney, 1991, 1995). Companies can copy one another’s technologies
much more easily than they can their human resource capabilities.

The link between HRM and performance has been conceived in a
variety of ways. The simplest view is that practices are additive (the more
the better, e.g. Guest & Hoque, 1994), and that they enhance performance
regardless of circumstance (a universal effect, e.g. Pfeffer, 1994). Alternative
perspectives emphasize various kinds of ‘fit’, of which three main types have been identified (Wood, 1999). ‘Internal fit’ posits synergy among the practices, meaning that their collective effect will be greater than the sum of their individual parts. For example, it might be argued that selecting able people without training them, or training employees but not empowering them to use that training, will have little effect; whereas implementing the three practices together will. Indeed, Barney (1995) goes so far as to argue that individual practices ‘have limited ability to generate competitive advantage in isolation’ (p. 56).

The second type is ‘organizational fit’, which concerns the role of HRM in enhancing the effectiveness of other organizational practices or technologies, and vice versa. Lawler et al. (1995), for example, link HRM to total quality management (TQM), arguing that the two sets of practices are ‘complementary in their impact on organizational performance’ (p. 144). Similarly, MacDuffie (1995) sees high commitment HRM practices as integral to the effectiveness of lean production (Womack et al., 1990) initiatives and vice versa.

The third and final type is ‘strategic fit’. This assumes that HRM practices need to be aligned with the organization’s strategy (e.g. in the case of private businesses to their competitive strategy) to have their full effect on performance (e.g. Miles & Snow, 1984; Schuler & Jackson, 1987; Youndt et al., 1996).

Empirical investigation of fit typically involves testing for interaction effects (e.g. among different practices, or between HRM system measures and strategy). Because fit is central to many accounts of the effects of HRM on performance, it is relevant to determine the extent to which such interactions have been examined. In practice, as we shall see, much empirical work ignores or pays only limited attention to issues of fit, consciously or by default taking the universal thesis propounded by Pfeffer (1994) and others.

Is an HRM effect on performance presumed by researchers and practitioners?

We can now return to the first question we have been asked to consider as expert witnesses, whether authors portray research findings as confirming a causal effect of HRM on organizational performance, in support of the guiding theory. There can be little doubt that such an interpretation is widespread. Early reviews of the field fuelled this perception as they concluded that studies supported the universal thesis (e.g. Becker & Huselid, 1998; Guest, 1997). Becker and Gerhart (1996: 779), for example, argue that ‘conceptual and empirical work . . . has progressed enough to suggest that
the role of human resources can be crucial’. Such views have permeated subsequent studies. Way (2002: 765), for instance, states that ‘Theoretical and empirical HRM research has led to a general consensus that the method used by a firm to manage its workforce can have a positive impact on firm performance’.

A more subtle way in which authors imply an established performance effect lies in the labels many use to denote salient HRM systems. For example, in his seminal study, Huselid (1995) uses the expression ‘high performance work practices’ (p. 635), mirroring Lawler et al.’s (1995, 1998) choice of ‘high performance organizations’ for their book titles. Subsequently, Delaney and Huselid (1996) refer to ‘performance-enhancing’ (p. 949) HRM practices, Kalleberg and Moody (1996) to ‘high performance work organizations’ (p. 114), and Appelbaum et al. (2000) and Way (2002: 765) to ‘high performance work systems’. The titles of many articles carry the same message. For example, an early article by Arthur (1994) is entitled ‘Effects of human resource systems on manufacturing performance and turnover’. Following Huselid’s (1995) lead in his article on ‘The impact of human resource management practices on turnover, productivity and corporate financial performance’ (our italics), many other authors, also describing cross-sectional results, use equivalent causal terminology (e.g. Ahmad & Schroeder, 2003; Bae & Lawler, 2000; Delaney & Huselid, 1996; Jayaram et al., 1999; Vandenberg et al., 1999; Wright et al., 2003). The idea of causality is thus embedded within the choice of label for HRM systems and in the wording of titles. Such terminology is in danger of appearing to prejudge the very relationship under investigation, as if the researcher’s role were to find evidence for a widely expected (and one might say, even hoped for) relationship.

A further way in which the discourse of research implies an established effect of HRM on performance is through the argument that it is now time to move on to other issues. In this vein, Delery (1998) suggests that ‘Establishing that HRM practices are linked with firm effectiveness is an important first step . . . however, there is little understanding of the mechanisms through which HRM practices influence effectiveness’ (p. 289). Similarly, there have been calls to broaden the outcomes studied to cover those of more immediate interest to employees. Osterman’s (2000) study of ‘High performance work organization’ practices follows this lead, as he set his objective to determine ‘whether their productivity and quality gains redounded to employees’ benefit’ (p. 179). He examines the not unimportant question of whether these practices are associated with layoff rates and compensation gains; but in so doing he simply assumes, without proof, that there are gains to redound (i.e. share among employees).
A final indication of the way in which a causal effect is assumed relates to practice. In their review of research, Becker and Gerhart (1996) set their goal as being to ‘demonstrate to senior human resources (HR) and line managers that their HR systems represent a largely untapped opportunity to improve firm performance’ (p. 780). That message has clearly got through, as evident from Cooper’s (2000) article in a leading UK HRM practitioner journal, which concludes: ‘Academic studies have established a link between HR and profitability. The race is now on to find out how and why people policies make a difference to shareholders’ (p. 28). Again, Philpott (2002) concludes that:

What is missing, says the CIPD [the UK’s Chartered Institute of Personnel Development], is a genuine appreciation both that people management holds the key to increased productivity and that meeting the objective requires the appropriate application of a range of people management practices. In general terms these practices encompass, recruitment, training, job appraisal and reward, job design, job quality and communication with staff.

(p. 5)

One should record that some have been more circumspect in their assessments (e.g. Delaney & Godard, 2001; Marchington & Grugulis, 2000; Wood, 1999). Wright and Gardner (2003), for example, go only so far as to suggest that evidence mounts that HR practices are at least ‘weakly related to firm performance’ (p. 312). Godard (2004) is even more sceptical. Though not doubting that HRM ‘can be highly effective in some work places’ (p. 355), he concludes that, especially in liberal market economies, the generalizability is likely to be low. Nevertheless, the message conveyed by a large part of the literature is that HRM practices do promote performance.

Expert witness testimony: part 1

Our evidence on the first question is clearly in favour of the defence. It shows that it was reasonable for the consultants to assume a positive effect of HRM practices on performance on the basis both of researchers’ portrayal of their own findings and how these have been presented in non-academic circles.

Does the evidence support an HRM–performance effect?

To address the second question, concerning the extent to which the evidence can be interpreted as establishing a causal effect of HRM on performance,
we evaluate key empirical studies. What follows is an account of the method we adopted, and our findings.

Sampling the studies

We used five criteria to select the studies to be evaluated. First, we restricted our choice to those appearing in reputable refereed journals, to ensure that we were assessing high-quality studies. This meant that other influential work, published, for example, in books or reports, was excluded (e.g. Appelbaum et al., 2000; Ichniowski, 1990; Lawler et al., 1995, 1998). Second, we restricted our selection to studies published from 1994 onwards, when research on HRM and performance first came to prominence. Third, we included only those studies covering multiple HRM practices, because the focus is on whether the HRM system as a whole promotes overall organizational performance. This excluded any studies of single practices such as training, sophisticated selection, or job design. Fourth, we included only studies using measures of economic performance such as productivity, profit or return on assets. Thus we excluded studies limited to other performance indicators such as labour turnover, absence or employee well-being. Finally, because of the detailed assessment involved, we restricted the sample to highly cited milestone studies of the mid- to late-1990s, and a selection of more recent ones (whose citation rate is yet to be determined). The studies we used are shown in Table 1.

Assessing the studies

In order to assess the evidence, we identified key criteria relevant to generalization and causal interpretation against which to judge studies individually and collectively. Those criteria concern: the sample and response rate; the reliability and validity of the HRM measure and source of data for it; the adequacy of the research design; the extent to which other factors have been controlled; the strength of the findings on the HRM–performance link; whether there has been a test for fit; and finally if the effects of individual HRM practices have been considered alongside those of the composite HRM measure. We expand on these criteria and the rationale for their choice in the course of presenting findings.

Sample and response rate

The size of samples used affects the power to detect main effects, and even more so theoretically predicted synergies (interactions). Coupled with the
response rate, size also affects the extent to which one can generalize. It can be seen (Table 1, column 2) that the studies have diverse but often rather small (< 300 in 18 of the 25 studies, and < 100 in 9 studies) samples (column 2). At the same time, the response rates, though varying between 4 and 84 percent, are generally low – albeit better than the 6–28 percent (mean 17.4%) rate reported in an earlier review by Becker and Huselid (1998). This is not a very secure foundation from which to generalize.

**HRM dimensions**

Evidence that the studies use appropriate and broadly equivalent measures of HRM is important to justify treating them collectively. As previous commentators have noted (e.g. Becker & Gerhart, 1996; Dyer & Reeves, 1995; Wright & Gardner, 2003) there is diversity across studies in the particular practices covered. Nonetheless, there is much commonality as studies typically cover a substantial range of the following: sophisticated selection, appraisal, training, teamwork, communication, job design, empowerment, participation, performance-related pay/promotion, harmonization, and employment security (Table 1, column 3).

**HRM measure type and source**

In any empirical investigation reliable measurement of independent variables (in this case HRM practices) is of paramount importance; and in cross-sectional research, ensuring such measures are uncontaminated by those of the dependent variables (i.e. performance) is of particular significance. In studies of HRM, the measurement of relevant practices cannot be a straightforward process, because the focal construct is not of that nature. Whether there is a bonus scheme in operation may be easy to establish, but judgments of the extent of teamwork or empowerment are much more complicated to gauge. Thus, quantifying HRM practices involves judgement. Consequently, it is open to random measurement error. At the same time, it is vulnerable to correlated error in the form of rater bias. For example, more optimistic or organizationally committed individuals rating both the practices and performance in their own company might give systematically higher scores than their more pessimistic counterparts elsewhere. Such ‘common method variance’ alone could create spurious relationships among HRM practices and between them and performance.

Given these considerations, the measurement of HRM practices would ideally involve: assessments from two or more persons (to determine reliability), the use of the same raters across different organizations (to
Table 1 Measurement features of empirical studies on HRM and organizational performance

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample and response rate (%)</th>
<th>HRM dimensions and illustrative components</th>
<th>HRM measure type and source</th>
<th>Dependent variable measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Arthur (1994)</td>
<td>30 US mini steel mills; 56 percent&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Two contrasting types of HRM system: control and commitment (low and high respectively on decentralization, participation, training and skill)</td>
<td>Questionnaire; single-source; HR managers</td>
<td>Self-report: productivity and scrap rate; same respondent as for independent variables; for previous year</td>
</tr>
<tr>
<td>2 Guest and Hoque (1994)</td>
<td>119 UK greenfield mainly manufacturing companies; 39 percent&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Four a priori types: 2 × 2 – whether or not claim HRM strategy and use more or less than half of 22 HRM practices (e.g. selection, job design)</td>
<td>Questionnaire; single-source; principal HR manager or senior line manager</td>
<td>Self-report: productivity and quality; same respondent as for independent variables; concurrent</td>
</tr>
<tr>
<td>3 Huselid (1995)</td>
<td>968 US companies with 100+ staff; heterogenous sample; 28 percent</td>
<td>Two scales: skills and structures (e.g. communications, QWL, training, grievance procedures); and motivation (e.g. performance appraisals, promotion on merit)</td>
<td>Questionnaire; single-source; mailed to senior HRM professional</td>
<td>Objective: productivity, Tobin’s Q and GRATE for subsequent year</td>
</tr>
<tr>
<td>4 MacDuffie (1995)</td>
<td>62 car-assembly plants, worldwide; 69 percent&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Two scales: work systems (participation, teams, quality role) and HRM policies (selection, performance-related pay, training). Also a use of buffers (lean production) scale</td>
<td>Questionnaire; ‘a contact person, often the plant manager’; sections completed by different people; later visits/telephone checks</td>
<td>Objective: productivity (labour hours per vehicle) – adjusted for vehicle size and number of welds; quality from independent market report; concurrent</td>
</tr>
<tr>
<td>5 Delaney and Huselid (1996)</td>
<td>390 for-profit and non-profit US firms, heterogenous; 51 percent</td>
<td>Five scales: staffing selectivity, training, incentive pay, decentralization, internal promotion</td>
<td>Telephone survey; single-source (multiple respondents in a few cases); unspecified ‘representatives’</td>
<td>Self-report: organizational and market performance; same respondent as for independent variables; concurrent</td>
</tr>
<tr>
<td>Study</td>
<td>Sample Size</td>
<td>Measures</td>
<td>Sample</td>
<td>Data Collection</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>-------------</td>
<td>---------------------------------------------------------------------------</td>
<td>--------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Delery and Doty (1996)</td>
<td>216 US banks (down to 101 in some analyses); 18 percent</td>
<td>Seven scales: internal promotion, training, appraisal, profit-sharing, security, participation, job specification; and two strategy measures</td>
<td></td>
<td>Questionnaire; single-source; senior HR manager (+ strategy from president)</td>
</tr>
<tr>
<td>Koch and McGrath (1996)</td>
<td>319 US business units; 4 percent</td>
<td>Separate HR planning index, Hiring index and Development index; and composite HR sophistication index (previous three added together)</td>
<td></td>
<td>Questionnaire; single-source; sent to executives</td>
</tr>
<tr>
<td>Youndt et al. (1996)</td>
<td>97 US manufacturing plants (metal-working industry); 19 percent</td>
<td>Two scales: administrative HR (e.g. appraisal, incentives) and human capital-enhancing HR (e.g. selection and training for problem-solving, salaried pay)</td>
<td></td>
<td>Questionnaire; multiple source (at least two respondents per plant, mean score); general and functional managers</td>
</tr>
<tr>
<td>Huselid et al. (1997)</td>
<td>293 US firms: heterogenous (e.g. manufacturing, finance, misc.); response rate unclear</td>
<td>Two scales: strategic HRM (teamwork, empowerment) and technical HRM (recruitment, training). Ratings are not of reported use, but of perceived effectiveness</td>
<td></td>
<td>Questionnaire; single-source; executives in HR (92%) and line (8%) positions (effectiveness of HR practices, not use per se)</td>
</tr>
<tr>
<td>Ichniowski et al. (1997)</td>
<td>36 steel production lines in 17 companies; 60 percent</td>
<td>Four types of line ranging from innovative HRM (high on incentive pay, selection, teamwork, employment security, training) to traditional HRM (low on all above components)</td>
<td></td>
<td>Interviews; multiple respondents; labour relations and operations managers, line workers and trade union reps</td>
</tr>
<tr>
<td>Wood and de Menezes (1998)</td>
<td>Representative sample of 1693 UK workplaces (806–926 in analyses); 84 percent</td>
<td>Four types of workplace, ranging from high to low HCM (high commitment management) – (high HCM showing more appraisal, information disclosure, communication, monthly pay)</td>
<td></td>
<td>Interviews; single-source; HR manager or senior manager responsible for HR</td>
</tr>
<tr>
<td>Study</td>
<td>Sample and response rate (%)</td>
<td>HRM dimensions and illustrative components</td>
<td>HRM measure type and source</td>
<td>Dependent variable measures</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------</td>
<td>------------------------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>12 Hoque (1999)</td>
<td>209 UK hotels; 35 percent</td>
<td>Overall HRM (21 practices used, including harmonization, job design, training, merit pay)</td>
<td>Questionnaire; single-source; respondents unclear</td>
<td>Self-report: productivity, service quality and financial performance; same respondent as for independent variables; concurrent</td>
</tr>
<tr>
<td>13 Jayaram et al. (1999)</td>
<td>57 firms supplying automotive industry; 39 percent</td>
<td>Five scales: cost, quality, flexibility and time HRM (commitment to and training, goals and support for), plus a generic scale (broad jobs, autonomy)</td>
<td>Questionnaire; single-source; CEOs</td>
<td>Self-report: cost, quality, flexibility and time performance; same respondent as for independent variables; concurrent</td>
</tr>
<tr>
<td>14 Vandenberg et al. (1999)</td>
<td>49 North American life insurance companies; response rate unclear</td>
<td>Five scales: flexibility, incentive practices, direction sharing, training, work design</td>
<td>Questionnaire; single-source; head of HR</td>
<td>Objective: return on equity (ROE); for subsequent year</td>
</tr>
<tr>
<td>15 Wright et al. (1999)</td>
<td>38 US petrochemical refineries; 20 percent</td>
<td>Four scales: selection, training, pay and appraisal; with participation as a contingency variable</td>
<td>Questionnaire; single-source; HR manager</td>
<td>Self-report: financial performance; independent source (refinery or operational manager); concurrent</td>
</tr>
<tr>
<td>16 Bae and Lawler (2000)</td>
<td>138 firms in Korea; response rate unclear</td>
<td>Single ‘high involvement HRM strategy’ index (five sub-scales covering training, empowerment, selection, performance-related pay, broad job design)</td>
<td>Questionnaire; single-source; head of HR</td>
<td>Self-report global measure; same respondent as for independent variables; concurrent. Objective (ROIC) for sub-sample (n = 68) – concurrent and subsequent year</td>
</tr>
</tbody>
</table>
17 Fey et al. (2000) 101 foreign-owned Eight single-item measures for managers and Questionnaire; single-source; Self-report; overall performance firms in Russia; non-managers separately: promotion on 38 percent HRM manager, (poor to outstanding); same heterogenous sample; merit; job security; technical training; 63 percent respondent as for independent variables. Data collected by US Bureau of the Census.

18 Cappelli and Neumark (2001) US manufacturing 164 New Zealand Single high-involvement work practices Questionnaire; single-source; Self-report: productivity (annual plants; response rates meetings, and TQM (as index of involvement) respondents as for independent variables. Data collected by US Bureau of the Census.

19 Guthrie and Johnson (2001) 13–15 percent Single high-involvement work practices Questionnaire; single-source; Self-report: productivity (annual plants; response rates meetings, and TQM (as index of involvement) respondents as for independent variables. Data collected by US Bureau of the Census.

20 Richard and Batt (2002) 260 call centres; HR director and productivity (net income per HR director and productivity (net income per

21 Batt (2002) 260 call centres; HR director and productivity (net income per

22 Wall (2002) 260 calls centres; HR director and productivity (net income per

23 Wall (2002) 260 calls centres; HR director and productivity (net income per
Table 1 continued

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample and response rate (%)</th>
<th>HRM dimensions and illustrative components</th>
<th>HRM measure type and source</th>
<th>Dependent variable measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>22 Way (2002)</td>
<td>446 small firms; 70 percent</td>
<td>Single index of ‘High Performance Work Practices’ covering seven practices (e.g. self-directed teams, job rotation, performance pay, training, involvement)</td>
<td>Telephone interview; single-source; plant manager. Data collected by US Bureau of the Census</td>
<td>Self-report: labour productivity (income/labour costs) plus perceived relative (to competitors) productivity; concurrent</td>
</tr>
<tr>
<td>23 Ahmad and Schroeder (2003)</td>
<td>107 manufacturing plants in USA, Germany, Italy and Japan; 60 percent</td>
<td>12 practices including: selection, teamwork, performance-related pay, training, employment security, harmonization</td>
<td>Self-report questionnaires; respondents differ between practices, but single-source for any given measure; managers, engineers, supervisors and workers</td>
<td>Self-report: single index covering – cost, quality, delivery, flexibility, new product introduction; different respondent to that for the independent variables; concurrent</td>
</tr>
<tr>
<td>24 Guest et al. (2003)</td>
<td>366 manufacturing and service sector UK companies; response rate unclear</td>
<td>Single overall use of HRM practices (48 items on: recruitment and selection, training and development, appraisal, financial flexibility, job design, two-way communication, employment security and internal labour market, harmonization, quality)</td>
<td>Telephone survey; single-source; HR director or most senior HR person</td>
<td>Objective: productivity (sales per employee) and profit for years before and after the period when the independent variables were measured</td>
</tr>
<tr>
<td>25 Wright et al. (2003)</td>
<td>50 business units, US food services company; response rate unclear</td>
<td>Single overall HR practices scale: nine items covering selection, pay for performance, training, participation</td>
<td>Employee attitude survey; multiple source; rank-and-file employees</td>
<td>Objective (from company records), productivity and profit – subsequent, for period 3–9 months after measurement of HR practices</td>
</tr>
</tbody>
</table>

* Estimated by current authors as the number of companies in relevant analyses as a percentage of total available sample.
* It is a requirement to ensure the accuracy of all data supplied to the US Bureau of the Census.
* See also Guthrie et al's (2002) article on the same data.
reduce error resulting from differences in the use of rating scales), raters who through knowledge of the use of practices in a range of organizations are able to benchmark the development of the practice against that elsewhere (to better calibrate the scale), and raters who are ignorant of the performance or the organization in question (to eliminate the possibility that knowledge of the outcome will contaminate measurement of the independent variables, resulting in common method variance bias). In contrast, the least satisfactory method is to use reports from single respondents, each describing their own organization, and who have knowledge of, or worse still, are the source of, the performance data.

The profile of the large majority of studies corresponds to the least satisfactory measurement approach. Twenty-one of the 25 studies use single respondents (see Table 1, column 4). Moreover, in only two of the four cases with multiple respondents (Wright et al., 2003; Youndt et al., 1996) is any information on inter-rater agreement provided. In the other two cases (Ichniowski et al., 1997; MacDuffie, 1995), it is unclear how the information was combined, and the extent to which it was consistent across sources. This is not a trivial issue, given findings by Gerhart et al. (2000) and Wright et al. (2001) of very low levels of reliability in HR measures.

It is also the case that 24 of the 25 studies rely exclusively on respondents describing only their own context. The one exception is Ichniowski et al. (1997) who characterized the HRM system across the 36 US steel production lines in their sample themselves, on the basis of interview data from multiple respondents (though quite how is unclear). Moreover, in those 24 studies it is reasonable to assume that those rating the extent of use of the HRM practices were also aware of their organization’s performance. In fact, this is certain in 12 cases in which data on HRM practices and performance were obtained from the same respondent.

Thus, across the studies as a whole, measures of HRM are generally of unknown reliability, are likely to contain considerable random measurement error, and at the same time in the majority of cases are open to contamination from knowledge of performance and demand characteristics. Random error will weaken the assessment of any underlying relationship between HRM and performance; whereas contamination of the data will create correlated error that could result in spurious associations.

*Dependent variable measurement*

Measures of the dependent variable (performance) minimally should come from a different source from that used to measure HRM practices, and ideally would be ‘objective’ – to reduce the likelihood of common method
variance. In many cases, such as in for-profit businesses, measures of productivity and profit taken from official accounts will be appropriate, perhaps supported by more industry-specific indicators (e.g. defects per 1000 cars in auto assembly, as in MacDuffie, 1995). In other instances different measures will be appropriate, for example in charitable organizations the percentage income redistributed will be relevant. So diversity of performance measures consistent with the nature of the business is to be expected. Of course such measures are not necessarily totally accurate (e.g. organizations can bring forward or postpone costs for a given financial year), but they are generally subject to external audit and, most importantly, are collected independently from the measurement of the HRM practices. This contrasts with the least desirable approach that relies on reported performance from a single employee who is also the source of information about the HRM practices.

The picture is more encouraging in this respect than for the measurement of HRM, but is still far from ideal (Table 1, column 5). Ten of the 25 studies do use objective data, ensuring measures of performance come from a different source than those of HRM practices. The other side of the coin, of course, is that 15 of the studies rely on reported performance, in 12 of which the information comes from the same person who assesses the HR practices. It should be noted, however, that recent findings (Wall et al., 2004), suggest that such common source self-report performance data may not necessarily be as biased as one might expect.

Research design

There are two main types of HRM-performance study. Cross-sectional studies are those in which both the independent and dependent variables are measured on one occasion only (be these concurrent or not); whereas in longitudinal studies either or both the independent and dependent variables are measured on at least two occasions. Both types of study are of value. Cross-sectional research is a cost-effective starting point for establishing that two or more variables are related, and the absence of a cross-sectional relationship would send warning signals that more costly longitudinal work might not be justified. Also, cross-sectional work often allows the use of much larger samples and hence augments generalizability. However, for reasons that are well known, cross-sectional studies provide a weak foundation for causal inference, for which longitudinal research designs are to be preferred.

Within longitudinal studies, however, there are alternatives. The simplest advance on a cross-sectional design is what we shall call the
‘quasi-longitudinal’ study. Studies of this kind examine how the extent of use of HRM practices at one point in time predicts subsequent performance, while controlling for prior (or concurrent) performance. The advantage of this over a simple cross-sectional study (e.g. concurrent performance only) is that it can help control for reverse causality, that prior performance fostered greater use of the HRM practices in question. This design, however, has some problems, especially when the timing of the introduction of the HRM practices is unknown. If those practices had been introduced, say, 5 years before the study, and had led to better performance over the 3 years prior to the study, then controlling for that prior performance would effectively be removing the effect of interest. Consequently, whereas quasi-longitudinal research that shows a positive relationship with performance adds to our understanding, if it fails to do so the interpretation is unclear.

A stronger option is what we shall call the ‘authentic longitudinal’ study. Research of this kind would involve measurement of both the independent and dependent variables on two or preferably more occasions. For example, at Time 1 one would measure both HRM practices and performance, at Time 2 the HRM practices again, and at Time 3 performance for a second time. The timing of the measurements would be determined according to a hypothesized lag for the effect of HRM on performance. Evidence consistent with a causal interpretation would require change in HRM to be associated with a subsequent change in performance. In this way, one is testing whether an increase (or decrease) in the use of the relevant practice(s) is associated with a subsequent improvement (or deterioration) in performance; and stable third factors (e.g. sector, product mix) are controlled for through the repeated measures design. The design would be further improved by focusing only on those organizations in which the HRM practices have been introduced (or substantially enhanced) between the performance measurement occasions; or by intervention studies (with controls).

Judged against the above criteria, the design of existing studies is disappointing. As shown in Table 2, 21 of the 25 studies are purely cross-sectional (Table 2, column 2), providing very weak grounds for causal inference. Moreover, 17 of these 21 cross-sectional studies are what Way (2002) describes as ‘temporally backward predictive’ (p. 779), that is to say they involve a relationship between current HRM and prior performance. This is self-evident where authors explicitly report having used previous performance data (e.g. Arthur, 1994; Batt, 2002). Moreover, even in those studies that use concurrent performance (e.g. Delery & Doty, 1996; Guthrie, 2001 – or the average of previous and concurrent performance data as in the case of Youndt et al., 1996), the grounds for inferring causality, were it to be linked to HRM, would be weak. This is because most concurrent measures,
Table 2  Research design and findings of empirical studies on HRM and organizational performance

<table>
<thead>
<tr>
<th>Study</th>
<th>Research design</th>
<th>Control for third factors</th>
<th>Main effects (summary, statistical significance, hit rate)</th>
<th>Interaction effects</th>
<th>Individual practice effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Arthur (1994)</td>
<td>Cross-sectional</td>
<td>Commitment systems show higher reported productivity ( (p &lt; .05) ) and lower scrap rates ( (p &lt; .05) ). 2/2 predicted effects</td>
<td>Not examined</td>
<td>Not examined</td>
</tr>
<tr>
<td>2</td>
<td>Guest and Hoque (1994)</td>
<td>Cross-sectional</td>
<td>No differences across four types. 0/6 predicted effects</td>
<td>Not examined</td>
<td>Not examined</td>
</tr>
<tr>
<td>3</td>
<td>Huselid (1995)</td>
<td>Cross-sectional</td>
<td>For example size, capital, R&amp;D, intensity, unionization, sector</td>
<td>Internal fit for skills and structures ( \times ) motivation for ( \text{GRATE}; \text{strategic fit for motivation} \times \text{HR strategy} - 2/14</td>
<td>Not examined</td>
</tr>
<tr>
<td>4</td>
<td>MacDuffie (1995)</td>
<td>Cross-sectional</td>
<td>Degree of automation, volume, complexity, product design, age</td>
<td>Work systems related to productivity ( (p &lt; .05) ); and work systems and HRM policies to quality ( (p &lt; .05 ) and ( p &lt; .01 ); respectively). 3/4 predicted effects</td>
<td>Internal/organizational fit ( - ) work systems ( \times ) HRM policies ( \times ) buffers for productivity; work systems ( \times ) buffers, for quality ( - 2/8 )</td>
</tr>
<tr>
<td>5</td>
<td>Delaney and Huselid (1996)</td>
<td>Cross-sectional</td>
<td>Many, e.g. size, age, product vs. service, unionization, sector</td>
<td>HRM practices collectively (block in regression) associated with perceived organizational and market performance ( (p \text{ values unspecified}) ). 4/10 (i.e. five HR practices ( \times ) two performance measures) predicted effects</td>
<td>Selected internal fit interactions examined ( (e.g. \text{staffing} \times \text{training}) ) but none found ( - 0/8 )</td>
</tr>
<tr>
<td>Study</td>
<td>Design</td>
<td>Measures</td>
<td>Findings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------</td>
<td>-----------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delery and Doty</td>
<td>Cross-sectional</td>
<td>Size (assets), age, district</td>
<td>Appraisals (p &lt; .05), profit sharing (p &lt; .01) and security (p &lt; .01) related to ROA; only profit-sharing with ROE (p &lt; .01). 4/14 predicted effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Koch and McGrath</td>
<td>Cross-sectional</td>
<td>Size, sector, unionization, R&amp;D, capital intensity</td>
<td>HR planning and hiring related to productivity (p &lt; .05 and p &lt; .01 respectively); development and sophistication not. 2/4 predicted effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Youndt et al.</td>
<td>Cross-sectional</td>
<td>Size, growth, change in sales, product complexity, strategy (cost, quality, delivery, scope)</td>
<td>Administrative HR no effects; human capital-enhancing HR related to reported productivity (p &lt; .01), but not to customer alignment or machine efficiency. 1/6 predicted effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huselid et al.</td>
<td>Quasi-longitudinal</td>
<td>Unionization, size, R&amp;D, sector, concurrent performance and others</td>
<td>Strategic HRM related to subsequent GRATE (controlling for concurrent, p &lt; .05) but not productivity or Tobin’s Q; technical HRM unrelated to any performance measure. 1/6 predicted effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ichniowski et al.</td>
<td>Cross-sectional and longitudinal</td>
<td>Up to 25 for technical reliability (e.g. age and speed of line)</td>
<td>Lines with innovative HRM systems have greater productivity (p &lt; .01); lines becoming more innovative in HRM over time showed increased productivity (p &lt; .01). 1/1 predicted results</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Strategic fit, individual HR practices with product-market innovation. One found (appraisal × innovation) – 1/14
HR sophistication × capital intensity, showing the effect of HR is stronger for higher capital intensity – 1/1
Single-instance of strategic fit – administrative HR × quality strategy at p < .05 – 1/8
Not examined
Not examined
All components (e.g. selection, training) have individual effects, none additional to system
Table 2 continued

<table>
<thead>
<tr>
<th>Study</th>
<th>Research design</th>
<th>Control for third factors</th>
<th>Main effects (summary, statistical significance, hit rate)</th>
<th>Interaction effects</th>
<th>Individual practice effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 Wood and de Menezes (1998)</td>
<td>Cross-sectional</td>
<td>Many, e.g. size sector, unionization</td>
<td>No difference across four types of workplace on productivity or change in productivity; some effect for financial performance but not as expected. 0/3 predicted effects</td>
<td>Not examined</td>
<td>Not examined</td>
</tr>
<tr>
<td>12 Hoque (1999)</td>
<td>Cross-sectional</td>
<td>Size, unionization, country of ownership, strategy</td>
<td>Overall HRM related to productivity, quality and financial performance ($p &lt; .01$ in all cases). 3/3 predicted effects</td>
<td>Not formally tested, but sub-group analyses suggest interaction with business strategy</td>
<td>Not examined</td>
</tr>
<tr>
<td>13 Jayaram et al. (1999)</td>
<td>Cross-sectional</td>
<td>Size</td>
<td>Cost ($p &lt; .01$), flexibility ($p &lt; .05$) and time ($p &lt; .05$). Each aspect of HRM related to its own outcome (e.g. HRM cost to cost performance); HRM generic related to time performance only ($p &lt; .05$). 4/8 predicted effects</td>
<td>Not examined</td>
<td>Not examined</td>
</tr>
<tr>
<td>14 Vandenberg, et al. (1999)</td>
<td>Cross-sectional</td>
<td>No</td>
<td>Flexibility positively associated with ROE; incentives and direction-sharing negatively related to ROE; training and work design unrelated to ROE: $p &lt;$ levels unclear, and insufficient detail of analysis. 1/5 predicted effects</td>
<td>Not examined</td>
<td>Not examined</td>
</tr>
<tr>
<td>Study</td>
<td>Design</td>
<td>Variables</td>
<td>Findings</td>
<td>Summary</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>-----------</td>
<td>----------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>15 Wright et al. (1999)</td>
<td>Cross-sectional</td>
<td>Technology, size, age</td>
<td>Only training related to financial performance – negatively (p &lt; .05). 0/4 predicted effects</td>
<td>Internal fit – selection, pay and appraisal interact with participation (positive effect where participation is high, strong negative effect where participation is low) – 3/4 predicted effects</td>
<td></td>
</tr>
<tr>
<td>16 Bae and Lawler (2000)</td>
<td>Cross-sectional</td>
<td>Unionization, sector, country of origin</td>
<td>HRM strategy positively associated with self-report (p &lt; .05) and both concurrent (p &lt; .01) and subsequent (p &lt; .05) objective (ROIC) performance measures. 3/3 predicted effects</td>
<td>Not examined</td>
<td></td>
</tr>
<tr>
<td>17 Fey et al. (2000)</td>
<td>Cross-sectional</td>
<td>Sector, size, years operating in Russia</td>
<td>Overall performance associated with promotion on merit for managers, and job security for other employees. 2/16 predicted effects</td>
<td>Not examined</td>
<td></td>
</tr>
<tr>
<td>18 Cappelli and Neumark (2001)</td>
<td>Longitudinal: two sets of panel data, 1977–93, n = 433 and 1977–96, n = 666</td>
<td>Many, including production function, age, sector</td>
<td>For none of the eight practices did their introduction relate to either productivity or change in productivity in the 1977–93 panel data set. Of the three practices in the 1977–96 panel data set, one (job rotation) related to both outcomes, but negatively. Thus 0/8 and 0/3 predicted effects</td>
<td>Internal fit between self-managed teams and profit sharing. No effects for nine other theoretically specified interactions (e.g. teamwork training × self-managing teams) – 1/10 As for main effects</td>
<td></td>
</tr>
<tr>
<td>19 Guthrie (2001)</td>
<td>Cross-sectional</td>
<td>Size, age, unionization, sector, relative pay</td>
<td>HIWP positively associated with reported productivity (p &lt; .01). 1/1 predicted effect</td>
<td>HIWP with labour retention (positive effect of HIWP for firms with high labour retention, negative for low) – 1/1 Not examined</td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Research design</td>
<td>Control for third factors</td>
<td>Main effects (summary, statistical significance, hit rate)</td>
<td>Interaction effects</td>
<td>Individual practice effects</td>
</tr>
<tr>
<td>-------</td>
<td>----------------</td>
<td>----------------------------</td>
<td>----------------------------------------------------------</td>
<td>-------------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>20 Richard and Johnson (2001)</td>
<td>Cross-sectional</td>
<td>Size, part of holding company/not, location, organizational stage (e.g. start up, growth)</td>
<td>SHRM not related to either ROE or productivity: 0/2 predicted effects</td>
<td>Strategic fit, HRM with capital intensity to predict ROE (HRM has stronger positive relationship with ROE the greater the capital intensity), but not productivity – 1/2</td>
<td>Not examined</td>
</tr>
<tr>
<td>21 Batt (2002)</td>
<td>Cross-sectional</td>
<td>Sector, unionization, customer segment</td>
<td>Work design and HR incentives positively related to sales growth (p &lt; .05 and p &lt; .01 respectively); no effect for skill level. 2/3 predicted effects</td>
<td>Composite HR index interacts with segment: stronger effect of HR on sales growth in residential and smaller centres</td>
<td>As for main effects</td>
</tr>
<tr>
<td>22 Way (2002)</td>
<td>Cross-sectional</td>
<td>Sector, unionization, size, capital intensity</td>
<td>No statistically significant relationship with productivity ratio; positive relationship (p &lt; .05) with overall perceived productivity</td>
<td>Not examined</td>
<td>Only, performance pay associated with perceived productivity (p &lt; .05)</td>
</tr>
<tr>
<td>23 Ahmad and Schroeder (2003)</td>
<td>Cross-sectional</td>
<td>Country, sector</td>
<td>Practices related to performance at p &lt; .05 or p &lt; .01 in 9/12 cases. But the analysis fails to control for other practices (all of which are highly intercorrelated), hence proportion of predicted effects unclear</td>
<td>Not examined</td>
<td>As for main effects, but not controlling for others</td>
</tr>
<tr>
<td>Study</td>
<td>Design</td>
<td>Context</td>
<td>HRM Practices</td>
<td>Outcome Measures</td>
<td>Findings</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-------------------</td>
<td>-------------------------------------------------------------------------</td>
<td>----------------------------------------</td>
<td>-----------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>24 Guest et al. (2003)</td>
<td>Quasi-longitudinal</td>
<td>Size, sector, unionization, consultation, HR cost strategy, prior performance, others</td>
<td>Use of HRM practices unrelated to subsequent productivity or profit (when prior productivity or profit controlled). 0/2 predicted effects</td>
<td>Not examined</td>
<td>Not examined</td>
</tr>
<tr>
<td>25 Wright et al. (2003)</td>
<td>Cross-sectional</td>
<td>None, but all business units in same corporation and fairly homogenous</td>
<td>HR practices not statistically significantly related to productivity but positively related to profit (p &lt; .05). 1/2 predicted effects</td>
<td>Not examined</td>
<td>Not examined</td>
</tr>
</tbody>
</table>
and certainly financial ones, though ostensibly covering the same period when the HR practices data were collected, necessarily reflect the result of prior performance. There is a lag built into performance data, as it takes time to sell, receive payment for and bank income, which is then aggregated over a financial year. At face value, therefore, such cross-sectional studies could be seen as examining the effect of performance on HRM practices rather than vice versa.

Of the four non-cross-sectional studies, two are quasi-longitudinal (Guest et al., 2003; Huselid et al., 1997). That is to say, they include repeated measurement of performance, but a single measurement of HRM practices. The first of these studies controls for concurrent performance, and the other for prior performance (Table 2, column 5). However, in neither case is it clear whether the earlier measurement of performance preceded the introduction of the HRM practices, so causal inference is difficult. Only two studies (Cappelli & Neumark, 2001; Ichniowski et al., 1997) have an authentic longitudinal design on which causal inferences about the relationship between HRM practices and performance could justifiably be based; and, as we shall see later, they yield divergent results.

**Control for third factors**

Especially in cross-sectional studies there is a need to control for third factors, that is, variables that may account for an association between HRM practices and performance. For instance, if larger organizations, those with greater (or lesser) trade union involvement, or those in particular sectors, tend to have both more sophisticated HR provisions and better performance, then a relationship between HR and performance would be found for one or other of these reasons. Better studies will use controls for third factors, less satisfactory ones will not.

Overall, the studies do well in this respect. Control for third factors is evident in all but two studies (Table 2, column 3). Size is included in the large majority of studies, but there is considerable variability with respect to which other factors are controlled. Our judgement, however, is that the variability is largely consistent with the nature of the study (e.g. controls for sector in heterogenous samples, and for product complexity when examining build-hours on assembly lines).

**Main effects**

Our summary of findings for the relationship between HRM and performance is based on the most sophisticated analysis presented by the authors,
but before any tests for interactions. Thus if both zero-order correlations and regression analyses controlling for third factors are presented, it is the findings from this that are used. However, if regression analyses excluding and including interaction terms are presented, we rely on the findings from the former. We adopt the latter position to increase comparability between studies that did and did not test for interactions (see next section). We also only accept findings that are statistically significant at $p < .05$. For ease of presentation in Table 2 (columns 4–6), and in places in the text that follows, we refer to findings as effects in line with the allocation of variables to predictor and outcome status within regression analysis. However, we do not intend this to signify a causal link.

The main effect findings are summarized in Table 2, column 4. These show that 19 of the 25 studies report some statistically significant positive relationships between HRM practices and performance. At the same time, however, six studies fail to find any such relationships (e.g. Cappelli & Neumark, 2001; Guest et al., 2003), and in one of these (Wright et al., 1999) there is a statistically significant negative relationship (between training and financial performance). In parentheses, it is interesting to note that there is no evidence that studies using self-report measures of performance provide more positive findings than those using objective measures.

The balance of evidence towards positive relationships is weakened by the lack of consistency within studies. The large majority used multiple performance indicators, and many also used multiple HRM measures. This means that there were multiple opportunities to show HRM–performance relationships, of which only a few are statistically significant. This is shown by the hit rate also reported in Table 2, column 4. In Huselid’s (1995) seminal study, for example, we should judge the finding that the employee motivation HRM scale was statistically significantly associated with productivity and Tobin’s Q against the fact that the employee skills and structures HRM scale was not statistically significantly related to either of these performance outcomes. Correspondingly, though the employee skills and structures HRM scale was related to GRATE, the employee motivation HRM scale was not. Similarly, Delery and Doty (1996) identified seven aspects of HRM and two performance indicators. Only 4 of the 14 relationships were statistically significant. Other studies involving more than a single relationship are equally patchy (e.g. Youndt et al., 1996, with one of six predicted effects).

Moreover, effect sizes are typically small, and the criteria used to judge statistical significance, and hence to draw conclusions about the reliability of findings, are often lenient, even in large sample studies. With a sample of more than 800, for example, Huselid (1995) recorded an effect size for his
two comprehensive measures of HRM combined on productivity of only 1 percent, and treated one-tailed \( p \)-values of < .10 as statistically significant. Similarly, Wood and de Menezes (1998), also with a sample of over 800 establishments, use a \( p < .10 \) statistical significance criterion.

The acceptance of small effect sizes and the use of moderate levels of statistical significance are perhaps indicative of a more general tendency in the literature towards over-positive interpretations of results. Guest and Hoque (1994), for instance, distinguished between four types of HRM system but found no statistically significant difference across them on reported productivity or quality. This did not prevent them concluding that ‘the results demonstrate that strategic HRM pays off’ (p. 11). Virtually all authors reach positive conclusions irrespective of the strength and consistency of their findings.

**Interaction effects (fit)**

All studies of the HRM–performance relationship would benefit from tests for interaction effects, and this is so for several reasons. If the hypothesis is that HRM practices have a universal effect on performance, then it adds strength to any observed main effect to show it is invariant across different circumstances (e.g. according to sector or company size). Conversely, if theory predicts internal, organizational or strategic fit, then tests of interactions are relevant for that purpose. Whatever the particular theoretical rationale, the basic point is that investigating possible interaction effects is a means of more fully understanding the nature of any observed relationship between HRM practices and performance, and enhancing the construct validity of the study.

Twelve of the 25 studies do not report tests for interactions. Of the other 13, 2 test for interactions using non-standard methods (e.g. sub-group analysis), and the remaining 11 deploy orthodox tests (e.g. moderated regression). Looking at those studies in more detail, the emphasis is on internal and strategic, rather than organizational, fit. However, the main impression is of considerable variation in the interactions examined and inconsistency in the findings recorded. Thus there is no compelling evidence either of synergy within HRM systems (i.e. that the effects of systems as a whole is greater than the additive sum of their parts), or of any systematic strategic fit effects.

**Individual practice effects**

Most studies of the HRM–performance relationship aggregate individual practices into multi-component scales (e.g. Huselid, 1995; Way, 2002), or
less commonly they group organizations on the basis of their HRM profile (e.g. Ichniosiowski et al., 1997; Wood & de Menezes, 1998), in order to test for effects. An argument for this approach is that one would expect a composite of HR practices to have a greater impact on overall organizational performance than any single practice. Nonetheless, all those composite measures include particular practices (e.g. training, empowerment, and performance-related pay) each of which may enhance performance in its own right. Thus the question arises as to whether one or a few among those constituent practices may account more parsimoniously for any observed effect of the overall HRM system on performance, or whether all are an integral part of the whole. This is not only of theoretical importance, but is also an issue with practical significance for managers wishing to stage the implementation of beneficial HR practices and avoid ineffectual ones. Thus examination of the relative effects of different component practices adds to the construct validity of the investigation.

This issue has been ignored in the majority of studies. Fifteen of the studies (Table 2, column 5) use composite measures of HRM without reporting any findings for their component practices. Results for component practices from the other 10 studies do not show any consistent pattern. Given this lack of evidence, it is difficult to be certain that it is the HRM system that relates to performance rather than one or particular sets of its components.

A wider consideration
A final methodological issue, not covered in Tables 1 and 2 because it is largely invariant, concerns the nature of the measurement of HRM practices. With only one exception (Huselid et al., 1997), studies have focused simply on the use of the practices. It could be argued, however, that effective use may be a better measure. This is because the inadequate implementation of a given practice may do more damage than no implementation at all. For example, adopting team briefings without allowing employees to express ideas may have a worse effect than not considering such involvement at all.

Expert witness testimony: part 2
The second question addressed to our expert witness concerned whether available evidence could be interpreted as establishing a causal link between HRM practices and performance. The advice has to be that the evidence is promising but only circumstantial. The unknown reliability of measures of HRM, the paucity of studies with adequate research designs, and the inconsistent results both across and within studies is troublesome. Taken together with the likelihood that positive findings are more likely to be published than
negative ones, it is clear that existing evidence for a relationship between HRM and performance should be treated with caution.

Summary and discussion

The conclusion from our analysis is that it is premature to assume that HRM initiatives will inevitably result in performance gains, either in all situations or even where deemed appropriate by contingency arguments. Consultants who promote the HRM model are reflecting a wider academic and business school perspective in which this approach is central to definitions of modernity in management. Governments, employers’ associations, professional associations such as those representing personnel managers, and even trade unions, throughout the world, have also promoted high-commitment HRM as the approach most suited to the assumed increasingly turbulent international economy.

Our assessment is that, although consultants are acting in good faith, and their views are seemingly reinforced by the presumption on the part of academics that HRM systems actually do promote organizational performance, the empirical evidence is as yet not strong enough to justify that conclusion. The cross-sectional evidence could be over-estimating such a relationship due to contamination between measures of the HRM and performance. Conversely, studies could be underestimating the strength and consistency of the relationship through inadequate measurement of HRM practices. We lean to the latter view, because while there is little evidence of common method bias leading to spurious conclusions (Wall et al., 2004) there are strong indications of poor measurement of HRM practices (Gerhart et al., 2000) that would produce attenuated effects. Moreover, regardless of the quality of data, the paucity of longitudinal studies also makes causal inference dubious. So the evidence is at once encouraging but ambiguous. Because there also remain strong theoretical grounds for believing an HRM system centred on enhancing employee involvement should be beneficial for organizational performance, research that overcomes the weaknesses of current studies is required.

Implications

There are three main implications of our analysis. First, there is a need, on the basis of the findings so far, to temper the claims being made. A first step is to moderate the language used. The term ‘high performance’ clearly presupposes the very effects researchers should be investigating, and should
be avoided. Other terminological changes would also help foster a more temperate debate. For example, use of words such as ‘impact’, ‘determinant’ and ‘effect’ when presenting cross-sectional findings, a practice too often evident in the existing literature, should be avoided in favour of terms such as ‘associated with’ or ‘related to’.

Second, adopting a strategy of articulating and investigating the relative merits of competing hypotheses may encourage more rigorous and disinterested research. This has long been advocated (see Chamberlin’s ‘method of multiple hypotheses’ as proposed in his presidential address to the US Academy of Science in 1890; reprinted in *Science*, 1965) as a means of combating bias in science, and is now more feasible than ever given developments in computing and statistical methods, such as structural equation modelling. By evaluating competing hypotheses it is less likely that researchers will align themselves with a particular hypothesis. One of the attractions of some of the early studies, most notably Huselid’s (1995) and Delery and Doty’s (1996), was precisely that they aimed to test between the universalistic and contingency theories of HRM. However, the selection of competing hypotheses needs to be much broader. For example, these could include hypotheses of differential impacts of HRM systems depending on the performance measure used. Such development must, in our judgement, go hand-in-hand with theoretical development.

Third, and most importantly, the need is for research designs that overcome the problems of existing studies. Among these, three weaknesses are paramount. The first is the reliance on single-source measures of HRM practices (e.g. from CEO or HRM manager) of unknown reliability, sensitivity and validity, often from the same source as the measure of performance. The second is the use of small samples coupled with low response rates. The third is the lack of sophisticated longitudinal studies, especially ones examining how change in HRM practices relates to subsequent change in performance. We now discuss the methodological implications of those three weaknesses for future research.

**Independent audit of HRM practices**

Others have already pointed to the need to improve the reliability and validity of measures of HRM by obtaining data from multiple sources (e.g. Wright & Gardner, 2003). We would go further to suggest that it is essential to have an independent audit of HRM and to cover comparison practices (e.g. R&D emphasis, total quality management). The audit should be conducted by individuals from outside the organization but familiar with the wider use of the practices in question, so that they can calibrate their
assessments accordingly. They would draw on multiple respondents and sources of data within organizations, and develop methods to achieve a high degree of concordance across different auditors. They would measure both the extent and effective use of practices, and be unaware of the performance of the organization. The use of selected comparator practices would enable determination of the importance of HRM practices relative to others. Because size of effect can have limited meaning, it is instructive to know if HRM practices have a stronger or weaker relationship with productivity than other practices or factors (e.g. R&D expenditure). Expanding predictor variables in this way also allows tests for possible interactions between HRM and such other practices (e.g. HRM with TQM, see Lawler et al., 1995).

Large samples, high response rates

Many of the important questions are ones that also require large samples. Testing for a relationship between HRM and performance benefits from the control of third factors, and also from examination of the effect of individual component practices relative to each other and to measures of overall HRM. The more such variables are taken into account, the more convincing any relationship is observed. However, the greater the number of variables involved, the larger the sample size required.

Sample size is even more of an issue when addressing theoretically predicted interactions (i.e. the different types of fit), where the lack of sensitivity of orthodox moderated regression means large samples are essential (Aiken & West, 1991; Busmeyer & Jones, 1993). Of course, one could argue that taking a universalistic perspective reduces the need for such large-scale research. Yet, this is not so. Following our earlier call for testing competing hypotheses, it is important to see whether the effect of HRM on performance is invariant across conditions (e.g. private vs. public sector; manufacturing vs. finance; strategic fit) rather than leave that issue unaddressed. Finally, of course, large sample sizes, assuming also high response rates, provide a firmer base for generalization.

Longitudinal research designs

The final requirement we emphasize, as do others (Guest, 1997; Wright & Gardner, 2003), is the need for longitudinal research. Some such investigations could be retrospective. Thorough independent audit of HRM practices, including evidence of when they were first introduced within organizations, could be set alongside available data on performance going back many years. Then the effect of the introduction of practices on
performance could be estimated. Other studies should be prospective, based on repeated independent audits, accepting the fact that findings will not be available for several years. The advantage of prospective over retrospective investigation is greater sensitivity and certainty in the measurement of change, and better coverage of both use and effective use of HRM practices.

The case for big science

The principles underlying our design criteria are well known and widely accepted as good practice. The weaknesses in the existing research designs cannot, therefore, be attributed to investigators being unaware of how better to conduct research. Rather, they stem from the almost inevitable limitations of the small-scale funding that characterizes social science. Funding for an individual or research team, over just a few years, does not provide the resources to deploy the recommended methods. Consider the need for independent audit, where experience suggests that it takes at least 2 days to obtain the requisite information for a single organization (see Patterson et al., 2004). Given such information should be obtained separately by at least two individuals (to determine reliability), and for large samples of organizations (say 400 in five key sectors), the data collection task alone would take many person years (i.e. 2 people × 2 days × 2000 organizations = 8000 days, which at 230 working days per year = some 35 person years). Add to this the final recommendation, for the repeated measurement of practices and performance as required by longitudinal research, and minimally the necessary resource is doubled.

Research on that scale is seemingly beyond the scope of current funding mechanisms. It is not, however, beyond the scope of more intensive collaboration than is the current norm among the many potentially interested parties, including academics, employer organizations, employee organizations, professional bodies and government departments. The involvement of government agencies is also important to encourage the desirable level of compliance (i.e. high response rate). Moreover, there are precedents for such activity. There are already numerous information-collecting exercises in many countries with which individuals (e.g. the census) and organizations (e.g. company financial reports) have to comply. There are assessments within organizations based on independent audits (e.g. ‘Investors in People’ in the UK); and there are surveys based on collecting data through interviews that are designed and managed by collaborative teams involving government departments, research councils and academics, as in the case of the Work Employee Relations Survey (WERS) series in the UK, or the equivalent Australian survey.
We propose that a big science project on management practices and performance builds on the best features of these exercises by devising a new study that covers practices well beyond employment relations and HRM, uses independent audit of those practices, and is at the level of the company or firm (rather than site) for which there is official financial performance data. If academics, practitioners and policy makers are serious in wishing to understand the effect of HRM and other practices on performance, then these kinds of data will be required. They must also face the possibility, of course, that no strong causal effects will be found.

Postscript

It will be appreciated that the focus of this article is on a particular issue concerning HRM that has dominated recent research – the relationship of high involvement forms of HRM with organizational performance. This should not be allowed to mask the fact that, even if such a performance effect were not demonstrable, the use of practices that enhance employee involvement may be an end in itself. Such personnel management practices also may be an important means of providing justice and equality of opportunity. Other reasons for studying HRM practices and systems could then include, for example, their role in enhancing employee well-being and promoting non-discriminatory practices. It will also be recognized that the large-sample quantitative approach taken here is a reflection of the question addressed. Where an HRM effect on performance is established, then important questions concerning the conditions under which this applies, the mechanisms through which it operates, and any unintended consequences of the HRM practices, will require supplementary quantitative and qualitative investigations. Such detailed research can also shed considerable light on why no performance effects accrue when that is found to be the case. In other words, our recommendation for big science should not be pursued to the exclusion of complementary options.

Acknowledgements

The support of the Economic and Social Research Council (ESRC) UK is gratefully acknowledged. The work was part of the programme of the ESRC Centre for Organisation and Innovation.
References


Toby Wall is Professor of Work Psychology at the University of Sheffield, where he is Director of the Institute of Work Psychology and of the ESRC Centre for Organization and Innovation. His interests span work design, innovation, employee well-being, absence, human resource management, work and organizational performance, and applied research design. He has published 12 books and over 150 articles in books and journals. The latter include papers in the Academy of Management Journal, Journal of Applied Psychology and Personnel Psychology. He has provided policy advice to government departments and to numerous organizations.

[E-mail: t.d.wall@sheffield.ac.uk]

Stephen Wood is Research Chair and Deputy Director, Institute of Work Psychology, Professor of Employment Relations, School of Management, and Co-Director, ESRC Centre for Innovation and Organisation, University of Sheffield. His current interests include high involvement management, lean production, work design, idea capturing schemes, family-friendly management, employment relations and the social and economic challenges of nanotechnology. Publications include (with H. Gospel) Representing workers: Trade union recognition and membership in Britain (Routledge, 2003); The transformation of work (Unwin Allen, 1989); and (with Richard Jones and Alison Geldart), The social and economic challenges of nanotechnology (Economic and Social Science Research Council, 2003). He has provided policy advice to government departments, professional associations and public and private sector organizations.

[E-mail: s.j.wood@sheffield.ac.uk]