DISCUSSION PAPER SERIES

IZA DP No. 9285

The Complementary Use of Experiments and Field Data to Evaluate Management Practices: The Case of Subjective Performance Evaluations

Patrick Kampkötter
Dirk Sliwka

August 2015

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

The Complementary Use of Experiments and Field Data to Evaluate Management Practices: The Case of Subjective Performance Evaluations

Patrick Kampkötter
University of Cologne

Dirk Sliwka
University of Cologne,
CESifo and IZA

Discussion Paper No. 9285
August 2015

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 9285
August 2015

ABSTRACT

The Complementary Use of Experiments and Field Data to Evaluate Management Practices: The Case of Subjective Performance Evaluations*

Most firms rely on subjective evaluations by supervisors to assess their employees' performance. This article discusses the implementation of such appraisal processes, exploring the use of multiple research methods such as the analysis of personnel records, survey data, and lab and field experiments to study them in detail. We argue that the complementary use of these methods helps to build a better understanding of how subjective evaluations are conducted and appraisal systems should be designed.

JEL Classification: D22, J33, M12, M52

Keywords: subjective performance evaluation, performance appraisals, management practices, experiments, field data, LPP

Corresponding author:

Dirk Sliwka
Seminar of Personnel Economics and Human Resource Management
University of Cologne
Albertus-Magnus-Platz
50923 Köln
Germany
E-mail: dirk.sliwka@uni-koeln.de

* We thank Andrew Kinder, Tommaso Reggiani, and participants of the 2014 annual meeting of the German Economic Association of Business Administration (GEABA) in Regensburg for helpful comments and suggestions. We thank the German Research Foundation (DFG) for financial support through priority program SPP 1764 (SL 46/2-1) and the research unit "Design and Behavior - Economic Engineering of Firms and Markets" (FOR 1371).

1. Introduction

Empirical studies applying different economic methods to evaluate the impact of management practices on firm and employee outcomes have become increasingly important in recent years. In management research, scholars have already for quite some time studied the connection between the use of so-called high performance work practices, i.e., combinations or bundles of human resource (HR) management practices, and employee or organizational outcomes, typically with quite mixed results (Huselid, 1995; Combs et al., 2006; Subramony, 2009). More recently, also economists have conducted large-scale survey studies to investigate the connection between rather general management practices and firm performance (Bloom and van Reenen, 2010, 2011; Bloom et al., 2012). Bloom and van Reenen (2007), for instance, use telephone interviews to evaluate firms' HR practices along various dimensions, such as monitoring, target setting, and people management and find that higher "management scores" are correlated with firm performance. An important "side-effect" of these survey studies, sometimes a bit overlooked in the economics literature, is that they give us a broad overview on what firms actually do and how frequently they do it which in itself is important as it informs researchers on the relevance of different management practices. But, due to the mainly cross-sectional nature of these data sets, these studies typically cannot establish causal effects. Some of the potential endogeneity issues can be addressed with panel data (Huselid and Becker, 1996; Black and Lynch, 2004), but when there are time-varying unobserved variables that simultaneously affect both the use of a management practice and firm performance the estimated effects will still be biased.

On the other hand, a recent literature has emerged in economics that focuses on field experiments in firms to evaluate the impact of individual management practices (see, for instance, Bellemare and Shearer, 2009; Bandiera et al., 2011; Englmaier et al., 2012; Hossain and List, 2012; Delfgaauw et al., 2013; Manthei and Sliwka, 2014; Friebel et al., 2015). Typically, these field experiments cover a single firm[1] and a specific form of a practice, but in contrast to the broader survey studies they allow a clean and credible identification of causal effects.

[1] An exception is the field experiment conducted by Bloom et al. (2013) among a large number of Indian textile firms, in which randomly chosen treatment plants received free management consulting services. The results reveal that this informational advantage leads to significant increases in productivity in the treatment firms compared to the control group.

Still, for very practical reasons these field experiments are limited in the number of feasible treatment variations, often study very specific industries and mostly cannot directly observe individual behavior but only infer information about behavior from performance indicators. Lab experiments, on the other hand, allow for observing behavior directly and, in principle, researchers can easily implement treatment variations that help to disentangle behavioral channels.

However, the "external validity" of lab evidence is sometimes called into question. If we identify a certain behavioral channel in the lab, it is of course legitimate to ask to what extent we can be sure that this mechanism is of equal importance in a natural setting in a firm and will not be dominated by other mechanisms not captured by the specific experimental design (see, for instance, the discussion in Levitt and List (2007)). Still, as Camerer (2015) argues, the primary goal of most experiments is to understand the general, underlying behavioral principles in a controlled environment and not to establish results that are generalizable from the lab to the field. Nevertheless, he presents ample evidence of lab findings that have proven to be consistent in comparable field settings. But a similar argument can also be made for field experiments: If we have clean causal evidence for a specific result in one firm or a specific industry, to what extent can we be certain that this result will also hold in different firms? In other words, when moving from the lab to a field experiment, we can of course now make a more precise claim about the impact of an intervention on the treated subjects in the specific firm – but we still cannot be sure to what extent this can be generalizable to other firms. Moreover, very often it is simply infeasible to run a field experiment since for legal or practical reasons certain instruments cannot be randomly assigned. In the context of development economics, Deaton (2010) somewhat provocatively states that the price for success of randomized field experiments in identifying causal effects of a specific program "*is a focus that is too narrow and too local to tell us 'what works'*" (p. 426).[2]

If we want to collect insights on how management practices affect outcomes in firms, and, if we ultimately want to help practitioners to design better management practices, we need to combine the relative strengths of all these approaches. The key reason is that most management practices affect the performance of organizations through different interlinked

[2] Heckman (1992), for instance, has argued that randomization itself could lead to biases in field settings. Imbens (2010) addresses the concerns raised by Deaton (2010) and acknowledges that cases exist where randomization is difficult or not feasible but strongly argues that, if they are feasible for the question we are interested in, "*randomized experiments are superior to all other designs in terms of credibility*" (p. 401).

behavioral and economic mechanisms. We therefore propose that economic research on management practices should focus on the following three key goals:

- to understand the different behavioral mechanisms at work when a practice is applied and to learn how these mechanisms may affect the impact of a management practice,
- to estimate the causal effect of its implementation,
- and to collect evidence about its relevance and frequency of use in companies.

To achieve these goals, researchers must necessarily apply multiple complementary research methods. Formal economic models help to develop a deeper and more precise understanding of potential behavioral mechanisms. Laboratory experiments are useful to isolate and disentangle these mechanisms in precisely controlled environments.[3] Field experiments and the econometric evaluation of quasi-experiments in firms help us to estimate the causal impact of instrument use on the performance of firms and the well-being of their employees. Moreover, they sometimes even allow us to estimate the magnitude of a performance effect, which can be used as a key ingredient for cost-benefit analyses. Finally, broad representative surveys among firms and employees give us more detailed information about the frequency and correlates of its use in real companies and thus generate further knowledge about the generalizability of insights from lab and field experiments. If there is no widespread adoption of a (well-known) practice for which there is a theoretical underpinning and experimental evidence showing that it causally affected performance in a specific environment, we have to ask why this is the case. Hence, in order to get a more comprehensive picture, it is important to exploit the complementary character of these approaches, instead of fighting scientific battles about their relative merits, a point that has already been stressed by Falk and Heckman (2009): *"[…] empirical methods and data sources are complements, not substitutes. Field data, survey data, and experiments, both lab and field, as well as standard econometric methods, can all improve the state of knowledge in the social sciences. There is no hierarchy among these methods, and the issue of generalizability of results is universal to all of them."*

In this paper, we want to illustrate and discuss potential applications of this mix of complementary methods using the example of one important HR management practice, namely subjective performance evaluations of employees (for other applications, see Englmaier and

[3] Ludwig et al. (2011) propose a further distinction stressing the importance of "mechanism experiments," i.e., field experiments that do not directly evaluate the impact of a policy but are designed to study a specific behavioral mechanism underlying a policy.

Schüßler, 2015). Subjective performance evaluations are widely used in many firms, but we still do not precisely know whether and how they affect various outcomes such as job performance or the satisfaction of employees. In detail, we present the results of several of our own studies using various methods to analyze the impact of differentiation in performance evaluations on the provision of individual efforts, as well as employee perceptions like job satisfaction or fairness perceptions in response to performance appraisals. Examples include the use of linked employer-employee data, an industry-wide field study, an "insider econometrics approach" combining data from personnel records and employee surveys, and a field and laboratory experiment.

2. Purposes of Performance Evaluations and Design Challenges

2.1. Purposes

Performance evaluations or performance appraisals[4] are mainly used to evaluate and monitor the contributions of individual employees to overall firm performance. They often combine the use of objective performance indicators and subjective evaluations. Firms typically use performance evaluations for multiple reasons (Landy and Farr, 1983; Murphy and Cleveland, 1991, 1995). First, in most incentive schemes individual bonuses are based on subjective and objective performance indicators generated through appraisal processes. Second, performance evaluations are often the starting point for employee "development" decisions, such as the assignment of training. Third, appraisal outcomes are used in the personnel planning process for decisions on promotions, reallocations, or dismissals. Frederiksen et al. (2012), for instance, analyze data sets on subjective performance ratings from six large, international companies that have been used in several prominent studies on internal labor markets and find that performance evaluations predict career outcomes such as promotions.[5] Fourth, performance appraisals give employees direct feedback about their performance and potential strengths and weaknesses. Feedback can show employees whether to reallocate efforts or to invest in new skills and, moreover, can have a direct impact on employee satisfaction and thus the decision to stay with an employer (Fletcher and Williams, 1996; Whitman et al., 2010; Kampkötter, 2015).

[4] In the following, performance appraisals, performance assessments, and performance evaluations are used interchangeably.

[5] See also Halse et al. (2011).

2.2. Design Challenges and Appraisal Formats

While performance appraisals sometimes also include objective performance information (such as financial key figures), most often subjective assessments by a supervisor play a dominant role. A key reason for this is that in many cases objective indicators of individual performance are not available (an exception is, for instance, the sales function, where objective performance measures are nearly always available and frequently used). This is typically the case in many cross-functional positions, such as human resources, controlling, finance, and marketing. Objectively measurable performance indicators can often only be derived jointly at the team level, and individual contributions to this team output are difficult to evaluate. Furthermore, individual performance is frequently rather complex and cannot be tracked with a small set of performance indicators. When individual performance strongly depends on external factors that are outside of the control of employees (such as the market situation, the state of the economy, etc.), objective performance measures can only be crude indicators for employees' efforts and talents. As a result, the majority of performance evaluations in practice are based on subjective assessments by supervisors.

There is substantial evidence, mostly from research in personnel psychology, showing that these subjective evaluations are typically "biased." Firms commonly use systems in which employees are assessed on a given scale (for instance, with evaluation grades ranging from 1 to 5), and often only a subset of the scale is actually used by supervisors. Psychologists have coined the terms of *centrality* and *leniency bias* to describe patterns that are frequently observed (Landy and Farr, 1980; Murphy, 1992; Prendergast and Topel, 1993; Kane et al., 1995; Murphy and Cleveland, 1995; Prendergast, 1999; Gibbs et. al., 2004; Moers, 2005; Frederiksen et al., 2012). In case of a centrality bias, the variation in performance appraisals is smaller than intended by the designer of the system, i.e., supervisors do not use the full range of the rating scale and especially avoid marginal grades. The so-called leniency bias describes a phenomenon where the mean of the appraisal ratings is higher than the mean of ratings intended by the firm, i.e., supervisors systematically evaluate their subordinates better than they are supposed to. Additionally, supervisors differ in the extent to which they are prone to these biases. Heterogeneity in the supervisors' types therefore leads to heterogeneous evaluation behavior even within the same firm (see, for instance, Bernardin et al. (2000) for a study on the role of personality factors in appraisal behavior).

In practice, firms have adopted several instruments to reduce potential biases which presumably lead to more differentiation among employees. Two instruments prominently used

in practice are *recommended distributions* and *evaluation panels*. When a firm adopts a recommended distribution, it tells managers the relative proportion of different grades that should be assigned. In the appraisal system of the multinational firm studied in Ockenfels et al. (2015), employees were rated on a 5-point scale, and the firm recommended the following distribution:

Grade 1 ("excellent"): ≤ 5% of employees
Grade 2 ("above average"): ≤ 25% of employees
Grade 3 ("fully meets expectations"): ~ 60% of employees
Grade 4 ("below average") and 5 ("inadequate"): together ≤ 10% of employees.

However, as recommended distributions are non-binding and just give a guideline on how to assess performance, they may be accompanied by more lenient and less differentiated actual ratings. In the studied firm, for instance, ratings were tied to (budgeted) bonus payments, and more than 65% of employees received a rating of "3" and more than 30% a rating of "2". On the other hand, less than 5% received a "4" and nearly nobody a "5", which is also a very common occurrence in other firms. Frederiksen et al. (2012), for instance, investigate typical patterns in subjective evaluations from several data sets used in the prior literature and find similar or even more extreme patterns.[6]

Some firms therefore adopt stricter so-called *forced distributions*, where these proportions are not guidelines, but rather the appraisal process is designed such that the evaluators must adhere to a given exact distribution. The most prominent example of such "grading on a curve" is General Electric, where Jack Welch's "vitality curve" forces managers to identify the top 20% and bottom 10% of employees each year (see Welch, 2001, chapter 11).

A key challenge in grading employees is that individual managers often supervise and therefore evaluate only a small set of employees. Even if a manager wants to be accurate in rating her employees and is able to rank them, the fact that somebody belongs to the top 20% in a certain unit does not guarantee that this person belongs to the top 20% of the whole firm. If the mix of talents and performance is unequally distributed across teams, this directly leads either to different evaluation standards (if managers have to stick to the distribution within their

[6] Among white collar employees of the former Dutch airplane manufacturer Fokker (Dohmen et al., 2004), 81% received the middle grade and 14% the top two grades. In the Baker-Gibbs-Holmstrom data set (Baker et al., 1994a, b) of a US-based service sector firm, 82% of the employees were rated with one of the two top grades, and in the Flabbi-Ichino data set (Flabbi and Ichino, 2001) of a large Italian bank, this fraction is 83%. Note that in all of these examples, 5-point rating scales are applied. See Frederiksen et al. (2012) for details.

team) or makes the process very complicated as evaluations have to be coordinated (if managers try to adhere to the distribution not within each team but across a larger number of teams). As a response, many bigger firms have in recent years adopted so-called *evaluation panels* (sometimes also called calibration meetings, management panels, or evaluation round-tables), in which a group of managers meet to discuss the performance evaluation of all of their employees. It is quite common, for instance, that a group of 60 or 80 employees are discussed in such a panel and that top management and HR representatives are involved in this process.[7] These panels serve to "calibrate" evaluations made by individual managers in order to generate common standards. Moreover, if a recommended or forced distribution is adapted, these panels make it easier to stick to this distribution reducing the likelihood that employees receive "unjustified" ratings because of the composition of their direct work group.[8]

In section 3, we present recent descriptive evidence of the use of these appraisal procedures from a novel representative data set on HR practices.

2.3. The Controversial Role of Differentiation

From an economics perspective, subjective evaluations entail a potential conflict of interest between the evaluating supervisor's personal interest and the interests of the firm in its role as the employer of both the supervisor and the evaluated employee. A large body of evidence in behavioral economics has shown that people have social preferences (see Fehr and Schmidt (2006) for a survey), i.e., their own well-being also depends upon the well-being of other people in their proximity. In this respect, there is evidence that both direct *altruism* (i.e., a person can ceteris paribus be better off when another person has a higher payoff) and *equity concerns* (i.e., a person is ceteris paribus better off when an outcome leads to a more equal payoff distribution) matter.[9] Moreover, in both respects *reference points* seem to play an important role in light of the substantial evidence that people often evaluate outcomes relative to a reference standard (Kahneman and Tversky, 1979). And these reference standards are affected, for instance, by peoples' own prior expectations (Bell, 1985; Loomes and Sugden,

[7] See, for instance, Michaels et al. (2001) or Welch (2001) for a description of typical processes.

[8] To give an example: If a manager has to evaluate 5 direct subordinates and can assign the highest evaluation to the top 20%, the likelihood that either no one or more than one of her direct subordinates belong to the top 20% in the firm is rather large. The bigger the group, the smaller is the likelihood that such "unjustified" ratings have to be assigned.

[9] But behavioral economics research has also established that individuals not only may have prosocial but some have also antisocial concerns such as spite, envy or even direct antisocial preferences (see, for instance, Zizzo and Oswald (2001), or Abbink and Sadrieh (2009)).

1986; Köszegi and Rabin, 2006) or by "social reference points," i.e., the outcomes of others (Bolton, 1991; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000).

If now a supervisor has to evaluate a direct subordinate, she has to trade-off the effects of her rating on the employee's well-being and the firm's performance. Hence, supervisors' preferences will often not be perfectly aligned with the interests of firms. For example, while an employer may prefer to have accurate evaluations that reflect differences in performance, supervisors may be tempted to assign generous ratings or not to differentiate between their subordinates. Very often subjective evaluations determine bonus payments to employees. Assigning better ratings thus leads to higher wages, often without substantial direct costs for supervisors.[10] A supervisor who cares about the well-being of her subordinate will therefore have a tendency to assign more lenient ratings (see Prendergast and Topel (1996) or Giebe and Gürtler (2012) for formal models on this issue). Breuer et al. (2013), for instance, analyzed longitudinal data from a call center organization where objective performance measures are available. They use variation in team composition over time to show that employees receive better ratings for the same objective performance when they have worked with the same supervisor over a longer time frame or when the supervisor manages a smaller team, demonstrating the effect of social ties between supervisors and subordinates on appraisals. Supervisors may also avoid poor ratings out of a reluctance to provide negative feedback, even if the actual performance was poor, because negative feedback typically has to be justified in more detail and may induce "psychological costs." Lenient ratings will also more likely prevent conflicts with subordinates (Varma et al., 1996). On the other hand, the social preferences of a supervisor may also affect the scope for relational contracts if supervisors and subordinates interact repeatedly (see Dur and Tichem (forthcoming) for a formal analysis of the role of a supervisor's altruism but also potential spitefulness in relational contracting under subjective evaluations).

Frequently, bonus payments have to be paid from a given budget, which is particularly true for the banking and financial services sector (Kampkötter and Sliwka, 2014). In this case, rater "leniency" is restricted as there is an upper limit on the assignable ratings. But when supervisors either have a preference for equity among employees or take employee's equity

[10] Sometimes there are of course indirect costs (as will be made clear below): for instance employees' performance may be lower which in turn may hurt the supervisor.

concerns into account, a "bias towards centrality" or reduced *differentiation* among employees directly follows.[11]

On the other hand, even if ratings assigned by supervisors are not fully "accurate," it is not clear at the outset whether an apparent "lack of differentiation" could in fact be due to reasonable behavior by supervisors that even may be to some extent in line with the firm's interest. Appraisal patterns interpreted as a "bias" from one perspective may in principle be beneficial from another perspective. Consider the following example. Suppose that we are looking at a system in which supervisors are asked to give the worst 20% of employees the worst performance grade. A supervisor who does not assign this worst grade to an employee he considers to be actually in the bottom 20% of course creates a bias if the purpose of the rating is to identify the employee's relative standing in the talent distribution. Moreover, from a neo-classical incentive perspective, such a bias may also be detrimental as low performance is then not sanctioned and high performance not adequately rewarded, which may reduce the incentive to exert higher efforts in the future. However, this supervisor may argue that his rating behavior is justified by another purpose. A large literature in experimental economics, starting with Fehr et al. (1993, 1997), has established that individuals have a preference for reciprocity. Thus supervisors' "leniency" may actually to some extent act as a trigger for a higher employee motivation through positive reciprocity, or the avoidance of low ratings may arise from the fear of demotivating agents and causing negative reciprocity. Sebald and Walzel (2014), for instance, show in a laboratory experiment that employees sanction their supervisors when assessments deviate negatively from an individual's self-evaluation of her own performance.[12] Hence, there may be trade-offs between the accuracy of the ratings and its other purposes. Even if we only take an incentive perspective, there may be a trade-off between triggering more social motives via positive reciprocity or extrinsic motives by punishing low performance. On top of that, excessive leniency may lead to unfair treatment of high performers and reduce their motivation: If high performers have a concern for equity not only in bonus payments but also in the exerted effort costs they may be tempted to reduce their efforts if low performers receive generous ratings at lower effort levels. This highlights the importance of studying these trade-

[11] See Grund and Przemeck (2012), Kampkötter and Sliwka (2014), or Ockenfels et al. (2015, online Appendix) for formal models analyzing the role of supervisors' or subordinates' preferences for equity in performance appraisals.

[12] Interestingly, the negative reciprocal action of employees to this perceived, unkind act by their supervisors also holds if the appraisals have no monetary consequences. See also Takahashi et al. (2014) who analyze the personnel records of sales representatives in a major Japanese car sales company and show that measures of rating biases are positively related to employee quits.

offs in detail. In the subsequent chapters, we will first show descriptive evidence on the use of performance appraisals and practices to foster differentiation before presenting a number of different studies that apply different methods to study the role and impact of differentiation in performance evaluations in greater detail.

3. The Use of Performance Evaluations in Firms: Descriptive Evidence

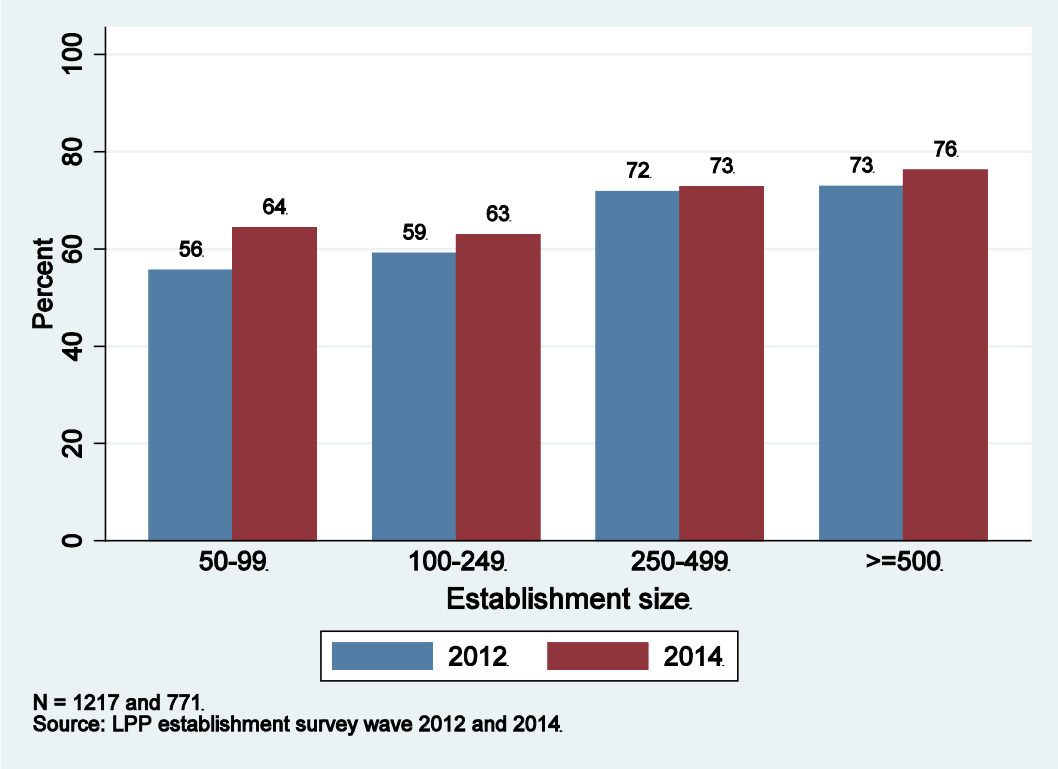
Performance appraisals are a core element of personnel policies in most firms. However, firms differ in the extent to which they apply appraisals. We start by analyzing early evidence from a new representative linked employer-employee data set of German firms, the *Linked Personnel Panel (LPP)*, which contains information about key elements of the appraisal process.[13] The authors designed this survey jointly with the Centre for European Economic Research (ZEW) Mannheim as part of a project on behalf of the German Federal Ministry of Labor and Social Affairs and the research institute of the Federal Employment Agency (IAB). So far, the first wave of both the employer and employee survey has been administered in 2012 and 2013, and the second wave will be completed in late 2015. The representative firm-level survey includes 1,219 private sector establishments employing at least 50 employees. The employee survey includes 7,508 randomly drawn respondents from 869 of these establishments. The key aim of this new data set is to develop a longitudinal infrastructure to assess the impact of HR practices on employees' "quality of work" (e.g., satisfaction, engagement, and turnover) and the economic success of firms. The LPP links employer-level information about HR policies with employee-level information about attitudes and behavior and enables researchers to analyze how individuals perceive and respond to HR policies. The data set provides information on various HR instruments on the firm level, including dimensions such as recruiting, performance management, employee and career development, training, corporate culture, and promotion of female employees. The employee questionnaire mirrors some of these practices such as training, promotion, and career development and additionally elicits information on employee perceptions such as job satisfaction, commitment, fairness perceptions, risk attitude, and personality traits. Over time, the survey will evolve into a panel data set that will allow to study within-firm variation of HR practices and link this to potential changes in employee perceptions and firm performance. Currently, the available data from the

[13] See Kampkötter et al. (2015) or Bellmann et al. (2015) for an overview on the structure of the data set.

first cross-section yields descriptive evidence on specific practices applied in performance evaluations.

Figure 1 shows the frequency of use of performance appraisals[14] by establishment size. We find that the majority of establishments use structured performance appraisals, with the frequency of use increasing from 62% in 2012 to 67% in 2014 across all establishment sizes. As the figure shows, larger establishments use systematic appraisals more often than smaller ones. However, we observe the most substantial increase in the smallest establishments.

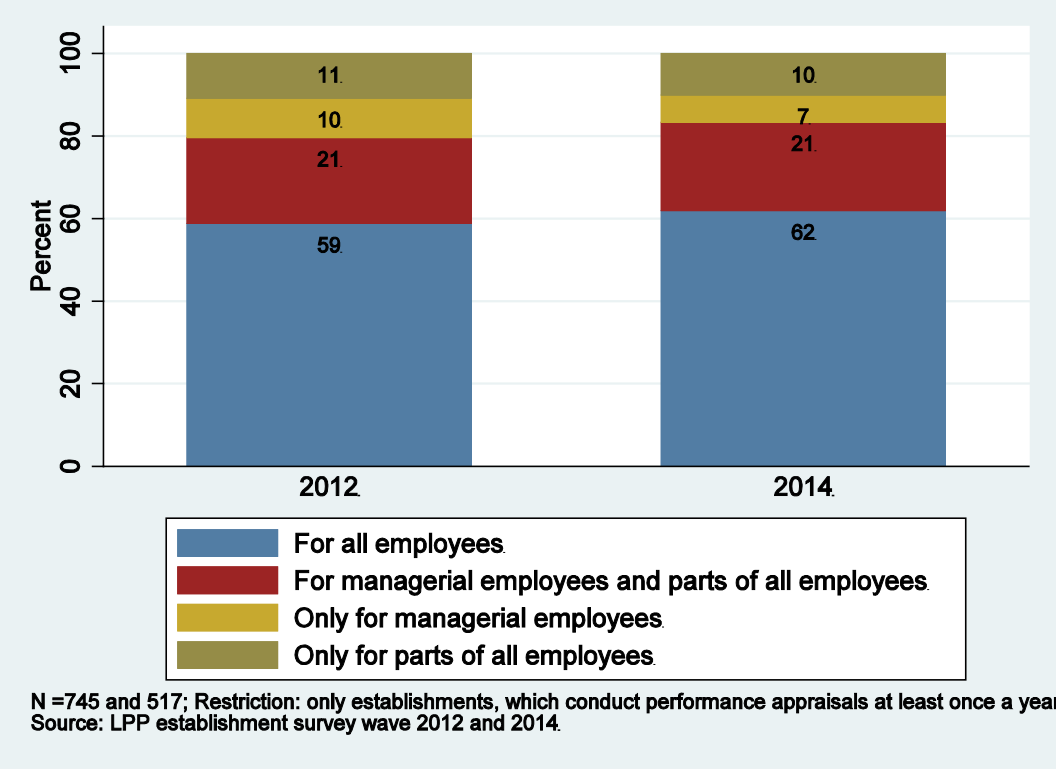
Figure 1: The Use of Performance Appraisals by Establishment Size



In a second step, we also asked about the appraisals' target group, for example, whether the practice is applied only to employees in a supervisory role (i.e., for managers) or for all employees. As Figure 2 shows, a majority of the establishments that use systematic appraisals indeed use these appraisals for all employees, and this fraction has (slightly) increased over the short time frame we consider.

[14] The exact item is as follows: "Is the performance of employees in your establishment evaluated by supervisors at least once in a year?"

Figure 2: Employee Target Groups of Performance Appraisals

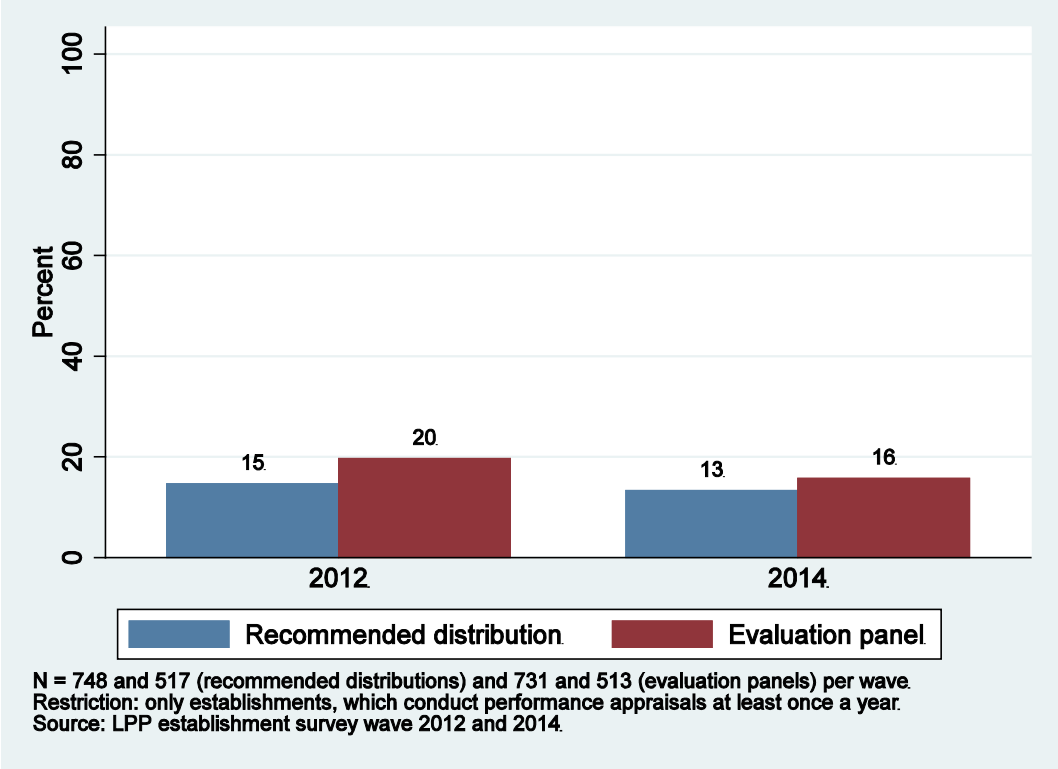


In the survey we also asked firms about the use of *recommended distributions* and *evaluation panels* (joint evaluation by more than one supervisor).[15] Figure 3 shows the distribution of the intensity of use over time for both practices. In 2012, only about 15%, and in 2014 only 13%[16] of the establishments using performance appraisals employed recommended distributions. However, the frequency of their use is higher in larger establishments: in establishments with 50-99 employees, only 10% employed recommended distributions, but this increases to 23% in establishments with more than 500 employees.

[15] The survey items are "Does a recommended distribution for performance assessments exist in your establishment? Recommended distributions convey information about the proportion of employees who should receive the best rating, the second-best rating, etc." and "Are employees typically assessed by *one* supervisor or jointly by *a group of supervisors* (management panels), i.e. not only by one supervisor?"

[16] We note that changes over time are partially due to sampling of firms as the LPP is an unbalanced panel.

Figure 3: Recommended Distributions and Evaluation Panels



In 2012, about 20% and in 2014 approximately 16% of the establishments using performance appraisals used evaluation panels or based their appraisals on joint evaluations by more than one supervisor. It is interesting to observe that it is more frequent in very small establishments (24% in 2012), more rarely done in medium sized establishments (13%) and again more often used in establishments with more than 500 employees (22%). This could possibly stem from the fact that in small firms the management team talks about the assessment of all their employees more often as the likelihood is sufficiently large that all members of the management team know all employees. This becomes more difficult in larger establishments. The larger the establishment becomes, the larger is then the likelihood that formalized procedures are used that establish a structure for group evaluations in panel discussions.[17]

The overall frequency of appraisal use shown in the above is in line with survey evidence from individual employees. Kampkötter (2015) uses the German Socio Economic

[17] In fact, as own discussions with HR managers of several big DAX30 companies in Germany have shown, structured evaluation panels are a typical element of their appraisal procedures and often there are tight guidelines regulating their implementation.

Panel (SOEP), a large and representative panel data set of a subset of the German population[18], to investigate the intensity of the use of performance appraisals among employees and their impact on job satisfaction. In the years 2004, 2008, and 2011, individual employees were asked whether their performance is evaluated regularly by their supervisor and whether these appraisals have monetary (impact on gross salary, annual bonus, wage increase, promotion) or non-monetary consequences. Descriptive results reveal a positive trend: While only 32% of all employees in the sample were subject to systematic performance appraisals in 2004, this number increased to 39% in 2011.[19] An even stronger increase of more than 50% can be found for appraisals linked to individual bonus payments (from 11% in 2004 to 17% in 2011).

4. The Use of Field Data and Experiments to Evaluate the Effects of Differentiation among Employees

As the previous chapter has shown, subjective performance evaluations are a core element of firms' HR practices. But there is only little empirical evidence on the incentive effects of performance appraisals, as noted by Rynes et al. (2005), who state in their survey that "although there is a voluminous psychological literature on performance evaluation, surprisingly little of this research examines the consequences of linking pay to evaluated performance in work settings." We have also seen that there are important trade-offs in the design of appraisals. In particular, whether driving supervisors to assign more differentiated ratings is indeed beneficial for performance represents a very important question for the design of appraisal systems in practice. In the following, we discuss different aspects of differentiation in subjective evaluations using our own exemplary studies applying different methods, comprising an industry-wide field study, an "insider econometrics approach" combining personnel and survey data, a laboratory experiment, the use of linked employer-employee data, and a field experiment.

[18] See also Grund and Sliwka (2009) for an earlier study on a related issue using cross-sectional data from the SOEP.

[19] Note that this is not the percentage of establishments using structured performance appraisals as in the firm-level data presented in the above, but rather the percentage of employees among a representative selection of employees in Germany. The fact that the fraction of employees with performance appraisals is smaller than the fraction of establishments using it is due to several factors. First, as laid out in the above, not all firms use it for all employees. Moreover, the frequency of use is lower in small establishments, and the smallest establishments with less than 50 employees are not part of the LPP survey.

4.1. Incentive Effects of Performance Appraisals: An Industry-wide Field Study

From a typical agency perspective, where a moral hazard problem has to be solved, performance appraisals should be structured in a way that high performance is rewarded and low performance sanctioned adequately. Biases in performance appraisals may therefore weaken the incentive effect, because the relationship between actual efforts and assigned ratings, i.e. the marginal return an employee gets from one unit more effort, decreases with the magnitude of the bias. However, as argued in the above, there may be countervailing behavioral effects as low ratings may also trigger negative reciprocity. In the already mentioned lab experiment by Sebald and Walzl (2014), agents had the opportunity to reduce the principal's pay-off at a cost in return to the feedback provided by the supervisor and actively made use of this option. In firms, negative reciprocity, i.e., punishing a supervisor, might manifest in different ways, for instance, by not providing sufficient effort, being absent from work, disturbing the working climate in the unit, or badmouthing or even sabotaging the supervisor.

Only a small number of studies have so far investigated the relationship between differentiation and indicators of performance in field settings, with nearly all of them looking at single firms. Bol (2011) analyzes a longitudinal sample of 200 employees working in five branch offices of a Dutch financial services firm. She finds a positive relationship between a higher differentiation in performance ratings in around 35 teams per year (i.e., a reduced centrality bias) and subsequent objective performance indicators. Engellandt and Riphahn (2011) also study personnel records from one firm, showing that a higher dispersion in performance ratings is positively associated with a higher performance in the future, proxied by paid and unpaid overtime.[20]

In Kampkötter and Sliwka (2014), we analyzed the impact of differentiation on subsequent performance in an industry-wide field study. We make use of a panel data set on compensation in about 40 German banks provided by an international management consultancy. The data set provides information on fixed salary, short-term performance-related bonus payments, age, firm tenure, hierarchical level (6 levels), functional areas (8 areas), career ladder (management and expert positions), and specific functions (about 60 functions) for the years 2005-2007. As a complementary survey study illustrates, the bonus schemes used in these

[20] Regular overtime work was not remunerated financially but used to substitute for working hours (paid overtime). As employees were not allowed to carry more than 120 overtime hours from one month to the next, those employees having accumulated more than 120 overtime hours provided free labor to the company (unpaid overtime).

banks are frequently so-called bonus pool arrangements, whereby the bank allocates a sum of money to individual units which is then distributed to the employees mainly based on subjective performance evaluations. The size of the bonus pool is typically a function of the financial success of the unit. The key idea of the empirical approach is to estimate the causal effect of differentiation in bonus payments within a unit on the size of the bonus pool in the subsequent year, which should reflect the financial success of this unit. In other words, the question is: does within-unit differentiation in bonus payments affect the success of a unit? To estimate the degree of differentiation, work units are identified by a unique combination of year, company, function, ladder, and hierarchical level. In a next step, the coefficient of variation in bonus payments is calculated for each work unit and year.

In the main specifications, fixed-effects models are estimated, where individual bonus payments in a period t are regressed on the work unit-level measure of differentiation (coefficient of variation) in the previous period t-1. The results show a positive and statistically significant average effect of a within-work unit change in differentiation on subsequent individual performance of employees. To evaluate the economic significance of this incentive effect, the degree of differentiation is divided into quintiles. Moving from a work unit that belongs to the 20% weakest "differentiators" to a work unit that belongs to the top 20% with respect to the degree of differentiation comes along with an increase in subsequent bonus payments by more than 30%. The results are qualitatively robust in instrumental variable regressions to account for the potential endogeneity of changes in the degree of differentiation. In particular, we try to identify factors affecting the dispersion in a unit that are exogenous to this unit's performance. We construct an instrument that measures the average degree of differentiation of other work units within a functional area of the same company and hierarchical level (excluding the work unit we are looking at) for each year. Changes, for instance, in a firm's general evaluation policies and guidelines should affect all departments in a company and therefore be reflected in this instrument. The identifying assumption is that the level of differentiation in other units does not have a direct impact on the bonus payments in a particular area, beyond the influence through the dispersion in the area itself.

Further analyses on subsamples show that the effect is the strongest at the intermediate and highest hierarchical levels. However, the picture changes at the lowest levels, where more differentiation is even associated with lower subsequent average bonus payments. Moreover, there are differences among functions and evidence in line with the idea that differentiation works better in functions where performance evaluation is less subjective (such as retail

banking). It is argued in a formal model that a lack of willingness to differentiate is more detrimental with more objective evaluations, where the potential loss in achievable extrinsic incentives is the largest. To better understand potential drivers of detrimental effects of differentiation, it is thus useful to dig deeper into the behavioral processes underlying the link between appraisals and employee behavior.

4.2. Forced Distribution and Performance: A Lab Experiment

Berger et al. (2013) analyze the impact of a forced distribution system on the differentiation of performance ratings in a controlled lab experiment. Of course, as stated in the above, we are sympathetic to the view that researchers should be cautious when deriving direct practical implications, as a lab experiment can never fully capture the richness of a real world employment relationship. However, lab experiments are uniquely suited to disentangle behavioral mechanisms by intentional design of different treatment variations and to measure specific individual reactions.

In the experiment, participants were assigned to fixed groups of 3 "workers" and 1 "supervisor." Workers had to work on a real effort task for 8 rounds in the main part of the experiment.[21] The timing in each round was as follows: First, workers had to perform a tedious real-effort task. Afterwards, workers and the supervisor learned the performance of all group members. The supervisors of each group then rated each worker on a 1-5 scale (with 1 being the best grade and 5 the worst) and workers were privately informed about their own rating at the end of each round. Performance ratings determined expected bonus payments paid to workers.[22]

In the main treatments of the experiment, bonuses were not paid by supervisors (as is common in most firms, where supervisors are typically not the owners of the firm). The supervisors' own payoff was a linear function of the workers' performance on the task, such that supervisors had some interest in evaluating workers in a way that increases performance. But of course, they may have also directly cared about the well-being of the participants in the role of workers.

---

[21] Prior to the treatment intervention all subjects had to work on the same task for a piece rate in order to obtain a measure for their ability. Workers were matched based on this ability measure in order to have homogenous groups.

[22] In the experiment bonuses were awarded for each period with a bonus of €10 for the highest rating and €0 for the lowest. One of the rounds was randomly drawn at the end of the experiment and paid out.

The main experiment consisted of two treatments. In the baseline treatment supervisors faced no restriction on how to assign ratings to their workers. This is compared to a forced distribution treatment, where it became mandatory for the supervisor to rate one worker with a grade of 1 or 2, one worker with a 3 and one worker with a grade of 4 or 5. Rating distributions in the baseline treatment show strong evidence for rater leniency, as more than 80% of all workers received a 1 or a 2, whereas less than 10% of all workers were given a 4 or 5.[23]

Interestingly, as a post-experimental elicitation of preferences shows, more altruistic subjects assign more lenient and more equity-oriented subjects assign less differentiated ratings, which is well in line with the idea that social preferences affect rating behavior. Moreover, it shows that heterogeneity in supervisors' types is an important element that has to be considered when appraisal systems are designed.

The main result of the experiment is that the forced distribution raised group output significantly by about 6% in the main experiment to 12% in a set of treatments where supervisors had to share the costs of the bonus payments. Further analyses of direct effort reactions to grading reveal that leniency indeed reduced performance. Hence, potential effects of positive reciprocity were apparently dominated by extrinsic incentive effects: When workers realized that they could earn high bonuses even without working harder, they apparently reduced their efforts. This was not possible in the forced distribution treatment, in which there was ongoing competition for the high grades.

But the forced distribution also had detrimental effects, as shown in a further treatment variation. In a third treatment, subjects had access to a technology where they could anonymously block their co-workers' computer screen for 20 seconds. This sabotage effort was costly because subjects' own screens were also blocked for 3 seconds. In this treatment group, output is significantly lower when a forced distribution is employed. Hence, when cooperation among employees is important (or sabotage is easy), a culture of a higher (forced) differentiation may lead colleagues to become competitors, creating negative side effects. This yields one potential explanation for the observation in Kampkötter and Sliwka (2014) that differentiation can be harmful at lower hierarchical levels. At these levels, employees in similar jobs are often direct colleagues who share offices. Hence, in these jobs competition may harm more than it helps to foster performance.

[23] Implementing the forced distribution of course reduced leniency, but supervisors still had the discretion whether to give one worker a 1 or a 2 and one worker a 4 or a 5. The vast majority of workers were rated 1 in the first case and 4 in the latter.

4.3. Evaluations and Reference Point Violations: Combining Personnel Records and Survey Data

As laid out above, social preferences are important drivers of human behavior and thus should affect subjective evaluations. Moreover, the perception of and reaction to subjective evaluations should be affected by reference standards, such as employees' prior expectations and social comparisons to the outcomes of others. Ockenfels et al. (2015) study this in detail, investigating the bonus and appraisal scheme for managers of a multinational company.

The study combines detailed data on performance evaluations from personnel records with survey data on employee perceptions. The panel data set on performance evaluations comprises information on compensation and bonus payments of all (several thousand) managers of the company for Germany (2004-2006) and the United States (2004-2007). This data is supplemented by an anonymous employee survey among managers, eliciting, for instance, their job satisfaction, which is then matched to the appraisal data on the individual level. The study thus follows an "insider econometrics" approach (see e.g., Bartel et al. (2004)) combining the econometric analysis of firm data with detailed institutional knowledge about the firm and survey evidence. The employee survey was administered to German managers in autumn 2007 and among managers in the United States in summer 2008.

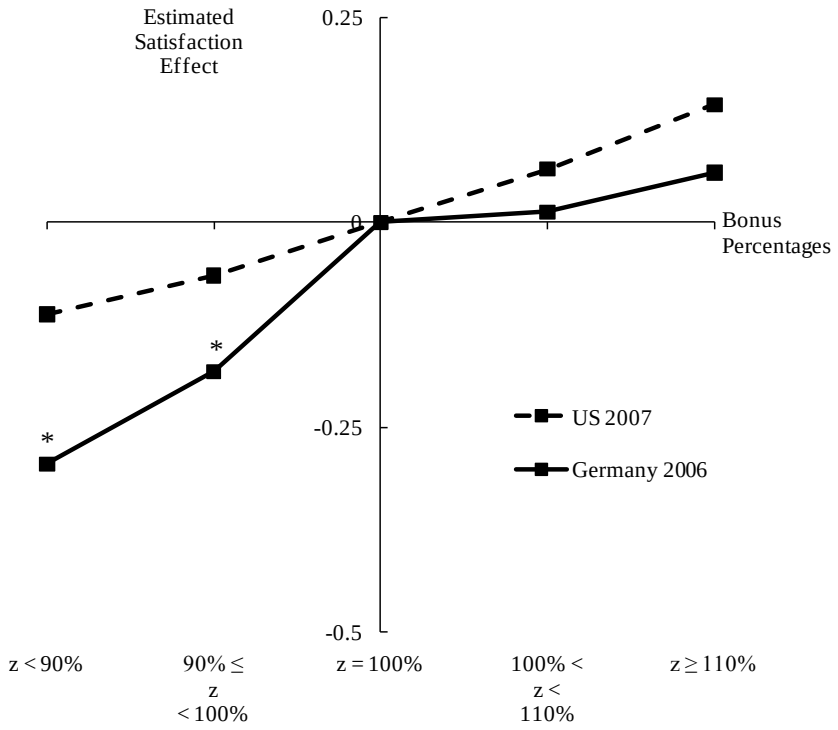
The bonus and appraisal system in this company was organized in the following way: Each year, managers were rated by their respective supervisors on a 5-point rating scale. Several weeks later, supervisors had to assign bonus payments to the managers. For each manager an individual "bonus budget" was determined, which depended on the manager's salary grade as well as the performance of the company and the respective division. Hence, when having the same salary grade, all managers in a unit also had the same bonus budget. Supervisors then determined individual bonuses, subject to the constraints that (i) the sum of bonuses did not exceed the sum of bonus budgets and (ii) bonus payments reflected the previously assigned performance grades. For instance, a manager with a rating of "Fully meets expectations" had to receive a bonus between 80% and 110% of the budget assigned for him. Better-rated managers had to be assigned at least 110%, and worse-rated managers less than 80% of the budget. While the rules were otherwise identical in Germany and the US, there was one key difference: In Germany, managers learned not only the amount of their bonus but also the payout percentage, i.e., what percentage of the budget allocated to them they actually received. In the U.S., managers were only informed about the absolute amount of the bonus and were not told the payout percentage. A key difference between the system in Germany and the US is thus

that in Germany managers directly could compare their bonus to (i) the average bonus in their team and (ii) their own prior expectations based on the assigned performance grade. It is now argued that a payout percentage of 100% is an important reference standard. Falling behind the 100% implies, for instance, that a manager received less than the average of her team. When inequity aversion plays a role, this should be accompanied by a utility loss beyond the monetary consequences.

The relationship between job satisfaction and absolute bonus payments and payout percentages is then analyzed. Figure 4 illustrates the key result. It shows the coefficients of a simple OLS regression with a unit normal transformation of the satisfaction score for Germany and the US, including dummies for intervals of the bonus percentages. The reference group consists of managers who receive exactly 100%. Hence, the graph normalizes satisfaction at the level of managers with 100% bonus in each country and displays the satisfaction effect of managers with other bonus percentages relative to this benchmark. In the German sample, both interval dummies below 100% are significantly smaller than zero. Both interval dummies above 100% are positive but statistically insignificant. In the US, none of the dummies is statistically different from zero.

Hence, bonus payments below 100% reduce employee satisfaction in Germany, where the system creates a salient comparison standard, but not in the US. Employees seem to use the target bonus of 100% as a reference point, and negative deviations from this reference point have a stronger impact on their well-being than positive deviations. Reference point violations here most likely have such a strong effect as a bonus below the 100% leads to both a violation of expectations but also of a social reference standard, as they reveal that a manager gets less than her colleagues.

Figure 4: Bonus Percentages and Employee Satisfaction



Source: Ockenfels et al. (2015)

Further analyses detect indications of a negative performance effect: supervisors who create more reference point violations among their subordinates themselves attain a lower performance rating in the subsequent year.[24] A complementary lab experiment replicating qualitative features of the studied environment shows that salient reference point violations trigger negative reciprocal reactions towards supervisors.

The tension between potential positive incentive effects of differentiation described in the previous sections and the potentially negative effect as differentiation may frequently come along with reference point violations is an interesting point to discuss. One insight is that fine-grained differentiation can be detrimental when performance evaluation is subjective. In the discussed study, the negative effects were basically driven by managers who were all rated as "fully meets expectations" but some received a bonus at 100% of their budget, while others, for instance, received only 96%. In monetary terms these are small differences, but they have a substantial effect on well-being. Here it is quite likely that a rather affective negative reaction

[24] Note that this is of course a subjective rating in itself and not an objective measure of performance.

to reference point violations may outweigh potential positive effort effects through higher-powered incentives because the latter effect is weak. Indeed, a further analysis of the data finds no indication that differentiation in the broader grades (i.e., the 1-5 performance ratings) is detrimental. It apparently is the differentiation within a grade, i.e., of managers with very similar performance levels, that is problematic for satisfaction and, in turn, performance.

4.4. Objective Performance Measurement: A Field Experiment

As already argued in section 4.1, objective performance information may help to facilitate differentiation and foster incentives. Manthei and Sliwka (2014) investigate a field experiment on the benefits of objective performance measurement in performance evaluations. A retail bank in Germany conducted the field experiment in 2003 in order to evaluate the causal impact of the use of objective performance measures on financial performance.

The bank had employed a bonus scheme for the employees in its retail branches based on quarterly financial targets. If the target was met, a branch manager had to allocate a bonus pool among the employees in the branch. Prior to the intervention, branch managers had no access to information on the sales made by individual employees. Hence, managers distributed individual bonuses based on subjective performance assessments. From July 2003 until December 2003, managers in a treatment group of 23 branches gained access to objective sales performance measures for each of their employees across different product categories, which was announced two months before the intervention. Nothing was changed in the control group of the remaining (more than 250) branches, and the rules of the bonus system remained also otherwise unchanged.

The analysis of the experiment reveals a causal effect of having objective performance measures on branch performance. The intervention increased the number of employee-initiated customer appointments by 11% after it was announced. It raised profits by about 2% on average and 5% in the largest branches, even though the intervention came at no costs for the bank. Interestingly, the intervention had no effect in smaller branches, which is in line with the idea that it is easier for supervisors in small branches to keep track of employee performance, even when no objective performance information is available. In larger branches, however, this is more difficult and here the accessibility of objective performance information had a

significantly larger effect on incentives.[25] Hence, the field experiment shows that providing objective performance information can indeed be beneficial, in particular when it is hard for supervisors to keep track of all employees.

4.5. Differentiation and Employee Perceptions: Descriptive Field Evidence

But of course, in many jobs it is simply infeasible or prohibitively costly to access objective performance indicators. In these settings, firms need to rely on subjective evaluations. The question then remains whether firms should foster differentiation. As we have seen above, there are trade-offs involved. Differentiation seems to help when employees work separately, but it may be detrimental when employees can easily harm each other without being observed. Indeed, when Yahoo recently introduced a forced distribution, many articles in the press complained about this change (example headlines are “Forced Ranking Is As Bad For Yahoo As It Was For Microsoft” (Forbes), “Yahoo's Latest HR Disaster: Ranking Workers on a Curve” (businessweek.com) or “Yahoo is ranking employees. When Microsoft did that, it was a disaster” (washingtonpost.com)). Hence, an interesting question is to see whether employee satisfaction is indeed lower in firms that foster differentiation.

We study this question with data from the first wave of the Linked Personnel Panel (LPP) described in section 3, where we observe whether a firm employs a recommended distribution for performance evaluations or not (we do not observe whether firms use a forced distribution which is very rare in Germany). As this is purely cross-sectional data, we caution that we cannot identify a causal effect by using such an instrument here. However, we can answer the question of whether, ceteris paribus, the use of this practice is a credible signal of lower employee satisfaction.

In the following, we analyze the relationship between the use of recommended distributions in performance evaluations and employee perceptions, such as job satisfaction and fairness preferences, by making use of the matched employer-employee character of the LPP. We estimate individual-level regressions, with different employee perceptions and attitudes as dependent variables. The first item, job satisfaction, is measured by the item “How satisfied are

[25] A detailed analysis of the performance effects also reveals substantial differences between product categories that also explain the branch size effects to some extent. There is less separation of labor in smaller branches which causes multitasking problems. Products where performance was not well measured before the intervention benefited also from a shift in efforts from the core product (consumer loans) where performance actually decreased in smaller branches. The effects are robust, for instance, when individual branches are taken out of the sample or size cut-offs are varied.

you with your job?" on an 11-point Likert scale. Affective commitment is measured with the six-item short-form introduced by Meyer et al. (1993). Work engagement is operationalized with the nine-item short scale of the Utrecht Work Engagement Scale (Schaufeli and Bakker, 2004). Helping and cooperation are reflected by two items measuring how often an employee offers help to her coworkers and how often coworkers themselves offer help in case this is needed. Finally, fairness of compensation is measured by a single item that reflects whether the employee perceives her compensation in the establishment as fair.

The main independent variable is a firm-level dummy variable indicating whether the employing establishment uses recommended distributions in their performance appraisal process. We control for establishment size, monthly net salary (in thousand euros), gender, part-time work, type of job (white-collar or blue collar, supervisory position or not), contract type (short- or long-term), age, highest educational attainment, highest professional qualification, as well as industry and region fixed effects. Robust standard errors clustered on the establishment level are reported in all regressions.

Table 1: Recommended Distribution and Employee Outcomes

	(1) Job satisfaction	(2) Commitment	(3) Work engagement	(4) Helping	(5) Fair compensation
Recommended distribution	0.0551	0.0569	0.0098	0.0606**	0.1208**
	(0.0787)	(0.0425)	(0.0343)	(0.0275)	(0.0563)
Monthly net salary	0.2301***	0.1486***	0.0408**	0.0117	0.279***
	(0.0419)	(0.0260)	(0.0207)	(0.0148)	(0.0527)
Constant	6.774***	3.041***	3.481***	4.268***	2.781***
	(0.1731)	(0.0898)	(0.0763)	(0.0610)	(0.1407)
Observations	3,627	3,586	3,521	3,617	3,623
R-squared	0.038	0.109	0.053	0.015	0.124

Additional control variables: Dummies for female, part-time, white-collar, short-term contract, management position, age, highest educational attainment, highest professional qualification, establishment size, industry, and region. Robust standard errors clustered on establishment level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 1 shows that recommended distributions do not come along with a reduced job satisfaction, commitment, and work engagement. The coefficients are even positive but statistically insignificant. Surprisingly, employees report significantly higher levels of helping behavior and perceive their compensation to be fairer in firms that use recommended distributions, as columns 4 and 5 reveal. Again it is important to note that this should not be

interpreted causally. We believe that the most plausible interpretation for this finding is that firms that are better managed have more professional appraisal systems in place.[26] Guidelines about the distribution of grades are an element of many professional appraisal systems, for instance, as otherwise different supervisors follow different standards in the same firm. The causal links can be further explored when longitudinal data on employee perceptions becomes available. But, we can now already conclude that the use of recommended distributions is not a negative but if anything a positive signal about the perceived quality of work in a firm.

5. Discussion

We argue that it is important to apply the toolbox of different research methods when studying management practices, also in order to help firms to design better practices. Formal economic models help to develop a precise understanding of potential behavioral mechanisms. Laboratory experiments allow for the isolation and disentanglement of these mechanisms in precisely controlled environments. Field experiments in firms help us to estimate the causal impact of instrument use on the performance of firms. And, finally, the use of broad representative surveys among firms and employees gives us more detailed information about the frequency and correlates of its use in real companies and to study the generalizability of the insights gained. Hence, it is important to stress the complementary character of these different approaches.

In our view, the use of these complementary methods is particularly necessary when studying performance appraisals, a core HR practice in most firms, as the behavioral and economic mechanisms involved can be surprisingly intricate and complex. But the presented research also reveals some robust patterns that imply rules of thumb for the design of appraisal systems: Differentiation increases performance when the interdependence between the assessed employees is not too strong, but it may increase incentives for counterproductive behavior, especially when cooperation and team work are important. Too fine-grained differentiation without objective performance information may hurt by violating reference points of employees and negative reciprocal reactions may then outweigh potential positive incentive effects.

[26] Further regressions additionally controlling for proxies of ‚better management' such as the existence of variable payment schemes, personnel development plans, written target agreements, employee feedback talks, and workforce planning show a reduced coefficient for recommended distributions, which supports the argumentation that the use of recommended distributions is rather a signal of good management and the coefficients are not estimates of a causal effect.

Objective performance measurement can help by avoiding “rating biases”. However, if objective performance information is not available and firms have to rely on subjective assessments, they should try to “manage” these assessments. Otherwise each supervisor is guided by her own individual social preferences, which leads to inconsistent evaluation standards across different units of a firm. This may well explain why employees in firms with recommended distributions are not unhappier and even perceive a higher fairness of compensation.

To conclude, we strongly believe that academic research using the presented mix of complementary methods can inform the practitioner’s debate and helps to gain a broader understanding of what drives individual behavior in firms.

References

Abbink, K., and A. Sadrieh (2009), "The pleasure of being nasty," *Economics Letters*, 105(3), 306-308.

Baker G. P., Gibbs, M., and B. Holmstrom (1994a), "The internal economics of the firm: Evidence from personnel data," *Quarterly Journal of Economics*, 109(4), 881-919.

Baker G. P., Gibbs, M., and B. Holmstrom (1994b), "The wage policy of the firm," *Quarterly Journal of Economics*, 109(4), 921-955.

Bandiera, O., Barankay, I., and I. Rasul (2011), "Field Experiments with Firms," *Journal of Economic Perspectives*, 25(3), 63-82.

Bartel, A., Ichniowski, C., and K. Shaw (2004), "Using" insider econometrics" to study productivity," *American Economic Review*, 94(2), 217-223.

Bell, D. (1985), "Disappointment in decision making under uncertainty," *Operations Research*, 33(1), 1-27.

Bellemare, C., and B. Shearer (2009), "Gift giving and worker productivity: Evidence from a firm-level experiment," *Games and Economic Behavior*, 67(1), 233-244.

Bellmann, L., Bender, S., Bossler, M., Broszeit, S., Dickmann, C., Gensicke, M., Gilberg, R., Grunau, P., Kampkötter, P., Laske, K., Mohrenweiser, J., Schröder, H., Schütz, H., Sliwka, D., Steffes, S., Stephani, J., Tschersich, N., and S. Wolter (2015), "LPP - Linked Personnel Panel. Quality of work and economic success: longitudinal study in German establishments (data collection on the first wave)," *FDZ-Methodenreport*, 05/2015.

Bernardin H.J., Cooke D.K., and P. Villanova (2000), "Conscientiousness and agreeableness as predictors of rating leniency," *Journal of Applied Psychology*, 85(2), 232-236.

Berger, J., Harbring, C., and D. Sliwka (2013), "Performance Appraisals and the Impact of Forced Distribution - An Experimental Investigation," *Management Science* 59(1), 54-68.

Black, S. E., and L. M. Lynch (2004), "What's driving the new economy? The benefits of workplace innovation," *Economic Journal*, 114(493), F97-F116.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and J. Roberts (2013), "Does Management Matter? Evidence from India," *The Quarterly Journal of Economics*, 128(1), 1-51.

Bloom N., Genakos C., Sadun R., and J. Van Reenen (2012), "Management Practices Across Firms and Countries," *Academy Of Management Perspectives*, 26(1), 12-33.

Bloom, N., and J. Van Reenen (2007), "Measuring and Explaining Management Practices Across Firms and Countries," *The Quarterly Journal of Economics*, 122(4), 1351-1408.

Bloom, N., and J. Van Reenen (2010), "Why Do Management Practices Differ across Firms and Countries?," *Journal of Economic Perspectives*, 24(1), 203-24.

Bloom, N., and J. van Reenen (2011), "Human resource management and productivity," *Handbook of Labor Economics*, 4, 1697-1767.

Bol, J. C. (2011), "The determinants and performance effects of managers' performance evaluation biases," *The Accounting Review*, 86(5), 1549-1575.

Bolton, G. (1991), "A comparative model of bargaining: Theory and evidence," *American Economic Review*, 81(5), 1096-1136.

Bolton, G., and A. Ockenfels (2000), "ERC: A theory of equity, reciprocity and competition," *American Economic Review*, 90(1), 166-193.

Breuer, K., Nieken, P., and D. Sliwka (2013), "Social ties and subjective performance evaluations: An empirical investigation," *Review of Managerial Science*, 7(2). 141-157.

Camerer, C. F. (2015), "The promise and success of lab-field generalizability in experimental economics: A reply to Levitt and List," In G. R. Fréchette and A. Schotter (Eds.) Handbook of Experimental Economic Methodology, Oxford: Oxford University Press.

Combs, J., Yongmei, L., Hall, A., and D. Ketchen (2006), "How Much Do High Performance Work Practices Matter? A Meta-Analysis of Their Effects on Organizational Performance?," *Personnel Psychology*, 59(3), 501-528.

Deaton, A. (2010), "Instruments, randomization, and learning about development," *Journal of Economic Literature*, 48, 424-455.

Delfgaauw, J., R. Dur, J. Sol, and W. Verbeke (2013), "Tournament incentives in the field: Gender differences in the workplace," *Journal of Labor Economics*, 31(2), 305-326.

Dohmen, T., B. Kriechel, and G. A. Pfann (2004), "Monkey bars and ladders: The importance of lateral and vertical movements in internal labor market careers," *Journal of Population Economics*, 17(2), 193-228.

Dur, R., and J. Tichem (forthcoming). “Altruism and relational incentives in the workplace”, *Journal of Economics and Management Strategy*.

Engellandt, A., and R. T. Riphahn (2011), “Evidence on incentive effects of subjective performance evaluations,” *Industrial & Labor Relations Review*, 64(2), 241-257.

Englmaier, F., Roider, A., and U. Sunde (2012), “The Role of Salience in Performance Schemes: Evidence from a Field Experiment,” IZA Discussion Paper No. 6448.

Englmaier, F., and K. Schüßler (2015), “Complementarities of HRM Practices - A Case for Employing Multiple Methods and Integrating Multiple Fields,” CESifo Working Paper No. 5249.

Falk, A., and J. J. Heckman (2009), “Lab Experiments Are a Major Source of Knowledge in the Social Sciences,” *Science*, 326(5952), 535-538.

Fehr, E., Kirchsteiger, G., and A. Riedl (1993), “Does fairness prevent market clearing? An experimental investigation,” *The Quarterly Journal of Economics*, 108(2), 437-459.

Fehr, E., Gächter, S., and G. Kirchsteiger (1997), “Reciprocity as a contract enforcement device: Experimental evidence,” *Econometrica*, 65(4), 833-860.

Fehr E., and K. M. Schmidt (1999), “A theory of fairness, competition, and cooperation,” *The Quarterly Journal of Economics*, 114(3), 817-868.

Fehr, E., and K. M. Schmidt (2006), “The economics of fairness, reciprocity and altruism - Experimental evidence and new theories,” Handbook of the economics of giving, altruism and reciprocity, 1, 615-691.

Flabbi, L., and A. Ichino (2001), “Productivity, seniority and wages: New evidence from personnel data,” *Labour Economics*, 8(3), 359-387.

Fletcher, C., and R. Williams (1996), “Performance management, job satisfaction and organizational commitment,” *British Journal of Management*, 7(2), 169-179.

Frederiksen, A., Lange, F., and B. Kriechel (2012), “Subjective Performance Evaluations and Employee Careers.,” IZA Discussion Paper No. 6373.

Friebel, G., Heinz, M., Krüger, M., and N. Zubanov (2015), “Team incentives and performance: Evidence from a retail chain,” Working Paper.

Gibbs, M., Merchant, K. A., van der Stede, W.A., and M. E. Vargus (2004), “Determinants and effects of subjectivity in incentives,” *Accounting Review*, 79(2), 409-436.

Giebe, T., and O. Gürtler (2012), “Optimal contracts for lenient supervisors,” *Journal of Economic Behavior & Organization*, 81(2), 403-420.

Grund, C., and J. Przemeck (2012), “Subjective performance appraisal and inequality aversion,” *Applied Economics*, 44(17), 2149-2155.

Grund, C., and D. Sliwka (2009), “The anatomy of performance appraisals in Germany,” *The International Journal of Human Resource Management*, 20(10), 2049-2065.

Halse, N., V. Smeets, and F. Warzynski (2011), “Subjective performance evaluation, compensation, and career dynamics in a global company,” Working Paper No 11-15, Aarhus School of Business, Department of Economics, University of Aarhus.

Heckman, J. (1992), “Randomization and Social Program Evaluation," In Manski, C., and I. Garfinkel, eds., Evaluating Welfare and Training Programs. Cambridge, Mass.: Harvard University Press, 1992, 201-230.

Hossain, T., and J. A. List (2012), “The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations,” *Management Science*, 58(12), 2151-2167.

Huselid, M. A. (1995), “The impact of human resource management practices on turnover, productivity, and corporate financial performance,” *Academy of Management Journal*, 38(3), 635-672.

Huselid, M. A., and B. E. Becker (1996), “Methodological Issues in Cross-Sectional and Panel Estimates of the Human Resource-Firm Performance Link,” *Industrial Relations* 35(3), 400-422.

Imbens, G. W. (2010), “Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009),” *Journal of Economic Literature*, 48(2), 399-423.

Kahneman, D., and A. Tversky (1979), “Prospect theory: An analysis of decision under risk,” *Econometrica*, 47(2), 263-291.

Kampkötter, P. (2015), “Performance Appraisals and Job Satisfaction,” *The International Journal of Human Resource Management*, forthcoming.

Kampkötter, P., and D. Sliwka (2014), “The Role of Differentiation in Performance Evaluations - Theory and Evidence,” University of Cologne Working Paper.

Kampkötter, P., Mohrenweiser, J., Sliwka, D., Steffes, S., and S. Wolter (2015), "The Use of HR Practices and Employee Attitudes: The Linked Personnel Panel", mimeo.

Kane, J.S., Bernardin, H.J., Villanova, P., and J. Peyrefitte (1995), "Stability of rater leniency: Three studies," *Academy of Management Journal*, 38(4), 1036-1051.

Köszegi B., and M. Rabin (2006), "A model of reference-dependent preferences," *The Quarterly Journal of Economics*, 121(4), 1133-1165.

Landy, F.J., and J. Farr (1980), "Performance rating," *Psychological Bulletin*, 87(1), 72-107.

Landy, F.J., and J. Farr (1983), *The measurement of work performance*, Academic Press, New York.

Levitt, S. D., and J. A. List (2007), "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?," *Journal of Economic Perspectives*, 21(2), 153-174.

Loomes, G, and R. Sugden (1986), "Disappointment and dynamic consistency in choice under uncertainty," *Review of Economic Studies*, 53(2), 271-282.

Ludwig, J., Kling, J. R., and S. Mullainathan (2011), "Mechanism Experiments and Policy Evaluations," *The Journal of Economic Perspectives*, 25(3), 7-38.

Manthei, K., and D. Sliwka (2014), "Multitasking and the Benefits of Objective Performance Measurement - Evidence from a Field Experiment," University of Cologne Working Paper.

Meyer J.P., Allen, N.J., and C. A: Smith (1993), "Commitment to organizations and occupations: extension and text of a three-component conceptualization," *Journal of Applied Psychology*, 78(4), 538-551.

Michaels, E., Handfield-Jones, H., & Axelrod, B. (2001). *The war for talent*. Harvard Business Press.

Moers, F. (2005), "Discretion and bias in performance evaluation: The impact of diversity and subjectivity," *Accounting, Organizations & Society*, 30(1), 67-80.

Murphy, K.R., and J. N. Cleveland (1991), *Performance Appraisal: An Organizational Perspective*, Allyn and Bacon, Boston, MA

Murphy, K.J. (1992), *Performance measurement and appraisal: Motivating managers to identify and reward performance*. Burns, W.J Jr, ed. Performance Measurement, Evaluation, and Incentives, Harvard Business School Press, Boston, 37-62.

Murphy, K. R., and J. N. Cleveland (1995), *Understanding Performance Appraisal*, Thousand Oaks, Sage.

Ockenfels, A., Sliwka, D., and P. Werner (2015), "Bonus payments and reference point violations," *Management Science*, forthcoming.

Prendergast, C., and R. H. Topel (1993), "Discretion and bias in performance evaluation," *European Economic Review*, 37(2-3), 355-365.

Prendergast, C., and R. H. Topel (1996), "Favoritism in organizations," *Journal of Political Economy*, 104(5), 958-978.

Prendergast, C. (1999), "The Provision of Incentives in Firms," *Journal of Economic Literature*, 37(1), 7-63.

Rynes, S. L., Gerhart, B., and L. Parks (2005), "Personnel psychology: Performance evaluation and pay for performance," *Annual Review of Psychology*, 56, 571-600.

Schaufeli, W.B., and A. B. Bakker (2004), *Utrecht Work-Engagement Scale*, Manual Version 1.1.

Sebald, A., and M. Walzl (2014), "Subjective Performance Evaluations and Reciprocity in Principal-Agent Relations," *Scandinavian Journal of Economics*, 116(2), 570-590.

Subramony, M. (2009), "A Meta-Analytic Investigation of the Relationship between HRM Bundles and Firm Performance," *Human Resource Management*, 48(5), 745-768.

Takahashi, S., Owan, H., Tsuru, T., and K. Uehara (2014), „Multitasking Incentives and Biases in Subjective Performance Evaluation," Discussion Paper Series A No. 614, Institute of Economic Research, Hitotsubashi University.

Varma, A., Denisi, A., and L. Peters (1996), "Interpersonal affect and performance appraisal: A field study," *Personnel Psychology*, 49(2), 341-360.

Welch, J. (2001), "Jack: Straight from the Gut," New York: Warner Books.

Whitman, D. S., Van Rooy, D. L., and C. Viswesvaran (2010), "Satisfaction, citizenship behaviors, and performance in work units: A meta-analysis of collective construct relations," *Personnel Psychology*, 63(1), 41-81.

Zizzo, D. J., and A. J. Oswald (2001), "Are people willing to pay to reduce others' incomes?," *Annales d'Economie et de Statistique*, No. 63/64, 39-65.