Subjective Performance Evaluations and Employee Careers

Anders Frederiksen (Aarhus University, CCP, ICOA and IZA)
Aarhus University
Department of Business Development and Technology
Birk Center Park 15
DK-7400 Herning
Email: afr@btech.au.dk

Fabian Lange (McGill University, CES-Ifo, and IZA)
McGill University
Department of Economics
Leacock Building, Room 511
Montreal, QC H3A 2T7
Email: fabolange@gmail.com

Ben Kriechel (ROA at Maastricht University and IZA)
Maastricht University, SBE-ROA
P.O. Box 616
6200 MD Maastricht
Email: ben@kriechel.eu

Abstract
Employees who work in complex environments are often evaluated by their supervisors. Data on these evaluations promise to be valuable for analyzing career dynamics and human resources practices. However, existing literature on subjective evaluations is based on data from individual firms. Furthermore, how supervisors evaluate workers and how firms use these evaluations might vary substantially with context, precisely because these evaluations are subjective. Thus, little is known regarding whether findings from single-firm studies generalize to broader settings. We examine personnel data from six large companies and establish how subjective performance ratings correlate with objective career outcomes. We find many similarities across firms in how these ratings correlate with base pay, bonuses, promotions, demotions, separations, quits, and dismissals and cautiously propose these as empirical regularities.

JEL: M5

Keywords: subjective performance ratings, personnel data, employee careers

1. Introduction

A central issue in personnel economics is how firms motivate and screen employees when they have limited information about employee actions and productive characteristics. Performance management systems rely on a diverse range of performance measures to alleviate the problems caused by limited information. Often the most important performance measures are provided by supervisors that evaluate the performance of their subordinates.

Theory suggests that firms can use subjective supervisor evaluations to better align the incentives of employees with their own (Baker, Gibbons, and Murphy, 1994; Prendergast, 1999).[1] Furthermore, Holmström and Milgrom (1991) showed how employees game incentive systems that rely on performance measures that cover only a subset of the tasks required of employees. This makes more comprehensive performance measures such as subjective ratings attractive as they can reduce opportunities for gaming in performance management systems.

While theory suggests that subjective ratings might help alleviate incentive problems, empirically very little is known about how subjective performance ratings are used in human resource practices. The existing empirical literature in personnel economics has focused on "simple" settings where simple measures can readily quantify the overall performance of employees reasonably well. For example, in his famous work on incentive pay at Safelite Glass, Lazear (2000) relied on the number of windshields installed by an employee as a measure of individual performance. Other measures analyzed in the literature include trees planted (Shearer, 2004), fruit picked (Bandiera et al., 2005, 2007), eggs collected the check-out speed of cashiers (Mas and Moretti, 2009), eggs collected (Amodio and Carrasco, 2016), and the sewing speed of textile workers (Hamilton et al., 2003). Oyer (1998) and Larkin (2007) used sales as the performance measure for salespeople.

However, it is difficult to objectively and comprehensively quantify the output of the vast majority of workers in developed economies. Objective measures are generally unavailable for workers who perform many different tasks in frequently changing environments or work in teams or in administrative and cross-divisional functions such as HR, legal, accounting, or finance. Similarly, objective measures are typically not available when employees' actions affect the value of the firm over the longer run. By focusing on simple measures in simple

[1] The studies listed here represent just a small subset of the large field studying incentives and hiring practices in organizations. For more comprehensive reviews of the field of personnel economics, see Prendergast (1999) and Oyer and Schaeffer (2010). For a social-psychological perspective, see Murphy and Cleveland (1995).

settings, the literature provides an incomplete picture of the compensation systems and personnel practices that apply to the majority of workers today.

Since objective performance measures are often inadequate, companies frequently require supervisors to subjectively evaluate their subordinates' performances. There is much work in social psychology, management science, accounting, and personnel psychology on subjective evaluations.[2] However, in economics, empirical research on performance appraisal systems that rely on subjective performance measures is thin,[3] which leads Oyer and Schaefer (2010) to conclude that "there is a great need for more empirical research on the use of implicit contracts and subjective performance evaluation in employment relationships."[4]

To bridge this gap, we study personnel data sets containing subjective performance ratings from six firms.[5] In isolation, each of these data sets has been studied before, but the focus was not typically on performance evaluations. Our main goal is to establish empirical regularities across firm data sets in how subjective performance measures are related to a

[2] Kampkötter and Sliwka (2015) review different empirical approaches for studying the use of subjective evaluations in firms. In the process they reference a broad range of studies from accounting, management science, personnel psychology, and related fields, as well as the economic literature on subjective evaluations.

[3] Notable exceptions include Engellandt and Riphann (2011), who in a very interesting study show that employees exert more effort (measured in overtime hours) when departments allow ratings to vary more flexibly over time.

[4] A similar call for more empirical research on subjective performance evaluations can be found in Prendergast (1999).

[5] These data sets cover all of those in the literature that contain subjective performance evaluation of which we are aware and to which we could gain access. Most notably, the data used in Medoff and Abraham (1980, 1981) can unfortunately not be located anymore. The earliest data set is analyzed by Baker, Gibbs, and Holmström (1993, 1994a,b). Their work inspired important theoretical contributions in personnel economics (e.g., Gibbons and Waldman, 1999, 2006). More recent studies based on this data set include DeVaro and Waldman (2012) and Kahn and Lange (2014). We also use data from Gibbs and Hendricks (2004), who examined the role of formal salary systems. The remaining data sets are from Europe. Flabbi and Ichino (2001) used data from a large Italian bank to replicate and expand on the analysis of Medoff and Abraham (1980, 1981). Dohmen (2004) and Dohmen et al. (2004) analyzed the personnel records from Fokker, a now defunct Dutch aircraft manufacturer. Frederiksen and Takáts (2011) used data from a large European pharmaceutical company to study the mix and hierarchy of incentives. Frederiksen and Takáts (2011) originally did not include subjective performance evaluations, but for our analysis, we obtained a second wave of data that included supervisor ratings. The last of our data sets was used by Frederiksen (2013) to analyze explicit and implicit incentives in a large service sector firm.

wide set of career outcomes, including base salaries, bonus pay, total compensation, demotions, promotions, and separations (sometimes distinguished by dismissals and quits).[6]

The theoretical literature in personnel economics distinguishes between effort and output. One might therefore be tempted to ask whether these performance evaluations are measures of the former or the latter. To take a stand on this issue at this point of the inquiry would, however, be counterproductive for two reasons—one empirical and one theoretical. Empirically, the stated aim of our paper is to investigate whether there are empirical patterns in the relation between performance evaluations and career outcomes in firms that are consistent across firms. Such empirical patterns can be discovered and reported without having to take a stand on the nature of performance evaluations as measures of effort or output. Theoretically, there is a sense in which all signals in the standard Principal-Agent models following Holmström (1979) are measures of effort. For setting incentives, what matters is only that a performance measure correlates with the hidden payoff relevant action of the agent. The principal would, for instance, only use output if in fact it correlates with unobserved effort. It is therefore not possible to meaningfully distinguish between a measure of input or output in the literature following Holmström (1979).

The first part of our analysis focuses on how performance ratings change with experience—a question that Medoff and Abraham (1980, 1981) first took up. They found that subjective performance ratings within job levels declined with experience. This result attracted significant attention because it was difficult to reconcile with standard human capital theory as it ran counter to the common pattern of rising wages with rising experience.[7] In our data, we find that performance ratings increase with experience (within job levels) in some firms, decrease in others, and vary in still others. Hence, we find that the performance-experience relation often deviates from that found by Medoff and Abraham (1980, 1981).[8]

Medoff and Abraham (1980) suggested a cardinal interpretation of subjective performance ratings, allowing them to compare average performance ratings both within and across experience levels. By contrast, Harris and Holmström (1982) and Lazear (1999) have argued

[6] It is of course possible that no such empirical regularities exist. Murphy and Cleveland (1995) emphasize that ratings systems are context specific and that so-called "distal" factors such the organizational culture and the competitive environment shape performance appraisal systems. To the extent that these distal factors vary widely, we might not be able to establish empirical regularities across firms.

[7] See, for example, Harris and Holmstöm (1982) and Gibbons and Waldman (1999).

[8] Medoff and Abraham (1980) also reported that experience profiles in log earnings regressions are not sensitive to including performance ratings. We also find this in our data.

for an ordinal interpretation, in which performance ratings reflect relative performance within more narrowly defined peer groups, such as peer groups defined by experience levels. We propose a methodology that accommodates both a cardinal and ordinal interpretation of performance ratings. When developing this methodology, we exploit the insight that cardinality implies ordinality and that subjective performance ratings are ordered random variables.

In the second, main part of the analysis we establish how performance ratings correlate with objective career outcomes. Our main findings are listed below. These findings are largely consistent across firms, and we therefore cautiously propose them as general empirical regularities.

We find that:

1. Performance scales tend to be very restricted. With only one exception, the companies use either a five- or a six-point scale. The effective scale is restricted further because supervisors are reluctant to give bad ratings; there is clearly a "Lake Wobegon" effect in which everyone is above average. Typically, more than 95 percent of ratings are concentrated on only three values at the upper end of the scale.[9]
2. Experience and firm tenure fail to explain the variation in performance evaluations. Instead, job levels explain a large component of the variation.
3. Without exception, individual performance ratings are highly autocorrelated at short lags. At one lag, the autocorrelations almost always exceed 0.4, typically exceed 0.6, and sometimes exceed 0.8. The autocorrelations decline with longer lags and tend to be between 0.1 and 0.4 after three or four lags. The autocorrelations in performance evaluations are higher for more experienced workers.

[9] Biases in performance ratings are well documented. Bol (2011) empirically investigates the importance of centrality and leniency bias and relate it to employee performance. Kane et al. (1995) document that supervisors differ in how lenient they are and that such biases are unrelated to actual performance differences across supervisors. Frederiksen, Kahn, and Lange (2016) likewise present evidence on leniency bias in one of the firms analyzed in this paper and considers the implications for individual earnings and career outcomes. In his discussion of Merck's performance evaluation system, Murphy (1992) notes that supervisors do not exploit the entirety of the available scale and tend to concentrate their ratings toward the upper end of the rating scale but away from extreme ratings. Murphy also discusses at length how Merck tried to get its supervisors to reveal more information through their performance appraisals and to enhance the incentives in its performance system.

Further, using the panel nature of the data, we can evaluate how pay correlates with past, current, and future performance ratings. Even though these correlation patterns vary somewhat across firms, we find several commonalities:

4. In all our firms, performance evaluations correlate positively with log total compensation, log base pay, and log bonuses.[10] We also find that these correlations increase with experience.
5. Base pay and total compensation tend to correlate more highly with current and past performance evaluations than with future performance evaluations for both younger and older employees.
6. The correlation between bonuses and performance evaluations differs substantially across firms. In some firms, bonuses correlate more highly with current than with past and future performance evaluations. These firms might tie bonuses directly to current performance. In other firms, however, there is little difference in how bonuses correlate with current, past, or future performance ratings.

Performance ratings also play a role in promotion and demotion policies and in the separation of employees from firms:

7. In all firms, promotions correlate positively with performance. Demotions correlate weakly and negatively with performance.
8. Separations correlate negatively with performance. In the two firms where we can distinguish dismissals from quits, we find that both are negatively correlated with performance ratings, and that the correlation between performance and dismissals is larger.

Our analysis of the six firm-level data sets proceeds as follows. In the next section, we introduce the firms and present descriptive statistics on subjective performance evaluations. Section 3 is inspired by Medoff and Abraham (1980, 1981) and considers how subjective performance ratings vary with experience and firm tenure. In Section 4, we propose a methodology that allows for further analysis of how subjective performance measures correlate with career outcomes and that is robust to different assumptions about whether performance measures are ordinal or cardinal. In Section 5, we analyze the autocorrelation patterns of performance ratings. In Section 6, we establish how performance ratings are related to earnings and their components (base pay and bonuses). Sections 7 and 8 address the

[10] Throughout the remainder of the paper we will refer to logarithms when using terms such as “base pay,” “bonuses,” and “total compensation.”

importance of subjective performance evaluations for employee mobility both internally (promotions and demotions) and externally (quits and dismissals). Section 9 connects our findings to some existing theoretical models in personnel economics, and Section 10 concludes.

2. The Firms

We analyzed personnel data from six large and very different companies. Either we or other researchers have analyzed data from these companies before, though typically the prior studies did not focus on performance evaluations. The appendix presents the firms in more detail, describes their personnel policies, and summarizes prior research conducted on the data from these firms.

With the exception of Fokker, we cannot reveal the identities of the firms. We therefore substitute with the names of the original research teams who used the data. We thus refer to the companies as Baker-Gibbs-Holmström (BGH), Gibbs-Hendricks (GH), Flabbi-Ichino (FI), Frederiksen-Takáts (FT), Frederiksen (F), and Fokker.

The six companies operate in different countries and industries, and our data covers different time periods (Figure 1). BGH and GH are based in the United States, and FI, FT, F, and Fokker are in Europe.[11] The companies span several sectors. BGH and F are in the service sector.[12] FI operates in the financial sector. FT is a pharmaceutical company, and Fokker was an aircraft manufacturer. For confidentiality reasons, we cannot reveal the industry GH belongs to. The BGH data cover the period from 1969 to 1988 and thus provide the earliest data available. FI, GH, and Fokker data span from the late 1980s to the mid-1990s. The most recent data, from FT and F, span the early 2000s to 2014. The companies' data cover only white-collar workers, with the exception of Fokker and FT, which have data on both blue- and white-collar workers.

[Figure 1]

Table 1 presents descriptive statistics. In this table, and throughout the article, we report all monetary values in 2000-prices US dollars. All six firms have more than 10,000 employees. For BGH, we have 56,000 person-year observations and 10,000 unique individuals. Because

[11] FI is located in Italy and Fokker operated out of the Netherlands until it went out of business in 1996. FT and F are still in operation and for this reason their precise locations and identities are unavailable.

[12] We are restricted from revealing the exact sector.

we only have information on managerial workers in this firm, the average salary is as high as $80,000. For GH we have information on more than 14,000 individuals and a total of 44,000 person-year observations. The data contain information on all employees, and the average salary is close to $60,000. Fokker is special in the sense that white-collar and blue-collar workers are governed by very different policies, and for this reason we conduct the analysis separately for these two groups. There are more than 10,000 blue-collar workers and around 4,000 white-collar workers. The average salary for blue-collar workers is $22,000 and for white-collar workers it is $40,000. For FI we have information on all non-managers, and those 13,000 employees earn, on average, $29,000. The most recent data are from FT and F with each containing around 20,000 unique individuals. For these firms we have information on both managers and non-managers. In FT, 10.7 percent of employees are managers, and they earn, on average, $71,000. The non-managers earn an average of $43,000. In F, 10.2 percent are managers with an average salary of $86,000. Non-managers have earnings comparable to those in FT ($45,000). In FT there are 65,000 person-year observations, and in F there are 149,000 person-year observations.

[Table 1]

2.1. Subjective Performance Measures

While performance evaluation processes vary across companies, wellknown textbooks such as *Compensation* by Milkovich et al. (2011) do describe some generic aspects. A typical review process follows a yearly cycle. At the beginning of the year, performance targets are formulated, but these can be reviewed and adjusted during the year. By the end of the year, performance is assessed by the immediate manager using either a performance scale or a performance matrix.[13] Our information about the actual process in the different firms is limited and we therefore focus on an empirical description of how performance ratings are distributed and on the relation between performance ratings and career outcomes.

Table 2 contains information on the performance scales used by the companies. With the exception of GH, the scale of the performance measures and their distributions are very similar. Most common is a five-point scale, with 1 corresponding to a low rating and 5 to a high rating. There are, however, slight variations in the scales used. For instance, Fokker applied a five-point scale for its white-collar workers and a six-point scale for its blue-collar

[13] A performance matrix combines an evaluation of both "performance over the year" and "future potential." This ranking is aggregated into an overall performance rating.

workers. The only firm applying a substantially different scale is GH, which uses a 27-point scale. However, of these 27 points, only 6 are used in significant numbers.

[Table 2]

In all firms, performance ratings are concentrated on a subset of the scale. Ratings are most concentrated for Fokker white-collar workers, where one category accounts for 81 percent of the ratings. For the other firms, typically all but 3 percent to 4 percent of ratings are concentrated in only three categories. It is also apparent that managers rarely give employees the lowest ratings.

The empirical distributions of the performance ratings may reflect "centrality bias," where supervisors are reluctant to give ratings that deviate from a particular norm, or "leniency bias," meaning that supervisors overstate their subordinates' performance. Murphy and Cleveland (1995, p. 245f) note that few organizations take pains to reward accurate reporting. Since raters are in repeat relationships with ratees, they are likely to bear costs from negative ratings. It is thus not surprising that raters inflate ratings.[14] However, it is also conceivable that the concentration of the performance ratings in the upper two or three categories reflects true employee performance. This distribution would be the outcome if low-performing employees continuously leave the firm.

Most employees are subject to performance appraisals each year. In some cases, however, an employee subgroup is exempted from evaluations. For instance, in FT, systematic performance evaluation is relatively new, and during the phase-in period, the company exempted various employee groups. In other companies, newly recruited employees are not evaluated. For example, in F, employees are not evaluated in their first year of employment. It is likely that similar rules are in place in other firms. In any case, the incidence of performance evaluations is not uniform and varies for reasons that are not well understood.[15] In what follows, we treat the incidence of evaluation as exogenous.

3. Performance Ratings Over the Life Cycle: Medoff and Abraham Revisited

In two well-known papers, Medoff and Abraham (1980, 1981) used personnel records containing subjective performance ratings from three different firms to answer the challenge

[14] Moers (2005) documents that specifics of the rating process such as the availability of multiple objective and subjective measures can affect centrality and leniency of ratings.

[15] Halse et al. (2011) study the use of performance measures in a global company and discuss why performance evaluations may differ in terms of quality and prevalence across countries.

raised by Mincer (1974, p. 11) of whether it can be "shown that growth of earnings under seniority provisions is largely independent of productivity growth." In their data, performance measures decline with experience, holding grade level constant. In addition, controlling for performance ratings did not attenuate the observed earnings-experience gradient. Because Medoff and Abraham interpreted the subjective performance measures as cardinal measures of productivity that can be compared across experience levels, they concluded that "the primary finding … appears to be at odds with what would be expected, given the human capital interpretation of the experience-earnings profile" (Medoff and Abraham, 1980, p. 704).

In Tables 3, 4, and 5, we provide evidence on the same question. Table 3 shows that there is no consistent pattern across firms in how mean performance ratings vary with experience, age, and firm tenure. Performance ratings increase with age, firm tenure, and experience in FI, they follow an inverted U-shape in GH, FT, and F, and they decline in BGH. Within Fokker, performance ratings increase for blue-collar workers, whereas among white-collar workers, they are almost perfectly flat.

[Table 3]

Table 4 presents regression analysis results similar to those of Medoff and Abraham (1981). That is, we regress performance ratings on polynomials in experience and firm tenure.[16] We orthogonalize tenure using experience and the other controls. The tenure coefficients can be interpreted as "within experience" effects of firm tenure. As in Table 3, we find that the performance-experience profiles are not consistent across firms. At average experience, performance ratings decline for BGH, FT, and F, and they increase for GH, FI, and for blue-collar workers at Fokker.

[Table 4]

Job-level indicators generally explain a significant fraction of the variation in performance. In BGH, FI, FT, and F, job-level indicators nearly double the R-square. In addition, the estimated performance gradients in experience and firm tenure are typically sensitive to controlling for job levels, which is evident in FI, FT, and F, where controlling for job levels attenuates the effect of experience on performance ratings considerably. However, in all cases R-squares remain low and the standard errors of these regressions are large, indicating

[16] We control for year and education dummies, and gender, and race when appropriate.

substantial variation in performance that does not correlate with experience, tenure, or position in the firm.

In Table 5, we present log earnings regression analogous to Medoff and Abraham (1980, 1981). Medoff and Abraham examined whether log earnings gradients in experience and firm tenure attenuate when performance ratings are included among the controls.[17] Flabbi and Ichino (2001) replicated these regressions for FI.[18] We consider the same specification for log earnings used in those papers. As do Medoff and Abraham (and FI), we find only weak evidence that controlling for performance evaluations reduces the magnitude of the experience and tenure effects on earnings.[19]

[Table 5]

The results presented in Tables 3 and 4 show that experience and firm tenure profiles in performance ratings vary considerably across companies even when controlling for job levels. This implies that the negative association between experience and performance found by Medoff and Abraham (1980) cannot be considered a stylized fact. However, our data confirm the second finding by Medoff and Abraham (1980), that experience profiles in log earnings regressions are unaffected by performance ratings (Table 5).

[17] Medoff and Abraham control for job levels in their regressions.

[18] Barmby and Eberth (2008) also provide evidence on this question using personnel records from a large financial-sector firm in the United Kingdom. In line with earlier findings, they show that the tenure coefficient in an earnings function is unresponsive to whether or not controls for performance ratings are included.

[19] Barmby and Eberth (2008) sketch an argument based on Stevens (2003) that rationalizes the empirical findings by Medoff and Abraham (1980, 1981). They argue that if workers accumulate specific human capital, then match quality and tenure might correlate negatively since specific human capital can sustain worse matches. Hence, when "match quality" is omitted from the earnings equation, the tenure coefficients will be biased downward. Barmby and Eberth then note that performance ratings may serve as a proxy for match quality. Thus, including performance ratings into an earning regression could reduce the bias in the tenure coefficient. Even if performance evaluations pick up human capital effects, including performance evaluations in the earnings regressions might leave the tenure coefficients unchanged since negative bias induced by the correlation between tenure and match quality is removed from the tenure coefficients.

Different promotion, selection, and turnover policies across firms might explain why experience-performance profiles differ across firms. Gibbons and Waldman (1999) argued that performance ratings might be negatively associated with tenure within a job level because high-ability workers are promoted earlier. When high-ability workers are promoted, the "quality" of the pool of non-promoted employees will decline with job-level tenure. Frederiksen and Takáts (2011) show that firm policies regarding promotions, layoffs, and demotions can lead both to a positive and a negative association between performance ratings and experience within job levels. As in Gibbons and Waldman (1999), performance-based promotions will result in a negative association. However, if low-performing employees tend to be laid off or demoted, then the average performance might increase with job-level tenure.

Facing the same question raised by the findings of Medoff and Abraham, Harris and Holmström (1982) argue that performance ratings reflect relative performance within peer groups (for instance, defined by experience and job levels). They write: "younger workers may be rated higher than older ones in a given job not because they performed better absolutely, but because they performed better *for their age*" (p. 326). Edward P. Lazear (1999) likewise stressed the relative nature of performance ratings in his 1998 presidential address at the Society of Labor Economists. He argues that "the essence of tournament theory is that relative performance matters" and that "theory predicts that workers should be judged and rewarded on the basis of their performance relative to others in their comparison group."[20] Thus, if performance ratings reflect relative performance, then it is only natural that experience-performance profiles differ across firms.

4. How to Use Subjective Performance Ratings

The discussion in the previous section left us with two competing interpretations of performance ratings: 1) a cardinal interpretation in which ratings reflect actual employee performance, and 2) an ordinal interpretation in which ratings reflect relative performance within narrowly defined peer groups. In this section, we propose a methodology that accommodates both interpretations.

A first observation is that cardinal rankings imply ordinal rankings, but ordinal rankings need not be cardinal. For this reason, we proceed using the weaker assumption that performance

[20] An additional observation is that specific-human capital may open up contractual arrangements that may distort the link between performance, experience, and pay. See for instance the early work on this issue by Hashimoto (1981).

ratings are ordinal. A second observation is that the categories for the performance ratings typically are coded using terms such “good,” “very good,” or “excellent,” and, therefore we treat them as ordered random variables.

We assume that the latent variable is normally distributed, which implies that performance ratings are “ordered probits.” We also allow the cut-offs of these ordered probits to differ across firms and we allow them to vary by the observable characteristics that define the employees’ peer groups. Hence, by regressing performance ratings on these variables, the residual retains the variation across individuals that is not explained by observable peer-group characteristics or firm membership.

We can now estimate how the underlying normally distributed latent index that describes performance is correlated with latent performance indices in other periods using maximum likelihood. The resulting correlations are called *polychoric* correlations. By imposing normality on log compensation measures, we can also estimate how the continuous compensation variable and the latent normal performance index correlate. These correlations are known as *polyserial* correlations. Below we refer to polyserial correlations when we report correlations of performance measures with compensation measures and we refer to polychoric correlations whenever autocorrelations in ratings are studied. In regressions, we apply the residualized performance ratings directly as covariates.

The methodology requires that we decide which variables to include when defining a peer group. To begin, we define the peer group using detailed experience, education, and year dummies, gender, and race as well as interactions of linear experience and year trends with gender, education, and race. These variables are predetermined in that performance ratings will not cause these variables. We then residualize performance ratings and compensation measures using this set of variables and estimate the polychoric and polyserial correlations.

Ideally, the peer groups should be defined using predetermined variables that are not themselves career outcomes. If we residualize performance measures on endogenously determined variables, we are likely to remove much of the variation in outcomes that we are interested in studying. For instance, if promotions are partially determined by performance evaluations and if a portion of compensation results from promotions, then studying patterns in compensation and performance after residualizing on the basis of the job hierarchy would be misleading. To see this, consider an example: Workers are either “young” or “old” and the firm has two job levels, Level 1 and Level 2. All young workers are recruited into Level 1, and subsequently some of them are promoted to Level 2. The promotion decision is based

entirely on the performance in the first period, and wages in the second period depend only on the job level and noise. In this highly stylized example, wages among the old are caused by and correlate with performance while young. However, once we control for job level, the correlation between wages and performance while young will disappear. Hence, when controlling for the job level, which is endogenous to the performance measures, we remove the variation in career outcomes that is of interest. It appears, however, that these arguments may be more of a theoretical concern. For instance, when using our most complete data set (F), results when we condition on job levels in addition to using the set of predetermined variables are very similar to those obtained when we only condition on predetermined variables.

It is also likely that performance ratings in a given peer group are systematically influenced by supervisors. This would be the case if some supervisors are generous and others are strict. We can address this issue using the (F) data set with its direct information on the supervisor-employee link. However, when we condition on predetermined variables and supervisor fixed effects, the results are qualitatively similar to those when we only condition on predetermined variables. Unfortunately, we are prevented from investigating more generally how important supervisors are since the other data sets lack information on the supervisor-employee link.[21]

The following sections present results based on performance ratings that are residualized on predetermined variables. We begin with autocorrelation patterns in performance ratings in Section 5, and in subsequent sections we investigate how performance ratings and career outcomes correlate.

5. Correlation Patterns in Performance Ratings

In this section, we consider the second moments of performance ratings. Figure 2, panels A–G, show how (residualized) performance ratings correlate for up to six lags.[22] For each firm, we show the correlations for younger workers (1–15 years of experience) and older workers (16–30 years of experience). We calculate these correlations using the unbalanced panels generated by the personnel data sets. We obtain the reported average correlations within the two experience levels by averaging across experience (within the two groups).

[21] See Frederiksen, Kahn, and Lange (2016) for an in-depth analysis of the role of supervisors in F.

[22] For some firms, the data do not allow us to calculate the autocorrelations across six periods.

The autocorrelation patterns in performance measures are quite similar across companies.[23] In all six data sets, the first-order autocorrelations are high. They range between 0.35 and 0.90 for more experienced workers and between 0.35 and 0.70 for younger workers. For all firms and all lags (except for one distant correlation in FT), the correlations are higher among more experienced workers. The age differences in these correlations are relatively small in BGH, GH, FT, and F. Looking across lags, we find that all the autocorrelations are positive (with one exception for the sixth autocorrelation among young white-collar employees at Fokker). Typically, the autocorrelations decay to about 0.2 to 0.3 for the higher-order autocorrelations, but among more experienced blue-collar workers in Fokker and among the more experienced employees in FI, the autocorrelations remain quite high. Thus, overall, we find that the autocorrelation patterns in ratings are very similar across all firms irrespectively of their country, whether we look at blue- or white-collar workers, and when the data was collected.

[Figure 2]

That autocorrelations are positive and large is a robust finding across firms. These autocorrelations are the result of a complex process that involves employees, supervisors, and incentive systems. Lazear (2004) discuss this process in the context of promotions where the "Peter Principle" (a regression to the mean story) is contrasted with Tournament Theory. From this discussion it becomes clear that regression to the mean in the measurement error (the transitory component of performance) itself would not generate the observed high degrees of positive autocorrelation in ratings, but would rather result in negative autocorrelations. Hence, a more plausible explanation for the observed patters is that performance ratings are subject to a significant degree of measurement error that are themselves correlated over time—maybe because employees tend to be supervised by the same supervisors over time (for a detailed discussion, see Kahn and Lange, 2014).

6. Correlations of Performance Ratings with Earnings Components

In this section, we consider how earnings and performance ratings are correlated. We consider total compensation and, to the extent possible, we look separately at bonus pay and

[23] Our results are robust to excluding new hires (tenure ≤2) from the sample.

base pay. We consider both contemporary correlations and how earnings and performance ratings correlate when they are separated by various leads and lags.[24]

For all earnings measures, we correlate the earning measure at *t* with performance ratings obtained in period *t*+*k*, where *k* is allowed to vary between (at most) -5 and +5. This is done for two groups: individuals with 0–15 vs. 16–30 years of experience.

[Figure 3]

In Figure 3A–E, we show how performance ratings correlates with base pay for the five data sets where we can break down total compensation into base pay and bonuses. A consistent finding across firms that is particularly pronounced among experienced workers is that base pay correlates more highly with contemporaneous ratings or ratings obtained in the near past than with future performance ratings. Kahn and Lange (2014) first noticed this pattern in their analysis of the BGH data. We find the same asymmetry in GH, FT, and F and among older workers in FI. Kahn and Lange also emphasize that in BGH, the base pay of older workers correlates more strongly with performance ratings than that of younger workers. We find the same patterns in the other firms (with a few exceptions for GH and FT).

[Figure 4]

We next turn to the correlations between performance ratings and log bonuses (see Figures 4A–E).[25] The observed patterns are quite different from those established for base pay. In FT and F, performance pay and bonuses are more highly correlated with current ratings than with ratings from other periods. This pattern is less pronounced but still discernible in BGH and FI. Only in GH is this pattern absent.[26]

[Figure 5]

---

[24] As discussed above, we report here polyserial correlations between residualized performance and log compensation measures.

[25] Bonuses typically make up 1.5–2.5 percent of total compensation, and the percentage is higher for more experienced workers and for workers at higher-level jobs. This finding is in line with Grund and Kräkel (2012), who study the chemical industry in Germany.

[26] Gibbs et al. (2004) use a sample of managers in car dealerships to study when firms tie bonuses to subjective criteria. In their data, they find evidence that bonuses are used to complement performance pay based on quantitative performance measures and to shield employees against risk in pay.

Finally, Figures 5A–G show how total compensation correlates with performance ratings. In all firms where we could separately study base pay and bonuses, we find that the correlations between total compensation and performance mirror those for base pay and performance. In Fokker, where this distinction was impossible to make, we find large differences between blue- and white-collar workers. Although the patterns for white-collar workers are in line with what we observe in other firms, the correlations for blue-collar workers are unusual. For them, past performance measures correlate less highly with current compensation than do future performance measures. These results are exceptional and can in part be explained by the very strict administrative rules governing pay that Dohmen (2004) described.

There are thus some common patterns in how performance measures correlate with bonus and base pay and total compensation. First, there is a clear tendency toward higher correlations between earnings and performance ratings for older rather than younger workers. For instance, we find that contemporary correlations between log total compensation and performance ratings are high, between 0.15 and 0.40 for more experienced workers, and relatively low, between 0.10 and 0.30 for less experienced workers. Second, in many, but not all, firms we find a step pattern in the correlations of total compensation and base pay across leading and lagging performance ratings. In particular, for older workers, correlations of log total compensation or log base pay with performance measures two or three periods into the past can be 0.05 points higher than the correlations two or three periods into the future. Finally, the step patterns in the correlations of total compensation and base pay with performance are not evident for bonuses. Instead bonuses tend to be more highly correlated with current performance. This provides some support for the hypothesis that bonuses are being used as explicit incentives.

7. Correlations of Performance Ratings with Promotions and Demotions

We next analyze internal employee mobility, specifically, the frequency of promotions and demotions and their relation to performance ratings. Our focus is on yearly transition rates. That is, we compare job levels at time t and t+1 for individuals who are employed by the firm in two consecutive years. When controlling for performance and individual characteristics, we always use information from time t.

Table 6 present statistics on the frequency of promotions and demotions in the different firms.[27] The first row shows how many levels the job hierarchy in the firm consists of and the

[27] We use job levels to construct promotions and demotions in BGH, Fokker, FI, FT, and F. For BGH, Fokker, and FI these job levels are those from the original studies. In BGH and FI, job hierarchies were constructed by

second and third rows contain promotion and demotion probabilities. The promotions frequencies vary substantially across firms, from 2.4 percent to 16 percent. Demotions are less frequent but typically they are more common than in BGH. This is important, because the original work on BGH has shaped the common perception that demotions are infrequent.

The fourth row presents log earnings difference between the current and next highest job level averaged across employees in the firm. We find significant differences across firms. For instance, in FI the average difference is 0.077 and in FT is it as high as 0.428. Row five shows the average earnings increase upon promotion, also known as the promotion premium. Again we find significant differences across firms with promotion premia ranging from 0.023 to 0.073. So, when employees are promoted, earnings increase by 2.3 to 7.3 percent, while average earnings differences between job levels are much higher (7.7 to 42.8 percent.) This finding is consistent with those of Baker, Gibbs, and Holmström (1994a,b) and Frederiksen, Halliday, and Koch (2016), who show that promotion premia are modest but that promotions open up the possibility of further increases in earnings. The result is that earnings differences across job levels by far exceed the promotion premium.

[Table 6]

In Table 6 (lower part), we show the time to first promotion. We restrict the sample to individuals who are recruited within the sample period and who stay with the firm for at least six consecutive years. Again, there is considerable variation across firms. Almost 80 percent of employees in BGH, but only 13.5 percent in F, are promoted within the first five years. For the other firms, the probability of being promoted during the first five years lies within this interval. Promotions are typically much more common within the first two or three years.[28]

---

the original authors (Baker, Gibbs, and Holmström, 1994a; Flabbi and Ichino, 2001). These authors deduced the job hierarchy from the patterns of transitions across jobs within the firm, which suggests caution in interpreting the transitions in this hierarchy. Job levels in Fokker are also those generated by Dohmen (2004) and Dohmen et al. (2004), as they were identified based on information about job transitions, job titles, reporting relations and team composition. GH provides direct measures of promotions and demotions, which we exploit. FT provides direct information on the job structure within the firm and F contains detailed information about pay grades from which we can construct promotions and demotions from the movements within this hierarchy.

[28] The main exception is FI, where a very large fraction of employees is promoted during the fifth year of employment. In general, we find that many patterns in FI point to a system that seems highly regulated and with little individual variation. The large heaping of promotions at particular points in individual careers as well as the lack of demotions and separations (see below) all point to a system that bases promotions and demotions on rules common to all workers. Because of the Italian context, it likely reflects union-based contractual rules.

However, it is also noteworthy that a large fraction is passed over for promotion in the first five years, and this group may never receive a promotion.

We observe so-called fast-track careers in all firms we study: employees who are rapidly promoted a first time tend to be promoted more rapidly a second time compared to those individuals that take a long time to be promoted a first time. These effects are pronounced in all firms we study.

Overall, it is difficult to explain the observed differences in the frequency of promotions and demotions except for the fact that they reflect differences in organizational structure and incentive systems and that they may be a consequence of differences in administrative and reporting practices across firms.[29] Nevertheless, several findings are consistent across firms: (i) there are many more promotions than demotions, (ii) recent hires are more likely to be promoted, (iii) the promotion premium is lower than the earnings gap between job levels, and (iv) fast tracks are common.

In the following we investigate how performance correlates with internal employee mobility. Our first finding is that high performance ratings are associated with greater likelihood of promotion. Table 7 reports partial correlations between (residualized) performance ratings and internal mobility. For all firms, we find positive correlations between performance ratings and promotions. The lowest coefficient (0.051) occurs among Fokker blue-collar workers, but otherwise the correlations are fairly similar and fall between 0.051 and 0.132. Correlations between performance and demotions are negative but very close to zero.

[Table 7]

In Table 8, we further explore the relation between performance and promotions and present odds ratios from regression analyses relating promotions to employee performance during the prior two periods. In all firms, an increase in recent performance significantly raises the odds of a promotion. In GH, Fokker, FI, FT, and F, an increase in performance today raises the promotion probability by between 20 and 134 percent. An even stronger relation is observed in BGH, where the odds ratio is 3.69. Lagged performance is, in general, less important for promotion. In BGH, FI, and Fokker, a test for the odds-ratio being 1 cannot be rejected. In

[29] Fokker provides an example of how promotion and demotions can be affected by the circumstances of the firm itself. After 1993, Fokker entered a period of reorganization with substantially more demotions. Some of these demotions are arbitrary reclassifications of departments within the firm without entailing changes in the job responsibilities or classification according to the union wage contracts.

GH, lagged performance has a negative effect on the probability of promotion, whereas in FT and F, the effect is positive.

[Table 8]

8. Correlations of Performance Ratings with Separations, Quits, and Dismissals

We now examine how performance relates to employee turnover. While we can examine job separations for all firms, only in two firms (F and FT) do we know whether the separation was initiated by the employee (a "quit") or the firm (a "layoff"). For this reason, we examine separation patterns across all companies and complement this with an analysis of quits and layoffs in FT and F.

In Table 9, we present job separation probabilities for the six firms. The separation rates in American firms (10.7 percent and 12.5 percent) exceed those in European firms. The lowest separation rate is in the Italian firm, FI, with just over 2.2 percent. Excepting the period of downsizing that Fokker underwent after 1992, separation rates in the European firms range from 5.9 to 7.6 percent. These separation rates line up with the stereotypical view that European labor markets are characterized by less mobility than the US labor market, and in particular the perception that there is very little labor mobility in Italy.

For FT and F, the majority of separations are classified as quits and in both firms only about 1.8 percent of workers are dismissed every year.

[Table 9]

Table 9 shows that the correlation between separations and job performance is uniformly negative. The correlations are particularly strong in BGH, GH, FT, and F and very weak in FI. In FT and F, where it is possible to disentangle quits from dismissals, both types of exits are negatively correlated with performance, and the correlation between performance and dismissals is stronger.

Table 10 explores in more detail how performance relates to separations. We use the same specification as in Table 8 (except for the change of the dependent variable) and find that higher performance implies a lower separation probability. A test for the odds ratio for lagged performance being 1 cannot be rejected in most firms. Only in Fokker (blue-collar) does lagged performance reduce the exit rate. Estimating similar regressions for quits and dismissals in FT and F reveals a strong negative relationship between performance and dismissals and a negative yet less strong relationship between performance and quits. As for

separations, lagged performance does not play a role in relation to quits and dismissals in FT and F.

[Table 10]

9. Discussion

We have documented at length how subjective performance evaluations relate to individual career outcomes. At this point, it might be useful to informally draw a few connections to existing conceptual and theoretical contributions in personnel economics.

Performance ratings arise in the complex interplay between supervisors, employees, and the incentive system, and they reflect the employee's innate ability, effort choice, but also transitory shocks (or measurement error) in performance. In the context of promotions, Lazear (2004) illustrates this interplay by contrasting the Peter Principle Theory with Tournament Theory and shows that the theories have different implication for how effort will evolve in the window around a promotion. Irrespective of the underlying dynamics, our finding that performance ratings are highly autocorrelated (Figure 2) implies that future theoretical work in this area should strive to fit this robust empirical pattern.

The foundation for the large literature on incentives in personnel economics is Holmström (1979). He analyzed a simple static model of incentive pay in which firms tie compensation to variables that correlate with productivity. Interpreting performance evaluations as one such variable, the model implies that incentive pay will correlate positively with contemporaneous performance evaluations. Interpreting bonuses as incentive pay, we find (Figure 3) exactly this pattern in F, FT, and to a lesser degree FI and BGH (but not GH).

However, it is also clear from our results that there is a dynamic component to the relationship between performance evaluations and compensation. In particular, the correlation patterns between compensation and performance measures across several leads and lags in F, FT, white-collar Fokker, GH, and BGH (Figures 4–5) are consistent with employer learning about unobserved productivity differences. Dynamic incentives and employer learning is central to models of career concerns (see Gibbons and Murphy, 1992), which makes this class of models particularly suitable for further analysis of incentives in firms. That learning also occurs among more experienced workers suggests that models of

career concerns need to be adapted for lifelong employer learning to fully accommodate the patterns observed in the data (see Kahn and Lange, 2014).

The dynamic perspective also proves to be important for the theory of promotion tournaments. In that model, incentives are contingent on the hierarchical structure, the promotion probability, and the wage spread between job levels. While these variables vary quite a bit across firms (Table 6), our results suggest that current performance is much more important than past performance in predicting promotions, demotions, and separations. However, we also find evidence for fast-track careers and find that much of the value of a promotion to an employee lies in subsequent earnings growth rather than the promotion premium. These findings suggest that promotions are related to performance evaluations but also that monetary returns to promotions cannot simply be summarized by the promotion premium.[30]

In summary, a robust feature in all our data is that rating histories matter. Hence, future theoretical work on incentives through pay and promotions need a more dynamic perspective to shed light on the complex mix of static and dynamic incentives.

10. Conclusion

In most employment relationships, objective performance measures are unavailable. For this reason, supervisors are often asked to subjectively evaluate workers' performance. Because personnel data with performance ratings are still rare, very little is known about how employers use the ratings when they sort, select, and create incentives for their employees. This paper has documented how subjective performance ratings are used in six different firms. We hope it will provide an empirical basis that can be used to evaluate, test, and modify theories of employment relationships.

Across six companies, we find many similarities in how performance scales are structured and used and in how performance ratings correlate with total compensation and base pay and bonuses. We also find many similarities in the relation between performance and employee mobility. For example, promotions are always positively correlated with recent performance, whereas demotions and transitions out of the firm are negatively correlated with performance.

[30] In this context it is also important to stress that demotions may not be as rare as Baker, Gibbs, and Holmström (1994a,b) made us believe. This off course complicates matters because future tournament models might have to allow for downward mobility.

We identified several exceptions and idiosyncrasies that likely stem from specific circumstances in the firms. For instance, among blue-collar workers in Fokker, compensation tends to be more highly correlated with future rather than past performance measures. We believe this is a consequence of a set of very stringent rules negotiated with the unions, as described in Dohmen (2004) and Dohmen et al. (2004). Nevertheless, the similarities across firms in their use of performance ratings far outweigh such exceptions.

Past research has raised the concern that the information in subjective performance measures may be limited by collusion (Tirole, 1986), influence costs (Milgrom, 1988), bias (Prendergast and Topel, 1993; MacLeod, 2003), and favoritism (Prendergast and Topel, 1996). Although these concerns are certainly valid, our empirical findings show that performance ratings correlate significantly with career outcomes, and that there are many qualitative similarities across firms in how performance measures and career outcomes correlate, even if there are exceptions. For this reason, we believe that subjective performance measures contain important information about employee performance.

We hope that our empirical work provides an impetus for model testing and theoretical work that examines how firms collect and use information on worker performance in settings where objective performance measures are unavailable. Ideally, such work can explain the similarities we observe across firms but also the factors that determine differences in how firms use performance ratings.

11. References

Amodio, Francesco and Miquel A Martinez Carrasco 2016, "Input Allocation, Workforce Management and Productivity Spillovers: Evidence from Personnel Data, mimeo available at: https://sites.google.com/site/fscoamodio/ .

Baker, G.P., Gibbs, M., Holmström, B., 1993. Hierarchies and compensation: a case study. European Economic Review 37(2–3), 366–378.

_____. 1994a. The internal economics of the firm: evidence from personnel data. Quarterly Journal of Economics 109(4), 881–919.

_____. 1994b. The wage policy of the firm. Quarterly Journal of Economics 109(4), 921–955.

Baker, G.P., Gibbons, R., Murphy, K.J., 1994. Subjective performance measures in optimal incentive contracts. Quarterly Journal of Economics 109(4), 1125–1156.

Bandiera, O., Barankay, I., Rasul, I., 2005. Social preferences and response to incentives: evidence from personnel data. Quarterly Journal of Economics 120(3), 917–962.

_____. 2007. Incentives for managers and inequality among workers: evidence from a firm level experiment. Quarterly Journal of Economics 122(2), 729–773.

Barmby, T., Eberth, B., 2008. Worker turnover and matching—implications for estimating returns to tenure. Economics Letters 101, 137-139.

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

DeVaro, J., Waldman, M., 2012. The signaling role of promotions: further theory and empirical evidence. Journal of Labor Economics 30(1), 91–147.

Dohmen, T., 2004. Performance, seniority, and wages: formal salary systems and individual earnings profiles. Labour Economics 11(6), 741–763.

Dohmen, T., Kriechel, B., Pfann, G.A., 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.

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

Flabbi, L., Ichino, A., 2001. Productivity, seniority and wages: new evidence from personnel data. Labour Economics 8(3), 359–387.

Frederiksen, A., 2013. Incentives and earnings growth. Journal of Economic Behavior and Organization 85(1), 97–107.

Frederiksen, A., Halliday, T., Koch, A.K., 2016. Within- and cross-firm mobility and earnings growth. Industrial and Labor Relations Review 69(2), 320–353.

Frederiksen, A., Kahn, L., Lange, F., 2016. Supervisors and performance management systems. MIMEO. McGill University.

Frederiksen, A., Takáts, E., 2011. Promotions, dismissals and employee selection: theory and evidence. Journal of Law, Economics and Organization 27(1), 159–179.

Gibbons, R., Murphy, K.J., 1992. Optimal incentive contracts in the presence of career concerns: theory and evidence. Journal of Political Economy 100(3), 468–505.

Gibbons, R., Waldman, M., 1999. A theory of wage and promotion dynamics inside firms. Quarterly Journal of Economics 114(4), 1321–1358.

_____. 2006. Enriching a theory of wage and promotion dynamics inside firms. Journal of Labor Economics 27(1), 59–107.

Gibbs, M., 1995. Incentive compensation in a corporate hierarchy. Journal of Accounting and Economics 19(2–3), 247–277.

Gibbs, M., Hendricks, W., 2004. Do formal salary systems really matter? Industrial and Labor Relations Review 58(1), 71–93.

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

Grund, C., Kräkel, M., 2012. Bonus payments, hierarchy levels and tenure: theoretical considerations and empirical evidence. Schmalenbach Business Review 64, 101–124.

Halse, N., Smeets, V., Warzynski, F., 2011. Subjective performance evaluation, compensation, and career dynamics in a global company. Unpublished paper, Aarhus University, Aarhus.

Hamilton, B.H., Nickerson, J.A., Owan, H., 2003. Team incentives and worker heterogeneity: an empirical analysis of the impact of teams on productivity and participation. Journal of Political Economy 111(3), 465–497.

Harris, M., Holmström, B., 1982. A theory of wage dynamics. Review of Economic Studies 49(3), 315–333.

Hashimoto, M., 1981. Firm-specific human capital as a shared investment. American Economic Review 71(3), 475–82.

Holmström, B., 1979. Moral hazard and observability. Bell Journal of Economics 10(1), 74–91.

Holmström, B., Milgrom, P., 1991. Multi-task principal-agent problems: incentive contracts, asset ownership and job design. Journal of Law, Economics and Organization 7(special issue), 24–52.

Kahn, L., Lange, F., 2014. Learning about employee and employer learning: dynamics of performance and wage measures. Review of Economic Studies 81(4), 1575–1613.

Kane, J.S., Bernardin, H.J., Villanova, P., Peyrefitte, J., 1995. Stability of rater leniency: three studies. The Academy of Management Journal 38(4), 1036–1051.

Kampkötter, P., Sliwka, D., 2015. The complementary use of experiments and field data to evaluate management practices: the case of subjective performance evaluations. IZA DP # 9285.

Larkin, I., 2007. The cost of high-powered incentives: employee gaming in enterprise software sales. Working paper. Harvard University.

Lazear, E.P., 1999. Personnel economics: past lessons and future Directions. Journal of Labor Economics 17(2), 199–236.

Lazear, E.P., 2000. Performance pay and productivity. American Economic Review 90(5), 1346–1361.

Lazear, E., 2004. The Peter Principle: a theory of decline. Journal of Political Economy 112(S1), S141-S163.

Mas, A., Moretti, E., 2009. Peers at work. American Economic Review 99(1), 112–145.

MacLeod, W.B. 2003. Optimal contracting with subjective evaluation. American Economic Review 93(1), 216–240.

Medoff, J., Abraham, K., 1980. Experience, performance, and earnings. Quarterly Journal of Economics 85(4), 703–736.

_____. 1981. Are those paid more really more productive? Journal of Human Resources 16(2), 186–216.

Milgrom, P.R., 1988. Employment contracts, influence activities, and efficient organization design. Journal of Political Economy 96(1), 42–60.

Milkovich, G.T., Newman, J.M., Gerhart, B., 2011. Compensation. 10th edition. New York: McGraw-Hill.

Mincer, J., 1974. Schooling, experience and earnings. Cambridge, MA: National Bureau of Economic Research.

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., Cleveland, J.N., 1995. Understanding Performance Appraisal. Thousand Oaks, CA: Sage Publishing.

Murphy, K.J., 1992. Performance measurement and appraisal: motivating managers to identify and reward performance. In: Burns, W.J. Jr. (Ed.). Performance Measurement, Evaluation, and Incentives. Boston, MA: Harvard Business School Press, 37–62.

Oyer, P., 1998. Fiscal year ends and non-linear incentive contracts: the effect of business seasonality. Quarterly Journal of Economics 113(1), 149–185.

Oyer, P., Schaefer, S., 2010. Personnel economics: hiring and incentives. In: Ashenfelter, O., Card, D. (Eds.). The Handbook of Labor Economics. Amsterdam: Elsevier.

Prendergast, C., 1999. The provision of incentives in firms. Journal of Economic Literature 37(1), 7–63.

Prendergast, C., Topel, R.H., 1993. Discretion and bias in performance evaluation. European Economic Review 37, 355–365.

_____. 1996. Favoritism in organizations. Journal of Political Economy 104(5), 958–978.

Shearer, B., 2004. Piece rates, fixed wages and incentives: evidence from a field experiment. Review of Economic Studies 71(2), 513–534.

Stevens, M., 2003. Earnings functions, specific human capital, and job matching: tenure bias is negative. Journal of Labor Economics 21, 783–805.

Tirole, J., 1986. Hierarchies and bureaucracies: on the role of collusion in organizations. Journal of Law, Economics, and Organizations 2(2), 181–214.

Figure 1. Location, Industry, and Time Period

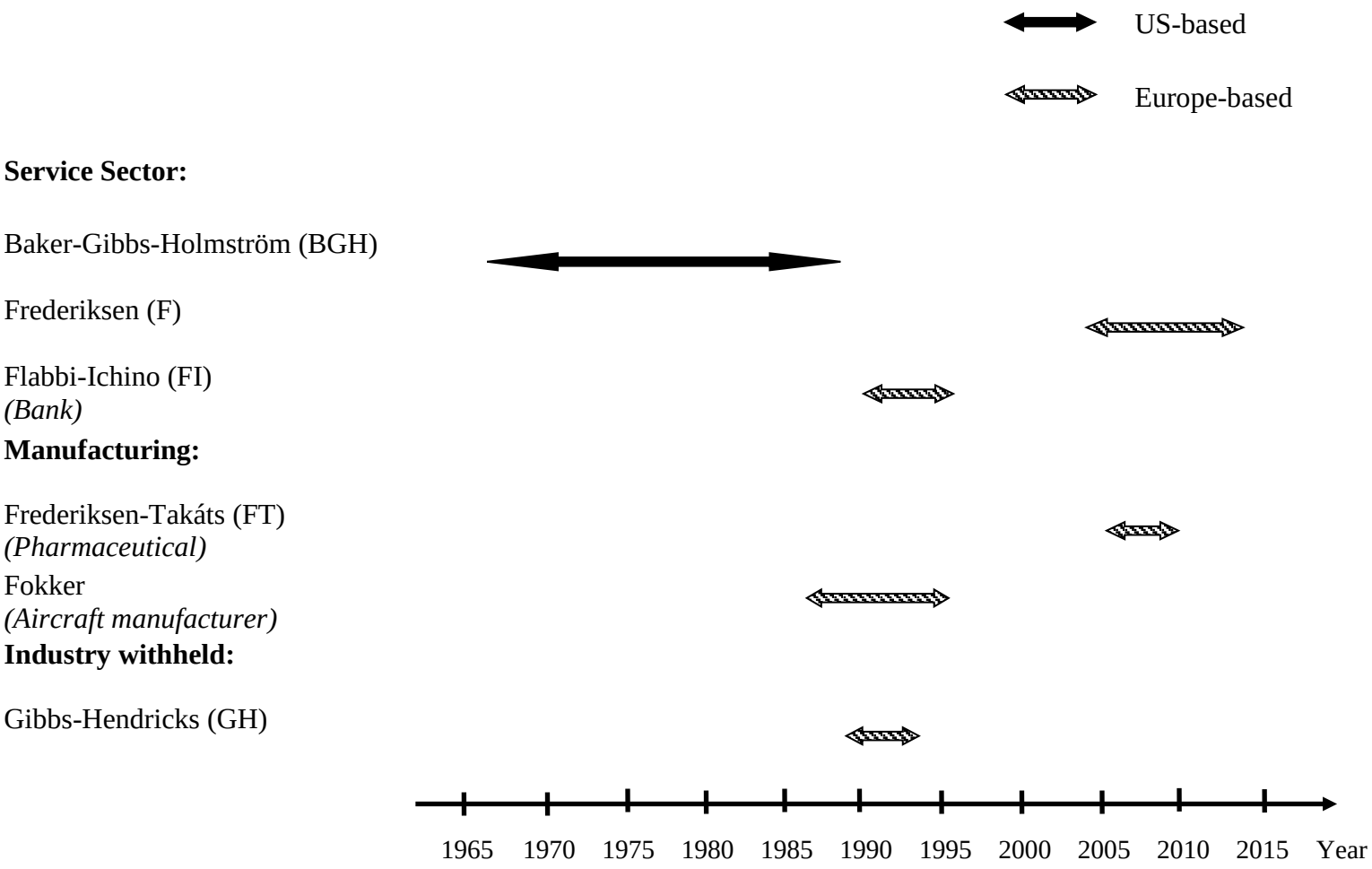


Figure 2. Performance Autocorrelations

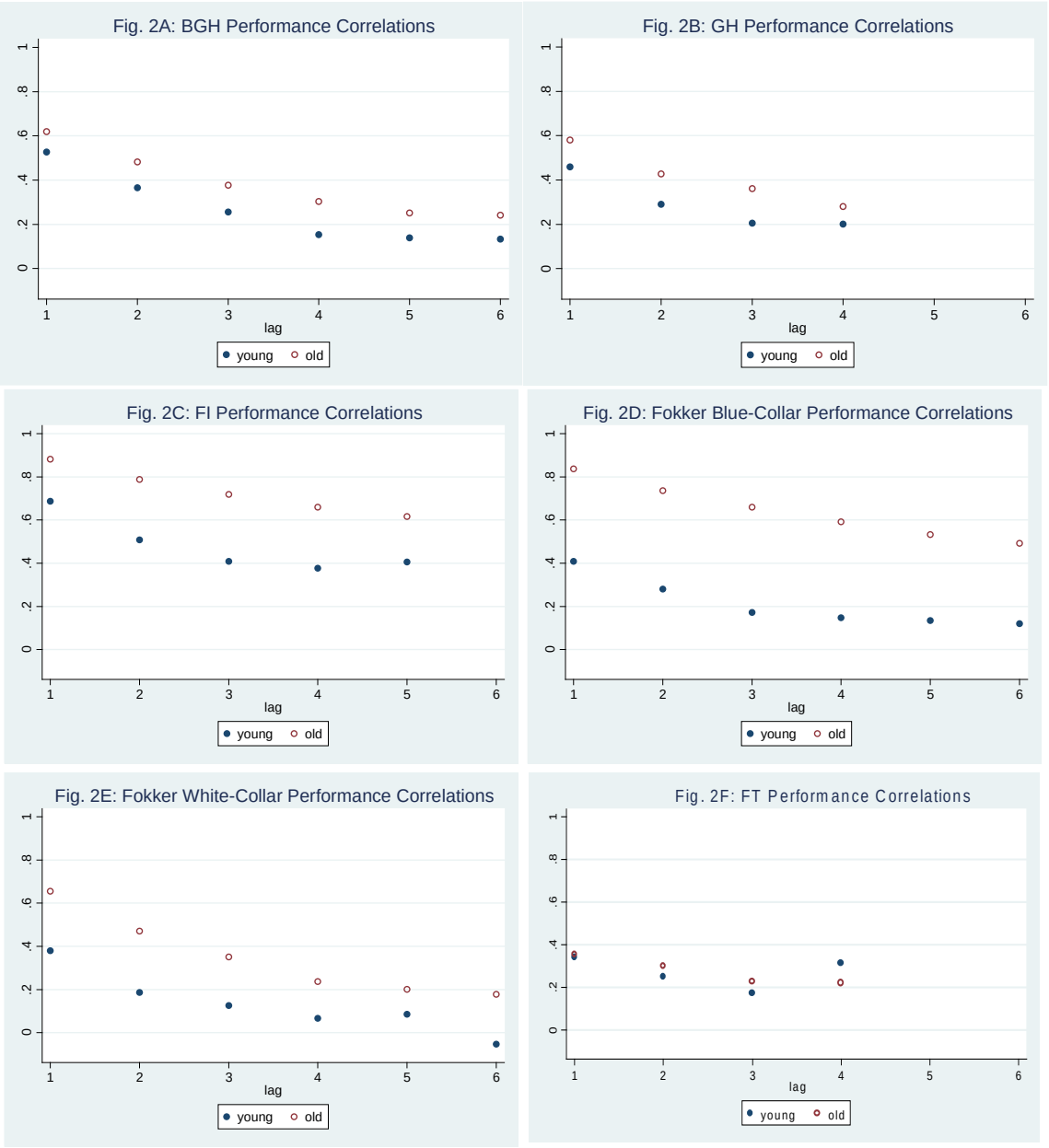


Fig. 2G: F Performance Correlations

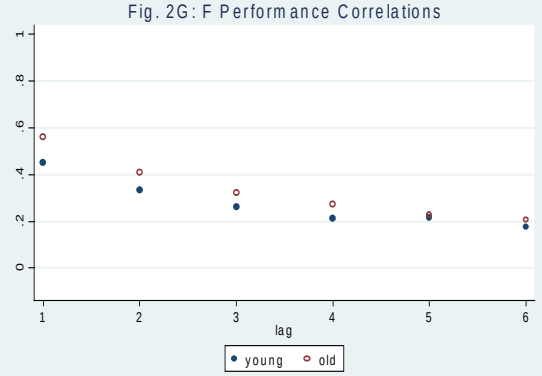


Figure 3. Performance–Base Pay Correlations

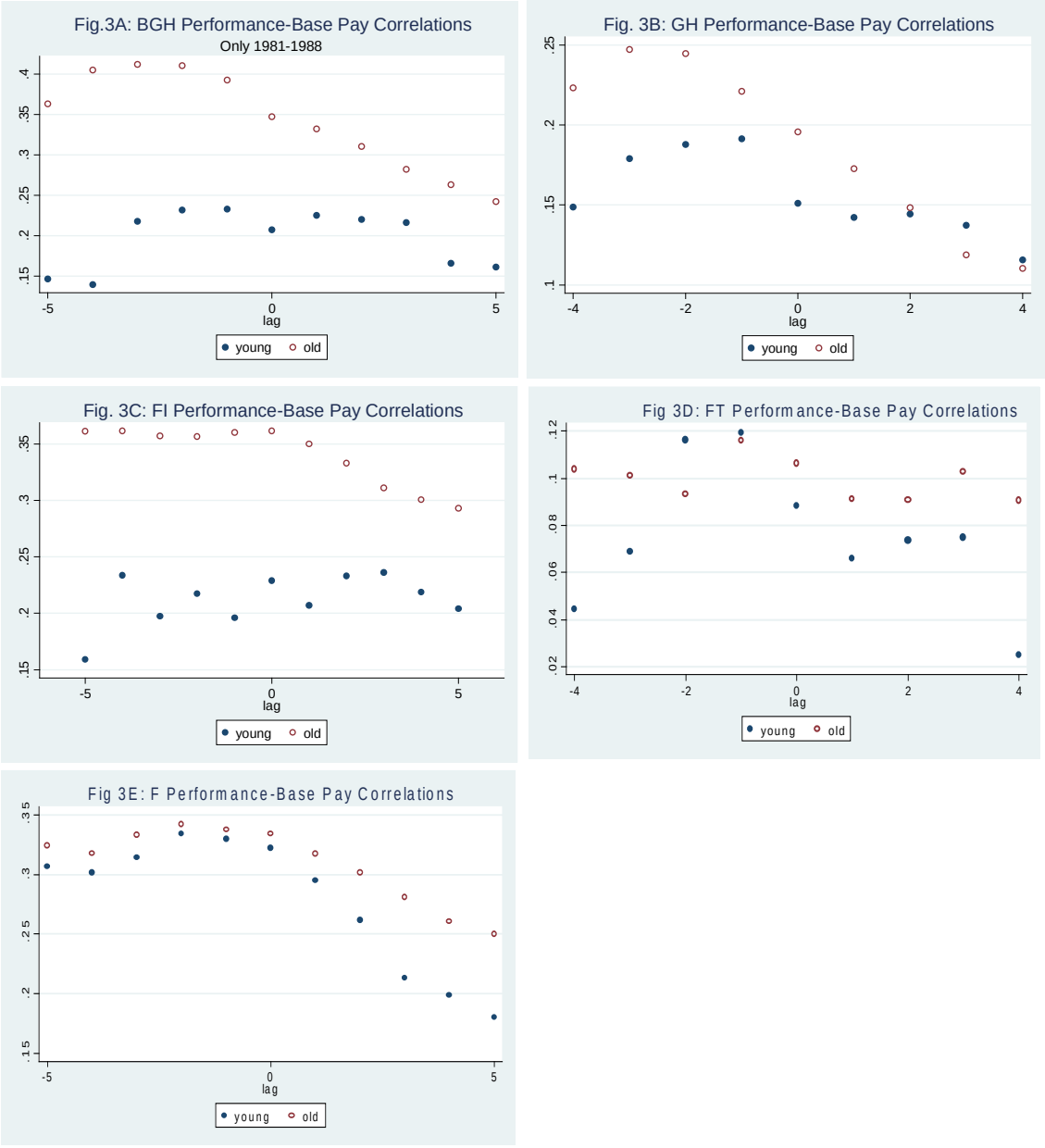
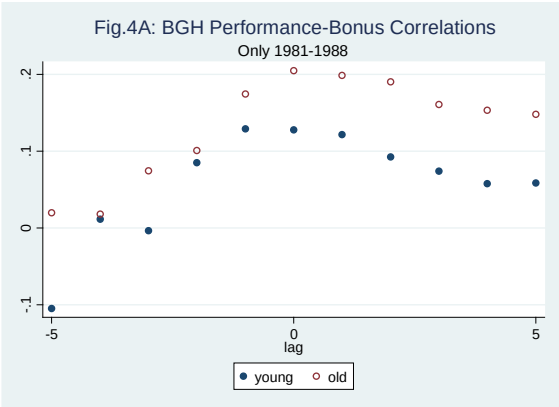
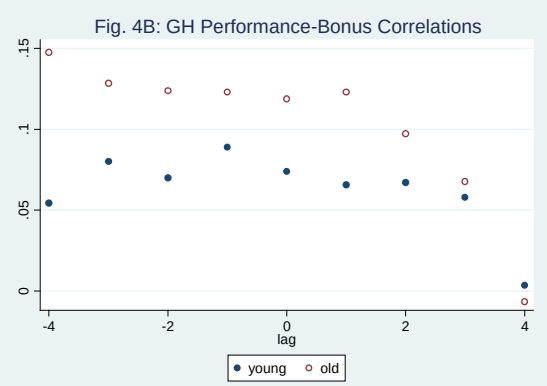
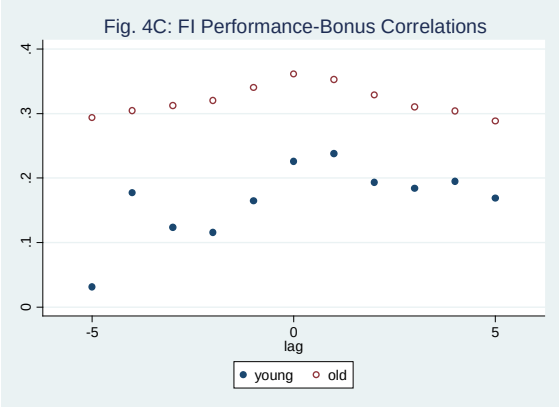
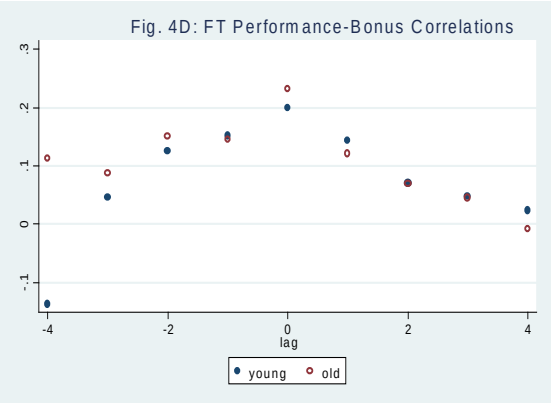


Figure 4. Performance-Bonus Correlations









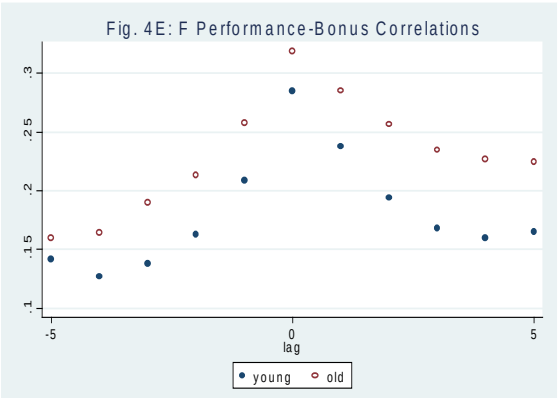
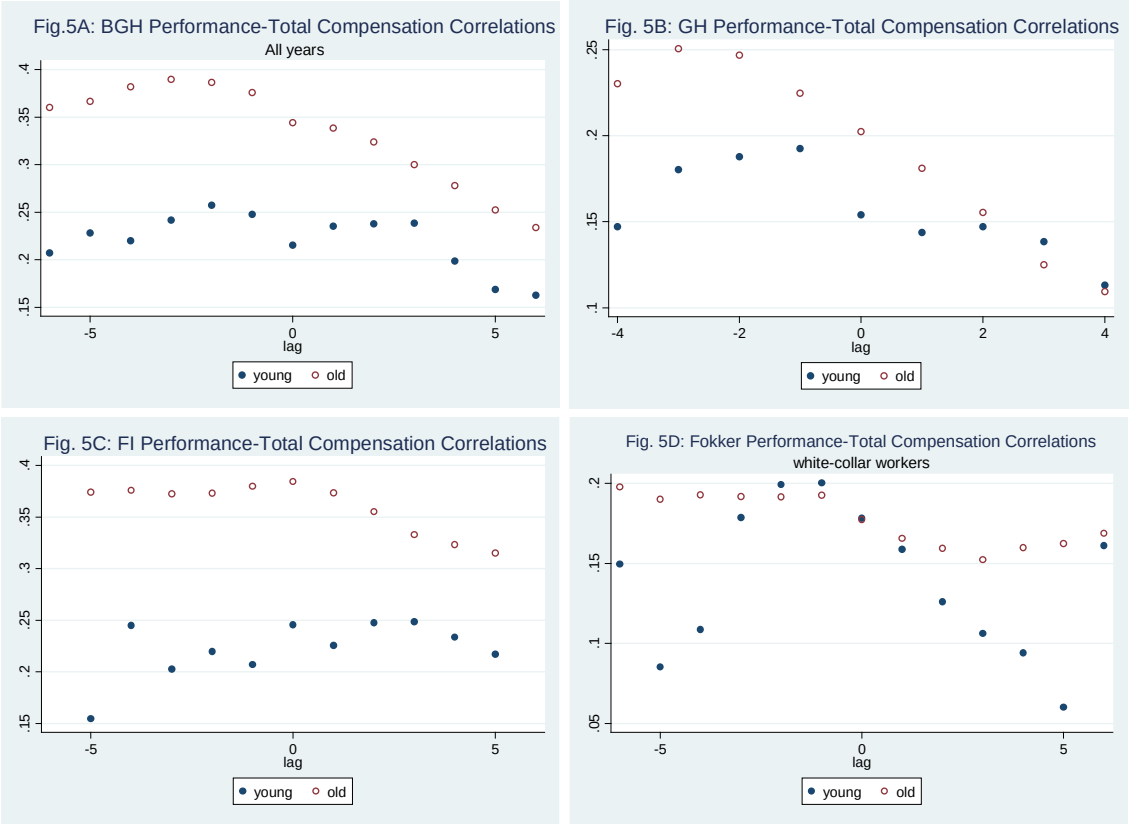


Figure 5. Performance–Total Compensation Correlations



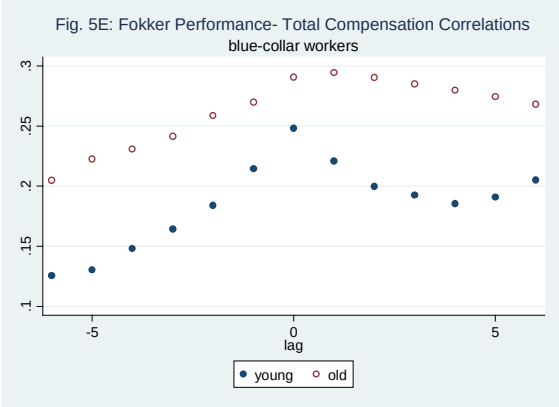


Fig. 5E: Fokker Performance- Total Compensation Correlations
blue-collar workers
.1
.15
.2
.25
.3
-5
0
5
lag
young
old

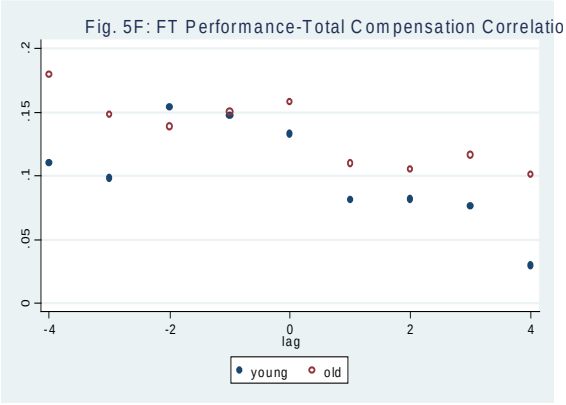


Fig. 5F: FT Performance-Total Compensation Correlatio
0
.05
.1
.15
.2
-4
-2
0
2
4
lag
young
old

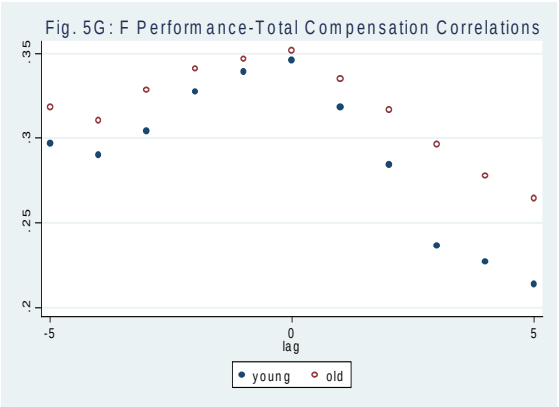


Fig. 5G: F Performance-Total Compensation Correlations
.2
.25
.3
.35
-5
0
5
lag
young
old

Table 1. Descriptive Statistics

	BGH[3]	GH	Fokker Blue Collar	Fokker White Collar	FI[4]	FT	F
Unique employees	9,747	14,372	11,516	4,102	12,996	17,933	23,532
Observations	55,754	43,964	71,086	25,771	63,390	64,976	148,565
Observations with performance ratings	36,428	36,337	70,851	25,731	62,428	23,442	97,299
Fraction managers	Only Managers	Breakdown not clear	Na	Na	Only Non-Managers	0.107	0.102
Compensation[1,2]							
All employees	Na	57,943 (37,055)	21,800 (4,103)	40,086 (12,851)	Na	45,550 (25,691)	48,825 (33,366)
Managers	80,069 (43,536)	Na	Na	Na	Na	70,921 (41,741)	85,869 (70,760)
Non-managers	Na	Na	Na	Na	29,128 (5,462)	42,566 (21,245)	44,618 (22,305)

[1] Averages (with standard deviations in parentheses) obtained using workers with fewer than 40 years of labor market experience.
[2] All earnings are in US$ (2000). US data are deflated using the CPI-U. For the other data sets, we use appropriate deflation indices and convert to US$ using December 31, 2000, exchange rates.
[3] The BGH data contains only managerial employees, composing about 20 percent of the total workforce. In GH and FI, the distinction between managerial and non-managerial employees is not clear from the information provided.
[4] FI data are available from 1975–1995, but performance data are only available from 1990. The statistics reported are based on the period 1990–1995.

Table 2. Distribution of Subjective Performance Measures

		BGH	GH[1]	Fokker Blue Collar	Fokker White Collar	FI	FT	F
Rating scale		**1–5**	**27 levels, but 93% on 6 levels**	**1–6**	**1–5**	**2–6**	**1–5**	**1–5**
Low	**1**	0.05	25	0.12	0.23	Na	0.06	0.13
	2	0.74	18	1.35	3.96	0.06	2.60	3.09
	3	17.05	4	43.83	81.33	2.59	50.73	50.84
	4	50.00	16	40.53	14.13	14.37	39.72	40.17
	5	32.16	24	12.70	0.35	38.01	6.89	5.77
High	**6**	Na	6	1.48	Na	44.97	Na	Na

[1] GH applies a 2–15 point scale with some finer sub-gradation. However, six levels account for 93 percent of the ratings. For GH, only the rates pertaining to the six most common ratings are included.

Table 3. Average Performance by Age, Experience, and Firm Tenure

	BGH	GH	Fokker Blue Collar	Fokker White Collar	FI	FT	F
Rating scale	**1–5**	**2–15**	**1–6**	**1–5**	**2–6**	**1–5**	**1–5**
Age (years):							
≤ 30	4.35	8.86	3.42	3.09	4.74	3.43	3.35
	(0.64)	(1.82)	(0.59)	(0.37)	(0.76)	(0.64)	(0.62)
31–40	4.20	9.26	3.79	3.10	5.26	3.54	3.54
	(0.69)	(1.91)	(0.76)	(0.463)	(0.75)	(0.67)	(0.68)
41–50	4.02	9.24	4.00	3.12	5.44	3.52	3.54
	(0.73)	(1.96)	(0.83)	(0.49)	(0.74)	(0.67)	(0.66)
51+	3.90	9.13	4.29	3.11	5.58	3.44	3.43
	(0.72)	(1.93)	(0.91)	(0.51)	(0.70)	(0.66)	(0.64)
Experience (years):							
1–10	4.33	8.98	3.38	3.10	4.76	3.48	3.38
	(0.66)	(1.84)	(0.57)	(0.37)	(0.74)	(0.66)	(0.63)
11–20	4.17	9.26	3.69	3.10	5.22	3.53	3.55
	(0.69)	(1.94)	(0.73)	(0.42)	(0.77)	(0.67)	(0.69)
21–30	4.00	9.20	3.97	3.11	5.43	3.54	3.54
	(0.73)	(1.95)	(0.81)	(0.48)	(0.73)	(0.66)	(0.66)
31–40	3.83	9.08	4.24	3.11	5.59	3.49	3.43
	(0.74)	(1.90)	(0.90)	(0.51)	(0.67)	(0.68)	(0.64)
Firm tenure (years):							
0–5	4.18	8.87	3.35	3.14	4.66	3.49	3.38
	(0.70)	(1.85)	(0.57)	(0.50)	(0.74)	(0.66)	(0.66)
6–10	4.05	9.34	3.66	3.11	5.15	3.54	3.52
	(0.71)	(1.92)	(0.70)	(0.46)	(0.75)	(0.67)	(0.67)
11–20	3.97	9.36	3.94	3.12	5.35	3.51	3.58
	(0.77)	(1.95)	(0.77)	(0.43)	(0.75)	(0.66)	(0.67)
21+	Na	9.18	4.38	3.08	5.59	3.42	3.49
		(1.92)	(0.86)	(0.40)	(0.68)	(0.66)	(0.64)

Note: Experience refers to potential experience calculated as: Age minus 6 minus years of education. For BGH, firm tenure is only available for individuals entering the sample after 1969 and the firm-tenure statistics are therefore limited to the sample of those individuals. Standard errors are in parentheses. The rating scale for GH runs from 2 to 15. Supervisors had the option to provide some finer gradations within the 2–15 scale, which, however, they rarely did.

Table 4. Experience and Firm Tenure Profiles of Performance Ratings

	BGH[1]		**GH[2]**		**Fokker Blue Collar**		**Fokker White Collar**		**FI**		**FT**		**F**	
Rating Scale	**1–5**		**2–15**		**1–6**		**1–5**		**2–6**		**1–5**		**1–5**	
Experience	-0.013 (0.002)	-0.035 (0.002)	0.071 (0.004)		0.050 (0.001)	0.050 (0.001)	0.002 (0.001)	-0.005 (0.001)	0.070 (0.002)	0.034 (0.002)	0.018 (0.003)	0.010 (0.003)	0.033 (0.001)	0.009 (0.001)
Experience squared / 100	-0.011 (0.004)	0.028 (0.004)	-0.162 (0.011)		-0.045 (0.003)	-0.045 (0.003)	-0.005 (0.003)	0.006 (0.004)	-0.093 (0.003)	-0.043 (0.004)	-0.047 (0.006)	-0.032 (0.006)	-0.072 (0.002)	-0.020 (0.002)
Orth. firm tenure	-0.034 (0.003)	-0.095 (0.004)	0.101 (0.005)	Na	0.058 (0.001)	0.059 (0.001)	0.012 (0.001)	0.010 (0.001)	0.078 (0.002)	0.052 (0.242)	0.020 (0.002)	0.015 (0.002)	0.013 (0.001)	0.012 (0.001)
Orth. firm tenure squared / 100	0.285 (0.024)	0.489 (0.024)	-0.322 (0.020)		-0.081 (0.004)	-0.082 (0.004)	-0.013 (0.004)	-0.010 (0.004)	-0.157 (0.006)	-0.129 (0.006)	-0.048 (0.007)	-0.034 (0.007)	-0.023 (0.002)	-0.019 (0.002)
Job level controls	NO	YES	NO		NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Experience effect at the mean	-0.016	-0.025	0.019	Na	0.037	0.037	0.000	-0.003	0.033	0.017	-0.114	-0.137	-0.147	-0.055
R-squared	0.09	0.17	0.04		0.23	0.23	0.01	0.02	0.14	0.24	0.02	0.05	0.09	0.16
Reg. std. error	0.68	0.65	1.89	Na	0.65	0.65	0.41	0.41	0.73	0.69	0.66	0.65	0.63	0.60
N	36,290	36,290	36,316		54,761	54,761	20,737	20,737	62,428	62,428	23,442	23,442	93,366	93,366

Note: Experience refers to potential experience defined as: Age minus 6 minus years of schooling. In each column, we residualize firm tenure and firm tenure-squared using all other controls appearing in that regression. Each regression controls for education in a flexible manner, where the exact education controls depend on the data set used. In addition to education, all regressions control for gender and year as well as race dummies when appropriate.

1) In BGH, firm tenure is not available for those already in the firm in 1969. We substituted a value of 0 for the orthogonalized firm tenure measure for those with missing firm tenure.

2) GH does not have data on the hierarchical structure of the firm.

Table 5. Log-Earnings Functions with Pay Grades and Performance Ratings

Panel A: BGH, GH, Fokker												
	BGH[1]			**GH[2]**			**Fokker: Blue Collar**			**Fokker: White Collar**		
Experience	0.037	0.010	0.012	0.049	0.045		0.050	0.046	0.044	0.062	0.039	0.039
	(0.001)	(0.006)	(0.001)	(0.001)	(0.001)		(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
Experience squared / 100	-0.070	-0.020	-0.022	-0.092	-0.085		-0.092	-0.086	-0.084	-0.094	-0.057	-0.058
	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)		(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
Orth. firm tenure	0.054	0.004	-0.001	0.039	0.036		0.013	0.011	0.010	0.015	0.010	0.009
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)		(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
Orth. firm tenure squared / 100	-0.144	0.027	0.011	-0.097	-0.085		-0.019	-0.018	-0.015	-0.003	-0.023	-0.022
	(0.012)	(0.008)	(0.008)	(0.003)	(0.003)		(0.000)	(0.000)	(0.000)	(0.002)	(0.001)	(0.001)
Performance rating:												
1			Omitted		Omitted				Omitted			Omitted
2			-0.001		-0.056				0.010			-0.041
			(0.195)		(0.005)				(0.012)			(0.017)
3			0.091		-0.048				0.030			0.003
			(0.194)		(0.009)				(0.012)			(0.017)
4			0.114		0.063				0.073			0.056
			(0.194)		(0.005)				(0.012)			(0.017)
5			0.165		0.095				0.106			0.115
			(0.194)		(0.005)				(0.012)			(0.022)
6					0.137				0.154			
					(0.008)				(0.013)			
Job-level effects	NO	YES	YES	NO	NO	YES	NO	YES	YES	NO	YES	YES
R-square	0.394	0.737	0.742	0.293	0.626	Na	0.79	0.83	0.84	0.67	0.87	0.88
N	21,474	21,474	21,474	36,316	36,316	Na	54,761	54,761	54,761	20,737	20,737	20,737

Panel B: FI, FT, F

	FI			**FT**			**F**		
Experience	0.016	0.001	0.001	0.081	0.069	0.068	0.038	0.004	0.003
	(0.000)	(0.000)	(0.000)	(0.003)	(0.003)	(0.003)	(0.000)	(0.000)	(0.000)
Experience squared / 100	-0.009	0.010	0.011	-0.132	-0.110	-0.105	-0.081	-0.009	-0.008
	(0.000)	0.000)	(0.000)	(0.006)	(0.006)	(0.006)	(0.001)	(0.001)	(0.001)
Orth. firm tenure	0.025	0.025	0.009	0.096	0.087	0.085	0.000	-0.001	-0.001
	(0.000)	(0.032)	(0.000)	(0.002)	(0.002)	(0.002)	(0.000)	(0.000)	(0.000)
Orth. firm tenure squared / 100	-0.030	-0.009	-0.001	-0.202	-0.180	-0.175	-0.004	-0.000	0.001
	(0.001)	(0.000)	(0.000)	(0.007)	(0.007)	(0.007)	(0.001)	(0.000)	(0.000)
Performance rating:									
1			Omitted			Omitted			Omitted
2			0.115			-0.180			-0.015
			(0.016)			(0.187)			(0.015)
3			0.165			-0.036			-0.000
			(0.016)			(0.185)			(0.015)
4			0.181			0.125			0.049
			(0.016)			(0.185)			(0.015)
5			0.200			0.170			0.173
			(0.016)			(0.186)			(0.015)
6			.						
Grade level controls	NO	YES	YES	NO	YES	YES	NO	YES	YES
R-square	0.622	0.806	0.811	0.478	0.506	0.512	0.301	0.783	0.796
N	61,825	61,825	61,825	23,442	23,442	23,442	93,366	93,366	93,366

Note: Experience refers to potential experience defined as: Age minus 6 minus years of schooling. In each column, we residualize firm tenure and firm tenure-squared using all other controls appearing in that regression. Each regression controls for education in a flexible manner, where the exact education controls depend on the data set used. In addition to education, all regressions control for gender and year as well as race dummies when appropriate.

[1)] BGH uses only the years 1981–1988 where full information on log compensation is available.

[2)] GH does not have information on job levels. The regression with performance ratings includes dummies for all performance ratings available in GH. We report the effects for the six ratings reported in Table 1.

Table 6. Promotions and Demotions

	BGH	GH	Fokker Blue Collar	Fokker White Collar	FI	FT	F
Levels in hierarchy	4	-	3	5	8	8	11
Prob. of promotion	16.0%	7.7%	3.4%	9.2%	12.6%	2.4%	10.2%
Prob. of demotion	0.2%	0.4%	1.1%	2.0%	0.0%	0.8%	1.3%
Avr. log earnings difference across job levels[1]	0.290	-	0.088	0.128	0.077	0.428	0.130
Avr. earnings increase upon promotion[2]	0.031 (0.001)	0.073 (0.006)	0.026 (0.001)	0.023 (0.001)	0.026 (0.002)	0.047 (0.004)	0.025 (0.001)
Time to first promotion (if promoted within the first five years)							
Year 1	31.8%	21.2%	16.1%	25.9%	12.9%	25.9%	31.5%
Year 2	35.1%	27.8%	21.2%	21.9%	10.6%	34.6%	42.4%
Year 3	17.6%	30.3%	22.3%	23.9%	9.0%	14.8%	12.9%
Year 4	9.5%	12.6%	25.9%	22.7%	9.4%	24.7%	7.8%
Year 5	5.9%	8.1%	14.5%	5.7%	58.0%	-	5.4%
Never or later	21.3%	55.0%	77.9%	43.6%	40.5%	87.9%	86.5%

Note: To construct the "time to first promotion," we sample those individuals who are both recruited and stay with the firm for six consecutive years within the sample period. The sample period for FT is five years, and we consider time to promotion within the first four years for this company. Among blue-collar workers in Fokker, we very rarely observe promotions to white-collar jobs. Somewhat more often, but still rare, are demotions of white-collar workers to blue-collar jobs.

[1] This value is calculated as $\sum(\boldsymbol{\beta_{Level+1}} - \boldsymbol{\beta_{Level}}) * \boldsymbol{n_{Level}/N}$, where $\boldsymbol{\beta_{Level}}$ is the job-level dummy for a given job level in a log-earnings regression, $\boldsymbol{n_{Level}}$ is the number of employees observed at this job level, and N is the total number of employees across all years.

[2] Obtained from regressing wage changes on whether or not an individual was promoted in a given period controlling for individual effects, experience profiles interacted with demographics, and education and year effects.

Table 7. Correlations Between Performance Ratings and Internal Mobility

	BGH	GH	Fokker Blue-Collar	Fokker White-Collar	FI	FT	F
Scale	1–5	2–15	1–6	1–5	2–6	1–5	1–5
Performance at *t* and promotion between *t* and *t*+1	0.124	0.060	0.051	0.084	0.062	0.132	0.078
Performance at *t* and demotion between *t* and *t*+1	-0.024	-0.016	-0.016	-0.030	Na	-0.012	-0.044

Note: The reported correlations are based on residualized performance measures.

Table 8. Promotions and Performance (Logit)

Endogenous variable: Promotion between *t* and *t*+1	**BGH**	**GH**	**Fokker Blue Collar**	**Fokker White Collar**	**FI**	**FT**	**F**
Performance at *t*	3.69 (0.19)	1.20 (0.03)	1.44 (0.07)	1.92 (0.13)	1.52 (0.06)	2.34 (0.23)	1.89 (0.05)
Performance at *t*-1	0.94 (0.05)	0.93 (0.02)	1.03 (0.05)	1.08 (0.08)	0.99 (0.05)	1.62 (0.16)	1.13 (0.03)
Pseudo R-squared	0.220	0.082	0.039	0.046	0.103	0.121	0.157
N	13,167	12,417	48,857	17,671	33,339	6,510	68,822

Note: The table reports odds ratios of logistic regressions of promotion between *t* and *t*+1 on residualized performance from time *t* and *t*-1. All regressions control for quadratics in experience and orthogonal firm tenure, together with education, gender, and year dummies, and race when appropriate. Each specification furthermore includes dummy variables for the job levels in *t* and *t*-1.

Table 9. Correlations Between Performance Ratings and Mobility out of the Firm

	BGH	GH	Fokker Blue Collar[1]	Fokker White Collar[1]	FI	FT	F
Scale	**1–5**	**2–15**	**1–6**	**1–5**	**2–6**	**1–5**	**1–5**
Separation rate	10.75%	12.48%	Overall: 9.91% Pre-1991: 6.06% Post-1991: 14.65%	Overall: 8.99% Pre-1991: 6.20% Post-1991: 12.33%	2.23%	6.47%	7.60%
Quit rate						4.77%	5.75%
Dismissal rate						1.70%	1.85%
Correlations							
Performance at *t* and separation between *t* and *t*+1	-0.084	-0.095	Overall: -0.067 Pre-1991: -0.046 Post-1991: -0.088	Overall: -0.055 Pre-1991: -0.049 Post-1991: -0.063	-0.018	-0.071	-0.064
Performance at *t* and quit between *t* and *t*+1	Na	Na	Na	Na	Na	-0.029	-0.046
Performance at *t* and dismissal between *t* and *t*+1	Na	Na	Na	Na	Na	-0.083	-0.051

Notes: The reported correlations are based on residualized performance measures.
[1] Fokker went through several downsizing episodes between 1992 and 1995. We therefore present statistics before, during, and after 1991.

Table 10. Separations and Performance (Logit)

Endogenous variable: Separation between *t* and *t*+1	**BGH**	**GH**	**Fokker Blue Collar**	**Fokker White Collar**	**FI**	**FT**			**F**		
	Sep	Sep	Sep	Sep	Sep	Sep	Quit	Dismissal	Sep	Quit	Dismissal
Performance at *t*	0.63	0.86	0.83	0.74	0.74	0.58	0.73	0.18	0.54	0.61	0.36
	(0.03)	(0.02)	(0.03)	(0.07)	(0.14)	(0.05)	(0.08)	(0.03)	(0.02)	(0.03)	(0.03)
Performance at *t*-1	0.98	0.98	0.80	0.90	0.95	1.07	1.12	0.92	0.93	0.92	0.94
	(0.04)	(0.02)	(0.03)	(0.09)	(0.18)	(0.10)	(0.13)	(0.17)	(0.04)	(0.04)	(0.09)
Pseudo R-squared	0.080	0.032	0.135	0.144	0.150	0.057	0.043	0.154	0.087	0.087	0.100
N	22,041	6,729	34,443	12,957	50,136	6,510	6,510	6,510	68,822	68,822	68,822

Note: The table reports odds ratios of logistic regressions where separations, quits, and dismissals between *t* and *t*+1 are regressed on residualized performance from time *t* and *t*-1. All regressions control for quadratics in experience and orthogonal firm tenure, gender, and year dummies, and race when appropriate. Each specification furthermore includes dummy variables for the job levels in *t* and *t*-1.

Appendix

Here we provide more detail about the firms analyzed in this paper and briefly summarize the prior research on these data.

Baker-Gibbs-Holmström (BGH)

In two groundbreaking papers, Baker, Gibbs, and Holmström (1994a,b) analyzed the personnel data of a US-based service-sector firm. The study focused on managerial employees (about 20% of the workforce) and covered a period when the firm experienced rapid growth in assets and employees. The authors described the internal personnel structure in detail, and looked for the existence of an "internal labor market." They also informally considered whether the data were consistent with models of employer learning, human capital acquisition, and simple incentives. In summarizing the findings of BGH, Gibbs and Hendricks (2004, p. 73) write that BGH "concluded that their evidence was inconsistent with simple models of learning and incentives. Instead, they suggested that many of their findings were consistent with a model in which employees accumulate human capital at varying rates."

BGH did not analyze the use of subjective performance ratings in this firm. That was first attempted by Gibbs (1995). He showed that performance ratings correlated strongly with pay, pay rises, and promotions, but they did not predict exit from the firm. Similar to BGH and based on the same data, Kahn and Lange (2014) re-established that heterogeneous human capital accumulation is important, but by using the information conveyed in the subjective ratings, they also showed that employer learning was taking place at all stages of the employees' careers. That is, employers were trying to "hit a moving target." Another recent paper by DeVaro and Waldman (2012) used the BGH data to test the promotion-signaling hypothesis.

Three peculiarities of BGH are worth mentioning. First, no variable in the original data explicitly identified the job hierarchy. Instead, BGH used the internal mobility patterns and some information on job titles to deduce the hierarchy. In our analysis, we rely on the hierarchy identified by BGH in their original work. Second, we have data on bonuses only from 1981 on. Bonuses make up a small fraction of total compensation and for this reason we use the compensation data from the entire 1969–1988 period for our analysis. When we look specifically at bonuses and base pay, we restrict the data to those years in which the two types of income are available separately. Third, tenure data can only be calculated precisely for

workers entering after 1969, when the sample period starts. Any statistics related to firm tenure that we present are based on those observations for which exact tenure can be determined. By contrast, experience is measured as potential experience (age minus 6 minus years of schooling). We use this measure of experience in the analysis of all data sets.

As shown in Table 1, BGH consists of 55,754 employee-year observations from a total of 9,747 unique employees.[31] Average total compensation (in 2000 dollars) is about $80,000, which far exceeds the average for the US population.[32] This, as well as the demographics and the high education levels of staff, is a reflection of only managerial employees in the data set.

Gibbs-Hendricks (GH)

Our description of GH is based on Gibbs and Hendricks (2004). GH use data on administrative rules governing pay to study the effect of different administrative pay systems (Grade, Hay, and PAQ, as described in GH) on the structure of wages in this firm. Gibbs and Hendricks asked to what extent these administrative rules simply reflected market forces (acting as a "veil"). Their overall conclusion is that the firm did not incur large costs from the nominal constraints imposed by the formal salary rules. This is consistent with the view that the ability to assign employees to different salary ranges combined with the use of bonuses and some discretion in pay suffice to accommodate market forces.

The data cover white-collar professional and managerial employees as well as clerical and technical office workers employed in a large US corporation active in several different businesses from 1989 to 1993. The data do not contain explicit information on the hierarchy, but rather contain indicators for promotions and demotions. GH draws on 43,964 employee-year observations from a total of 14,372 unique employees. The average compensation of $58,000 exceeds the US average.

Gibbs and Hendricks (2004) describe in detail the formal salary policies in place at GH. There are in fact three different compensation systems (Hay, Grade, and PAQ) in place at this firm, but all of these are quite similar in spirit. All three assign pay ranges to different jobs within the firm. The midpoint of these pay ranges is based on points attached to the job

[31] In our analysis of the firms, we only use employees with experience less than 40.

[32] All earnings measures are reported in 2000 dollars equivalents.

describing the tasks in the position.[33] The compensation associated with these points are based on market or historical within-firm data. The salary range for each job is typically between 80 percent and 120 percent of the midpoint. Within these salary ranges, raises are largely based on time in position and performance ratings.

Employees might also share in various bonus pools. Some bonus pools were based on corporate or division performance and only higher-level employees would fully benefit from these. Others, unrelated to corporate or division performance, would be open to more employees. How these bonus pools were split between employees depended, at least in part, on performance ratings.

Fokker

Fokker was a Dutch airplane manufacturer. The company faced financial trouble after 1991 and underwent several rounds of downsizing before finally going bankrupt in 1996. Dohmen (2004) and Dohmen et al. (2004) study the internal hierarchy and pay structure of this firm.

The firm's performance ratings were tied to compensation according to a very strict system of rules and regulations. Further, the data consist of both blue-collar and white-collar workers who were subject to very different personnel regimes. We therefore analyze the blue-collar and white-collar samples separately. If employees are represented in both groups at different points in time, we dropped them from the analysis.

The data span 1987 to 1996. We use 71,086 employee-year observations from 11,516 unique blue-collar workers. The white-collar sample is smaller, with 25,771 employee-year observation and 4,102 unique individuals. Average compensation in this firm was $40,086 for white-collar workers and $21,800 for blue-collar workers.

The internal hierarchy in Fokker was identified by Dohmen (2004) and Dohmen et al. (2004) using job transitions, job titles, team composition, and reporting relationships. The information used to identify the job hierarchy is thus richer than that in BGH, but the procedure itself does rely on job transitions to define the hierarchies. Thus, as in BGH and FI, caution is advised when interpreting movements across the hierarchy as this hierarchy itself is measured based on data on transitions in the firm.

[33] For the jobs covered by the Hay or Grade system, the points are based on Hay Associates, a consulting company specializing in compensation systems. For the more blue-collar type jobs covered by the PAQ system, the points are based on a position-assessment questionnaire.

Flabbi – Ichino (FI)

The company analyzed by Flabbi and Ichino (2001) is a large bank operating throughout Italy. Flabbi and Ichino used the data from this firm to replicate the analysis by Medoff and Abraham (1980, 1981).

As do Flabbi and Ichino, we restrict the sample to males. We also restrict the analysis to non-managerial workers, since subjective performance evaluations were only available for these. The data span 1990 to 1995 and consist of 63,390 employee-year observations from 12,996 unique employees. Reflecting the lower incomes in Italy and the restriction to non-managerial employees, average earnings in the firm are $29,000.

The approach to measuring the hierarchy in the firm follows BGH and a similar note of caution in interpreting the data on movements in the hierarchy applies.

Regarding the performance evaluations, Flabbi and Ichino (2001, p. 365) state that the "supervisors receive detailed instructions on how to rank their subordinates using a four-level scale. These instructions ... involve four possible choices labeled as *low, medium, good,* and *very good*." Flabbi and Ichino also show that these performance evaluations correlate with absences and "misconduct episodes" (Table 3, p. 367).

Frederiksen –Takáts (FT)

The company that Frederiksen and Takáts (2011) analyzed is a global pharmaceutical company headquartered in Europe but with production and sales activities on all continents. Frederiksen and Takáts study the firm's use of incentives and derive a hierarchy of incentives. In particular, they explain why firms often use a complex mix of incentives.

The FT data used in the analysis span 2007 to 2011 and thus constitute some of our most recent data. The data available for analysis comprises employees working in the country where the company's headquarters is located. A total of 64,976 employee-year observations are available for analysis, and these are based on information from 17,933 unique individuals. Average earnings in this firm are $46,000.

The use of a systematic and company-wide performance appraisal system is relatively new to the FT firm, and the sample period overlaps with the phasing-in of the performance management system. Consequently, only a fraction of employees received performance ratings in the early years. However, by the end of the sample period, more than two-thirds of employees received a rating.

In FT the hierarchy is derived from detailed information on the job structure and the data contains all relevant information on compensation and employee mobility. In FT (and in F) dismissals can be distinguished from quits. On that note, Frederiksen and Takáts (2011) state that the institutional settings, in the country where FT is operating, do not restrict layoff policies.

Pay setting in FT is described in some detail in Frederiksen and Takáts (2011). According to the firm, base pay depends on job functions, responsibilities, and competencies and is set to be competitive in the labor market. The firm claims that bonuses are related to well-defined targets known to employees and the bonus pools vary across employee groups. For more details, we refer the reader to the original paper.

Frederiksen (F)

The F firm is a service-sector firm that Frederiksen (2013) analyzed for implicit and explicit incentives. Using a dynamic moral hazard model (a career concerns model), Frederiksen predicted cross-sectional and individual earnings dynamics and established the mechanisms leading to earnings growth. The overall conclusion was that the model performed well in explaining early career earnings dynamics.

The F firm has some international activities, but our data cover only domestic operations. The data comprise more than 23,000 unique employees and almost 150,000 employee-year observations between 2004 and 2014. For the purpose of this study, F constitutes the most complete data set as it contains detailed information on wages, bonuses, performance ratings, and employee mobility, including information on quits and dismissals, over a long period of time. Average earnings in the firm are $48,825.

The performance appraisal system in F is based on a performance matrix where employees are rated on their performance (over the past year) and on their potential (over a two-year time frame). To make our analysis of F comparable to that of the other firms studied in this paper, we only use the information on performance.

Frederiksen (2013) does not explain in detail how pay is determined in F, but we have been able to obtain this information through conversations with the company. Base pay at entry is determined by the job description, the employee's skill level, and the associated pay grade and is to some extend influenced by unions. We use the pay grades (of which there are 11) in our analysis of the organization's hierarchy and mobility along this hierarchy. The company also explains that the bonus pool is determined by executive management and then it is

distributed downward in the organization through predefined channels until it arrives at the appropriate level, where it is distributed across the employees by the manager.

According to Frederiksen (2013), the regulations regarding dismissals are quite weak in the country that F operates in and are comparable to those regulations in place in the US. Frederiksen also writes that the costs associated with dismissals are lower than in most other developed economies. From conversations with the firm we have learned that unions are strong in this country and dismissal policies reflect this. In particular, policies in place tend to move low performers to different jobs with the purpose of a better job fit. Only if this fails will the firm take actions that might lead to a dismissal.