Subjective Performance Evaluations and Employee Careers

by

Anders Frederiksen (ICOA at Aarhus University, CCP and IZA)
ICOA Aarhus University
Fuglesangs Allé 4
DK-8000 Aarhus
Email: afr@asb.dk

Fabian Lange (Yale University, CES-Ifo, and IZA)
Yale University
Department of Economics
37 Hillhouse Ave, Room 8
Email: fabolange@gmail.com

and

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

Abstract

Firms commonly use supervisor ratings to evaluate employees when objective performance measures are unavailable. Doubts persist whether their subjective nature invalidates findings based on subjective performance measures. And, it is unclear whether findings from individual firm data-sets generalize to the labor force at large. We examine personnel data from six large companies and establish how subjective ratings, interpreted as ordinal rankings of employees within narrowly defined peer-groups, correlate with objective career outcomes. We find many similarities across firms in how subjective 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[1]

How firms motivate and select employees when facing limited information about their actions and characteristics is the central issue in personnel economics. In some settings, firms rely on objective measures of employee performance to form personnel policies.[2] However, objective measures are generally not available when workers perform many different tasks in frequently changing environments, when they work in teams or when their actions affect the value of the firm over both the short and long run. In such contexts, supervisors are often required to subjectively evaluate workers' performances. These subjective evaluations are presumably an important tool in creating incentives and screening workers. Unfortunately, empirical research on subjective performance measures is thin, which leads Oyer and Schaefer (2010, p. 11) 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."

A major obstacle in studying the use and consequences of subjective performance ratings is, of course, their subjectivity. Subjective evaluations can be influenced not just by the performance of the worker, but also by the characteristics of the supervisor and by the relationship between the supervisor and the employee. Furthermore, subjective measures are reported on arbitrary metrics. Therefore, abstract concepts such as effort and ability—concepts common in models of moral hazard and adverse selection— may map onto performance ratings in widely varying ways across

[1] This project would not have been possible without the exemplary willingness of a number of researchers to share their data. We thank Michael Gibbs for providing access to both the BGH and the GH data, Gerard Pfann for giving access to the Fokker data, and Lucca Flabbi and Andrea Ichino for allowing us to work on the FI data and sharing their code with us. We greatly appreciate the help we have received by all of these in working with and understanding the data. We are also indebted to the two companies referred to in this paper as FT and F for allowing us to work on their data and in particular to the employees in these companies who have made it possible. We are grateful for the comments and discussions we have had with Nikolaj Andreas Halse and Thomas Bech Hansen and members of the CCP network. We also thank Lisa Kahn for helpful comments. Michael Lykke Jensen has provided competent research assistance on this project.

[2] Over the last decades, a number of studies have examined the use of objective performance measures. In the most famous of these studies, Lazear (2000) explored how Safelite, a windshield repair-company, used the number of windshields replaced as a measure of performance for its repairmen. This landmark paper showed how a change in the incentive system affected both the sorting and selection of workers and their performance on the job. Other objective performance measures that have been investigated are the number of trees planted (Shearer 2004), the amount of fruit picked (Bandiera, Barankay, and Rasul 2005, 2007), the check-out speed of cashiers (Mas and Moretti, 2009) and the sewing speed of textile workers (Hamilton, Nickerson and Owan, 2003). These studies considered how objective performance varies with the pay system, work organization and social interactions. A burgeoning literature in education economics uses value-added measures of student test scores to examine issues in selecting and creating incentives for teachers (Barlevy and Neal 2012; Goldhaber and Hansen 2010).

firms and circumstances. All this makes it difficult to evaluate the standard models in personnel economics using subjective performance ratings. Thus, a natural question to ask is if individual firm data sets containing subjective performance evaluations can be used to test theory and if empirical findings from specific firms generalize to larger populations of employers. The answers to these questions, in part, hinge on whether or not there are empirical regularities across firms in how subjective performance ratings co-move with objective career outcome.

In this paper we study personnel data sets from six firms.[3] These data sets cover all data sets in the literature that contain subjective performance evaluation of which we are aware and that we could gain access to.[4] In isolation, each of these data sets has been studied by us or by other researchers before, but typically the focus was not on performance evaluations. Our main goal is to establish regularities across firm data sets in how subjective performance measures are related to a wide set of career outcomes, including base salaries, bonus pay, total compensation, demotions, promotions, and separations (sometimes distinguished by dismissals and quits).

The first part of our analysis focuses on how performance ratings on average change with experience – a question that was first taken up by Medoff and Abraham (1980, 1981). In our data, we find that performance ratings increase with experience (within job level) in some firms, decrease in some others, and are non-monotone in still others. Medoff and Abraham found that subjective performance ratings within job levels declined with experience. How to interpret these findings depends on whether performance ratings are assumed to be comparable across experience levels or whether they are assumed to be ordinal comparisons ("rankings") of employees within a more narrow peer-group defined, for instance, by the experience level. While Medoff and Abraham favored the first interpretation we favor the second, because of the large variation in performance-

[3] The most prominent dataset we use is the one analyzed by Baker, Gibbs, and Holmström (1993, 1994a,b). These articles inspired important theoretical contributions in personnel economics (e.g., Gibbons and Waldman, 1999, 2006). More recent studies based on this data are Kahn and Lange (2011) and DeVaro and Waldman (2012). We also use data from Gibbs and Hendricks (2004) who examined the role of formal salary systems. While these two datasets are from the United States the remaining datasets 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, Kriechel, and Pfann (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. These data do not include subjective performance evaluations, but for our analysis, we obtained a second wave that included supervisor ratings. The last of our data sets was used by Frederiksen (2010) to analyze explicit and implicit incentives in a large service sector firm.

[4] Most notably, the data used in Medoff and Abraham (1980, 1981) is unfortunately not accessible.

experience gradients that we observe across firms. This variation is hard to explain if one interprets performance scales as cardinal measures unless one assumes that average productivity varies with experience in vastly different ways across firms. By assuming that performance ratings are ordinal rankings of individuals within experience levels, we can reconcile the systematic correlations between career outcomes and ratings that we document below with the large variation in experience-ratings gradients across firms. Thus, our interpretation of performance ratings, which accommodates our empirical findings, is that they reflect ordinal measures of employee performance within narrowly defined peer groups (defined by experience, demographics and education).

A number of patterns in the data are sufficiently consistent across firms that we cautiously propose that these are general empirical regularities.

Regarding the distribution of performance ratings, 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 ranking scale.
2. Experience and tenure fail to explain much of the variation in performance evaluations. Instead, job levels explain a fairly large component of the variation in performance ratings.
3. We find, without exception, that the performance ratings that individuals receive are highly correlated 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 also found to be higher for more experienced workers.

Using the panel nature of the data, we can evaluate how pay components correlated with past, current, and future performance ratings. Even though there is some variation in these correlation patterns across firms, we do find a number of regularities:

4. In all our firms, performance evaluations are positively correlated with log total compensation, with log base pay and with log bonuses.[5] We also find that the correlations of performance evaluations with base pay, with bonuses, and with total compensation increase with experience.
5. Base pay and total compensation tend to correlate more highly with contemporaneous and past performance evaluations than with future performance evaluations. This finding is present among younger and older employees.
6. Bonuses tend to correlate more highly with current than with past and future performance evaluations. These finding could reflect that some firms 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 influence how employees move internally and out of the firm.

7. In all firms, promotions correlate positively and demotions negatively with performance.
8. Transitions out of the firm are negatively correlated 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. We introduce the firms and present descriptive statistics on subjective performance evaluations in the next section. Section 3 is inspired by the work of Medoff and Abraham (1980, 1981) and considers how subjective performance ratings vary with experience and tenure. In Section 4 we propose a methodology that improves comparability of subjective performance measures across different firm-level data-sets. In Sections 5 and 6, we analyze the autocorrelation patterns of performance and pay separately. In Section 7, we establish how total compensation as well as its components – base pay and bonuses – are related to performance ratings. Sections 8 and 9 address the importance of subjective performance evaluations for employee mobility both internally (promotions and demotions) and out of the firm (separations, quits and dismissals).

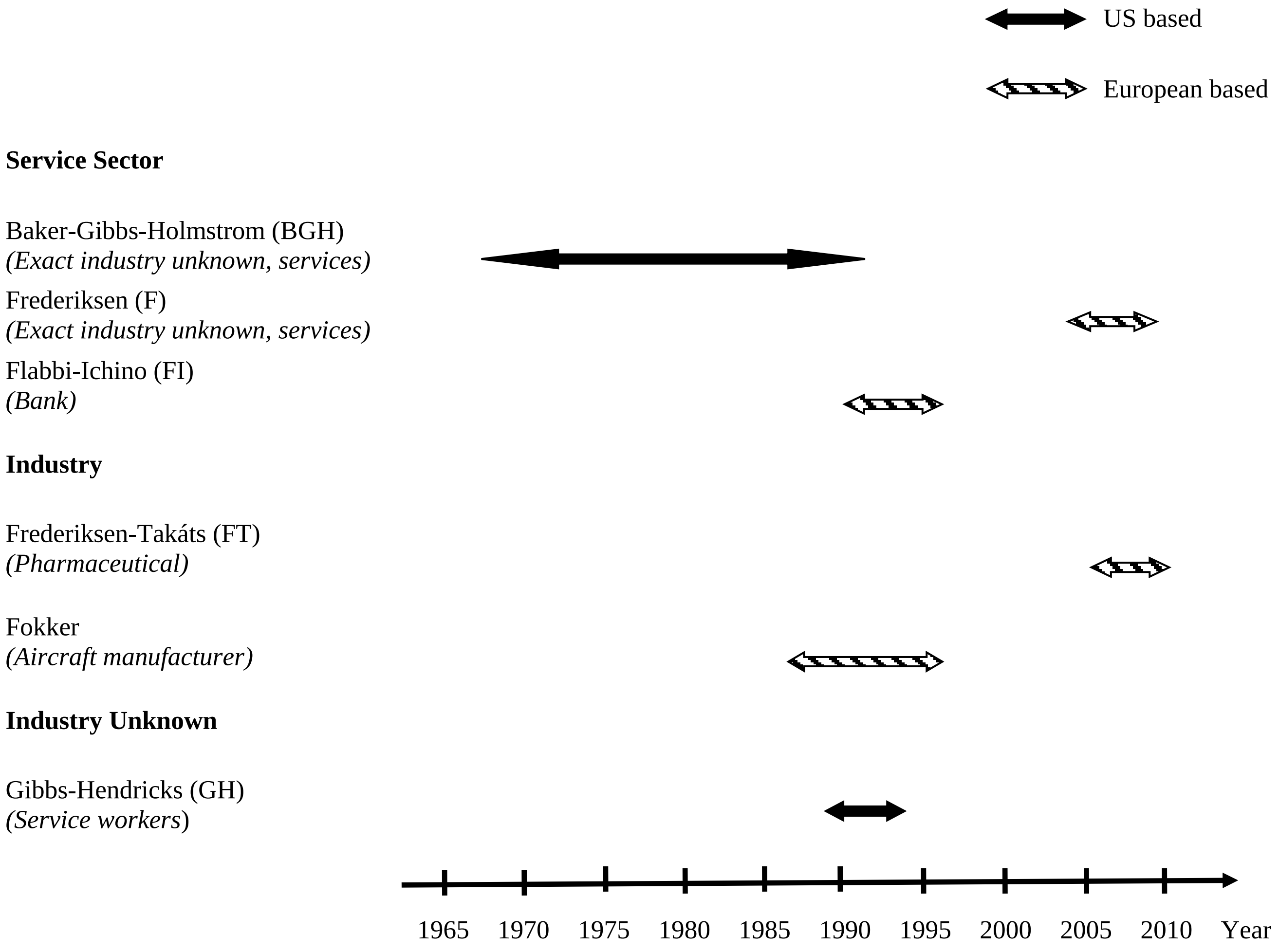
2. The Firms

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

We analyze personnel data from six large and very diverse companies. The data from these companies has been analyzed by us or by other researchers before, even though typically the focus of the prior studies has not been on performance evaluations. In this Section, we introduce these companies and briefly summarize the research on these data that precedes this study. We also present summary statistics on these data, paying particular attention to the performance evaluations.

With the exception of Fokker, we are not allowed to reveal the identities of the firms in our study. We will therefore use the names of the research teams that first analyzed these data to identify the different companies. We thus refer to the companies as Baker-Gibbs-Holmstrom (BGH), Gibbs-Hendricks (GH), Flabbi-Ichino (FI), Frederiksen-Takáts (FT), Frederiksen (F), and Fokker.

Figure 1. Location, Industry, and Time Period



The six companies are located in different countries, they operate in different industries and our data covers different time-periods. Figure 1 summarizes these differences. BGH and GH are based in the US, whereas FI, FT, F, and Fokker are located in Europe.[6] The companies span several sectors. BGH and F are in the service sector.[7] FI operates in the financial sector. FT is a pharmaceutical company and Fokker was an aircraft manufacturer. We do not know what industry GH belongs to. The data spans different time periods. The BGH data covers the period 1969 to 1988 and thus provides the earliest data available. FI, GH, and Fokker provide data from the late 1980s until the mid-1990s. The most recent data, from FT and F, cover the period from the early 2000s to 2011. With the exception of Fokker, a Dutch airplane manufacturer and the pharmaceutical company referred to as FT, both of which have data on blue-collar and white-collar workers alike, the other companies cover only white-collar workers.

We know turn to present the firms in more detail and to briefly summarize the prior research that has been conducted on these data.

Baker-Gibbs-Holmstrom (BGH)

In two ground-breaking papers, Baker, Gibbs, and Holmstrom (1994a,b) analyzed the personnel data of a U.S. based service-sector firm. The study focused on managerial employees (about 20 percent 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 considered in an informal way whether the data were consistent with models of employer learning, human capital acquisition, and simple provision of incentives. In summarizing the findings of BGH, Gibbs (1995) writes 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 (2011) reestablished that heterogeneous human capital accumulation is important,

[6] 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 location and identity remain unavailable.

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

but by using the information conveyed in the subjective ratings they also provided evidence 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) has 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. The second peculiarity is that, prior to 1981, we have data on base pay only. After 1981, we have data on both base pay and bonuses. Bonuses make up a small fraction of total compensation and for this reason we use the compensation data from the entire 1969 to 1988 period for our analysis of pay. 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 tenure that we present below are based on those observations for which tenure is available. 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.[8] Average total compensation (in 2000 dollars) is about $80,000, which far exceeds the average for the U.S. population.[9] This, as well as the demographics and the high education levels of staff, reflects the focus on managerial employees in this 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 suffices to accommodate market forces.

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

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

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

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 performance ratings in this firm 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 spans 1987 to 1996. We use 71,086 employee-year observations for the blue-collar workers, 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 for white-collar workers was $40,086 and for blue-collar workers it was $21,800.

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). The findings of these studies will be discussed in detail below.

Subjective performance evaluations are available only for non-managerial workers. Further, as do Flabbi and Ichino, we restrict the sample to males.

The data spans 1990 to 1995 and contains 63,390 employee-year observations that are based on 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.

Fredriksen –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 data available for analysis contains employees working in the country where the company's headquarter is located and besides these corporate activities they include employees in production, information-technology, and research and development. 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 measurement 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.

The FT data used in the analysis span 2007 to 2011 and thus constitute the most recent data among the six data sets. The data contain all relevant information on compensation and employee mobility, and a unique feature of the data is that we can identify separations as either quits or dismissals. 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.

Frederiksen (F)

The F firm is a service sector firm that Frederiksen (2010) analyzed for implicit and explicit incentives. Using a dynamic moral hazard model, Frederiksen predicted cross-sectional and individual earnings dynamics and 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 covers only domestic operations. The data comprises more than 20,000 unique employees and a total of 89,508 employee-year observations between 2004 and 2009. For the purpose of this study the F data constitutes the most complete dataset as it contains detailed information on wages, bonuses, performance ratings and employee mobility including information on whether separations are quits or dismissals. Average earnings in the firm are close to $50,000.

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	20,183
Observations	55,754	43,964	71,086	25,771	63,390	64,976	89,508
Observations with performance ratings	36,428	36,337	70,851	25,731	62,428	23,442	64,550
Fraction Managers	Only Managers	Breakdown not clear	Na	Na	Only Non-Managers	0.107	0.260
Compensation[1,2]							
All employees	Na	57,943 (37,055)	21,800 (4,103)	40,086 (12,851)	Na	45,550 (25,691)	48,334 (35,154)
Managers	80,069 (43,536)	Na	Na	Na	Na	70,921 (41,741)	60,930 (55,211)
Non-managers	Na	Na	Na	Na	29,128 (5,462)	42,566 (21,245)	43,738 (22,261)

[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). U.S. 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.

Subjective Performance Measures

Table 2 contains information on the performance scales and distributions 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, other scales as well. For instance, Fokker applied a five-point scale for its white-collar workers and a six-point scale for its blue-collar workers. The only firm applying a substantially different scale is GH, which uses an 18-point scale.

Table 2. Distribution of Subjective Performance Measures

		BGH	**GH[1]**	**Fokker Blue Collar**	**Fokker White Collar**	**FI**	**FT**	**F**
Rating scale		**1-5**	**18 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	2.58
	3	17.05	4	43.83	81.33	2.59	50.73	42.21
	4	50.00	16	40.53	14.13	14.37	39.72	47.38
	5	32.16	24	12.70	0.35	38.01	6.89	7.70
High	**6**	Na	6	1.48	Na	44.97	Na	Na

[1] GH applies a 1–18 point scale but six levels account for 93 percent of the ratings. For GH, only the rates pertaining to the six most common ratings are included.

In all firms, performance ratings are concentrated on a subset of the scale. The concentration is most extreme 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. From the distributions, it is clear that managers are very reluctant to give employees low ratings as these are rarely used.

The clear majority of employees is 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 subgroups from the evaluation program. In other companies, newly recruited employees are unlikely to have performance evaluations. For example, in F, employees do not receive ratings 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.[10] In what follows, we treat the incidence of evaluation as exogenous.

3. Performance Ratings over the Lifecycle: Medoff and Abraham Revisited

We begin our analysis of subjective performance ratings by investigating how they are related to experience and tenure. 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 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.[11] Thus, because they interpreted the subjective performance measures as cardinal measures of productivity, 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" (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 tenure. Performance ratings increase with age, 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.

[10] 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.

[11] Using the omitted variable bias formula, it should be clear that both of these findings are directly related in that controlling for performance ratings will attenuate the earnings-experience gradient if (a) performance ratings correlate positively with experience and (b) performance ratings correlate positively with wages.

Table 3. Average Performance by Age, Experience, and 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:							
– 30	4.35	8.86	3.42	3.09	4.74	3.43	3.40
	(0.64)	(1.82)	(0.59)	(0.37)	(0.76)	(0.64)	(0.64)
31 – 40	4.20	9.26	3.79	3.10	5.26	3.54	3.68
	(0.69)	(1.91)	(0.76)	(0.463)	(0.75)	(0.67)	(0.69)
41 – 50	4.02	9.24	4.00	3.12	5.44	3.52	3.66
	(0.73)	(1.96)	(0.83)	(0.49)	(0.74)	(0.67)	(0.67)
51+	3.90	9.13	4.29	3.11	5.58	3.44	3.56
	(0.72)	(1.93)	(0.91)	(0.51)	(0.70)	(0.66)	(0.66)
Experience:							
1-10	4.33	8.98	3.38	3.10	4.76	3.48	3.42
	(0.66)	(1.84)	(0.57)	(0.37)	(0.74)	(0.66)	(0.65)
11-20	4.17	9.26	3.69	3.10	5.22	3.53	3.69
	(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	3.65
	(0.73)	(1.95)	(0.81)	(0.48)	(0.73)	(0.66)	(0.67)
31-40	3.83	9.08	4.24	3.11	5.59	3.49	3.55
	(0.74)	(1.90)	(0.90)	(0.51)	(0.67)	(0.68)	(0.66)
Tenure:							
0-5	4.18	8.87	3.35	3.14	4.66	3.49	3.47
	(0.70)	(1.85)	(0.57)	(0.50)	(0.74)	(0.66)	(0.70)
6-10	4.05	9.34	3.66	3.11	5.15	3.54	3.65
	(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	3.70
	(0.77)	(1.95)	(0.77)	(0.43)	(0.75)	(0.66)	(0.66)
21+	Na	9.18	4.38	3.08	5.59	3.42	3.60
		(1.92)	(0.86)	(0.40)	(0.68)	(0.66)	(0.66)

Note: Experience refers to potential experience calculated as: Age minus 6 minus years of education. For BGH, tenure is only available for individuals entering the sample after 1969 and the tenure statistics are therefore limited to the sample of those individuals.

Table 4 presents regression results similar to those of Medoff and Abraham (1981). That is, we regress performance ratings on a polynomial in experience, a polynomial in tenure, and controls. Among the controls are year and education dummies, gender, age, and race when appropriate. We orthogonalize tenure using experience and the other controls so that the experience coefficients include any effect that operates through tenure. The tenure coefficients can be interpreted as "within experience" effects of tenure.

As in Table 3, we find that the performance-experience profiles are not stable across firms. At average experience, performance ratings decline for BGH, FT, and F, and they increase for GH, blue-collar Fokker, and FI. The quadratic polynomials are more regular; in all firms except BGH and white-collar Fokker the quadratic experience and tenure profiles are concave. This implies that performance ratings increase more rapidly among newly hired and/or young workers.

Job-level indicators generally explain significant fractions 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 tenure are typically sensitive to controlling for job levels. In FI, F, and FT, controlling for job levels attenuates the effect of experience on performance ratings by one-third to one-half.

Finally we note that the R-squares are low and the standard errors of these regressions are large, indicating substantial variation in performance that does not correlate with either experience or tenure. One explanation is that performance ratings are noisy measures of actual productivity.

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 tenure attenuate when performance ratings are included among the controls.[12] Flabbi and Ichino (2001) replicated these regressions for the FI firm. We consider the same specification for log earnings used in those papers. As do Abraham and Medoff (and FI), we find only weak evidence that controlling for performance evaluations reduces the magnitude of the experience and tenure effects on earnings. Following the interpretation of Medoff and Abraham (1980, 1981) that performance ratings can be used to compare performances across individuals at different experience levels, one would be forced to conclude that earnings do not reflect worker productivity. We prefer a different interpretation of the performance measures.

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

Table 4. Experience and 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)	Na	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)	0.038 (0.001)	0.015 (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)	-0.079 (0.003)	-0.029 (0.003)
orth. tenure	-0.034 (0.003)	-0.095 (0.004)	0.101 (0.005)		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)	0.014 (0.001)	0.014 (0.001)
Orth. 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)	-0.027 (0.003)	-0.024 (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	-0.132	-0.026
R-squared	0.09	0.17	0.04	Na	0.23	0.23	0.01	0.02	0.14	0.24	0.02	0.05	0.07	0.14
Reg. std. Error	0.68	0.65	1.89		0.65	0.65	0.41	0.41	0.73	0.69	0.66	0.65	0.64	0.62
N	36,290	36,290	36,316		54,761	54,761	20,737	20,737	62,428	62,428	23,442	23,442	54,793	54,793

Note: Experience refers to potential experience defined as: Age minus 6 minus years of schooling. In each column, we residualize tenure and 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, tenure is not available for those already in the firm in 1969. We substituted a value of 0 for the orthogonalized tenure measure for those with missing tenure.

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

The results in Tables 3 and 4 show that experience and tenure profiles in performance ratings vary considerably across companies even when controlling for job levels. In addition, the results in Table 5 show that experience profiles in log earnings regressions are not sensitive to including performance ratings. Following Medoff and Abraham (1980), these results would implausibly imply very large differences across firms in how worker productivity evolves with experience and that earnings are unaffected by performance. We believe that the results are better explained if performance ratings are interpreted as noisy measures of relative performance within narrowly defined peer groups—where the peer group is defined by demographics, education, and experience. This interpretation attributes the large difference in the performance-experience gradients across firms to differences in the use of the performance scales at different experience levels across firms. It also accommodates the finding that experience profiles in log earnings regressions are robust to the inclusion of performance ratings.

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 (0.001)	0.010 (0.006)	0.012 (0.001)	0.049 (0.001)	0.045 (0.001)		0.050 (0.000)	0.046 (0.000)	0.044 (0.000)	0.062 (0.001)	0.039 (0.000)	0.039 (0.000)
Experience squared / 100	-0.070 (0.002)	-0.020 (0.001)	-0.022 (0.001)	-0.092 (0.001)	-0.085 (0.002)		-0.092 (0.000)	-0.086 (0.000)	-0.084 (0.000)	-0.094 (0.001)	-0.057 (0.000)	-0.058 (0.000)
Orth. tenure	0.054 (0.002)	0.004 (0.001)	-0.001 (0.001)	0.039 (0.001)	0.036 (0.001)		0.013 (0.000)	0.011 (0.000)	0.010 (0.000)	0.015 (0.001)	0.010 (0.000)	0.009 (0.000)
Orth. tenure squared / 100	-0.144 (0.012)	0.027 (0.008)	0.011 (0.008)	-0.097 (0.003)	-0.085 (0.003)		-0.019 (0.000)	-0.018 (0.000)	-0.015 (0.000)	-0.003 (0.002)	-0.023 (0.001)	-0.022 (0.001)
Performance rating:						Na						
1			Omitted		Omitted				Omitted			Omitted
2			-0.001 (0.195)		-0.056 (0.005)				0.010 (0.012)			-0.041 (0.017)
3			0.091 (0.194)		-0.048 (0.009)				0.030 (0.012)			0.003 (0.017)
4			0.114 (0.194)		0.063 (0.005)				0.073 (0.012)			0.056 (0.017)
5			0.165 (0.194)		0.095 (0.005)				0.106 (0.012)			0.115 (0.022)
6					0.137 (0.008)				0.154 (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. 034	0. 004	0. 003
	(0.000)	(0.000)	(0.000)	(0.003)	(0.003)	(0.003)	(0.001)	(0.000)	(0.000)
Experience squared / 100	-0.009	0.010	0.011	-0.132	-0.110	-0.105	-0.071	-0.007	-0.006
	(0.000)	0.000)	(0.000)	(0.006)	(0.006)	(0.006)	(0.001)	(0.001)	(0.001)
Orth. tenure	0.025	0.025	0.009	0.096	0.087	0.085	-0.001	-0.001	-0.001
	(0.000)	(0.032)	(0.000)	(0.002)	(0.002)	(0.002)	(0.000)	(0.000)	(0.000)
Orth. tenure squared / 100	-0.030	-0.009	-0.001	-0.202	-0.180	-0.175	-0.000	0.001	0.002
	(0.001)	(0.000)	(0.000)	(0.007)	(0.007)	(0.007)	(0.001)	(0.001)	(0.001)
Performance rating:									
1			Omitted			Omitted			Omitted
2			0.115			-0.180			-0.019
			(0.016)			(0.187)			(0.025)
3			0.165			-0.036			-0.020
			(0.016)			(0.185)			(0.025)
4			0.181			0.125			0.030
			(0.016)			(0.185)			(0.025)
5			0.200			0.170			0.134
			(0.016)			(0.186)			(0.025)
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	0.240	0.671	0.683
N	61,825	61,825	61,825	23,442	23,442	23,442	54,785	54,785	54,785

Note: Experience refers to potential experience defined as: Age minus 6 minus years of schooling. In each column, we residualize tenure and 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. Reported are the effects for the six ratings reported in Table 1.

Section 4: How to Use Subjective Performance Evaluations

At the end of Section 2, we showed that the support of the performance evaluations varies across firms, that the distribution is discrete with only a small number of support points and that labels attached to ratings are meaningless as typically more than 90% of workers are rated to be “above average”. Further, in Section 3, we have established that there is a lot of variation in experience and tenure gradients of performance evaluations across firms.

In this Section, we propose a protocol that accounts for these features of performance ratings and at the same time make them operational. The usable information in performance evaluations resides in their information about relative rankings of employees within narrowly defined peer groups. Such peer groups should be defined on the basis of predetermined variables that are not themselves career outcomes. Thus, performance measures should be residualized using detailed experience and year dummies, gender, and race as well as interactions of linear experience and year trends with gender, education, and race before use. Such residuals then reflect the employee’s relative performance within a peer group defined by demographics, education, calendar time, and experience.

The observed discrete nature of the support of the ranking distributions and the difference in the support of rankings across firms implies that there is no natural scale for performance ratings and thus we have to explicitly account for the ordinal nature of the data. In doing so we impose the assumption that performance ratings are based on normally distributed latent variables. The cut-offs that determine how these latent variables map into observed ratings will need to vary by the predetermined characteristics that define the peer groups. It is now possible to report the correlations between the latent indexes that are implied by the joint distribution of the performance measures over time. Such measures of performance ratings can also be correlated with, for instance, residualized compensation measures if it is assumed that the compensation measures are log-normally distributed.

When examining relations between performance and promotions, demotions, quits and layoffs such measures can be regressed directly on the residualized performance measures.

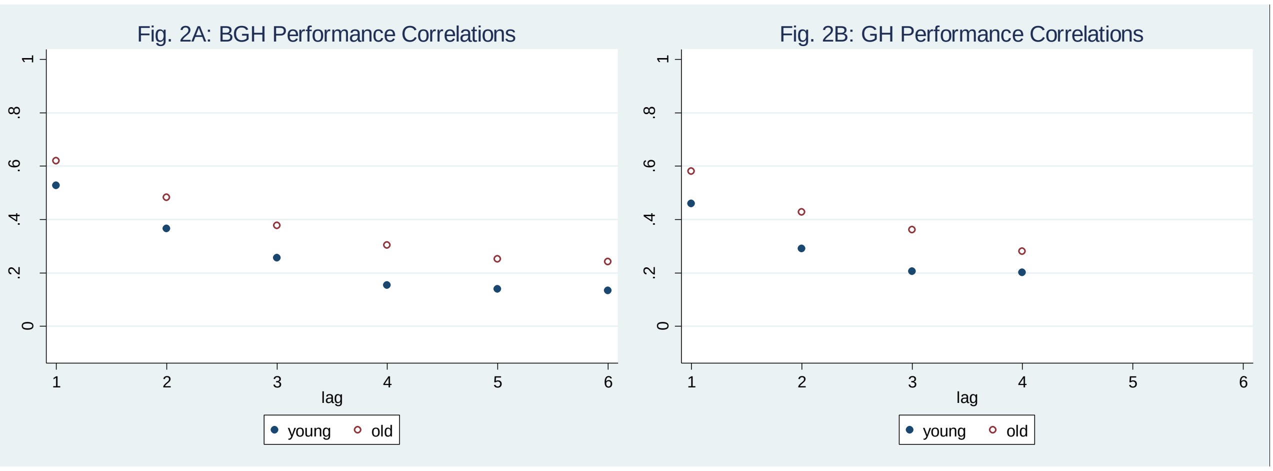
Using this methodology the next step in our analysis is to examine the data for consistent patterns in the joint distributions of performance ratings and career outcomes. We begin with autocorrelation patterns in performance ratings (Section 5) and then in subsequent sections we investigate how performance 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.[13] For each firm, we show the correlations for younger workers (years of experience 1-15) and older workers (years of experience 16-30). These correlations are calculated using the unbalanced panels generated by the personnel data sets, and they are averages across individuals within each of the two experience levels.

The autocorrelations in performance display many robust similarities across companies. To begin, the first-order autocorrelations tend to be high in all six data-sets. They lie 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 the more experienced workers. The age differences in these correlations are relatively small in BGH, GH, FT, and F. Looking across lags, we find (with one exception for the sixth autocorrelation among young white-collar employees in Fokker) that all the correlations are positive. Typically they 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 location (United States or Europe), whether they pertain to blue collar and white collar, and regardless when between 1969 to 2010 they were collected.

Figure 2. Performance Autocorrelations



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

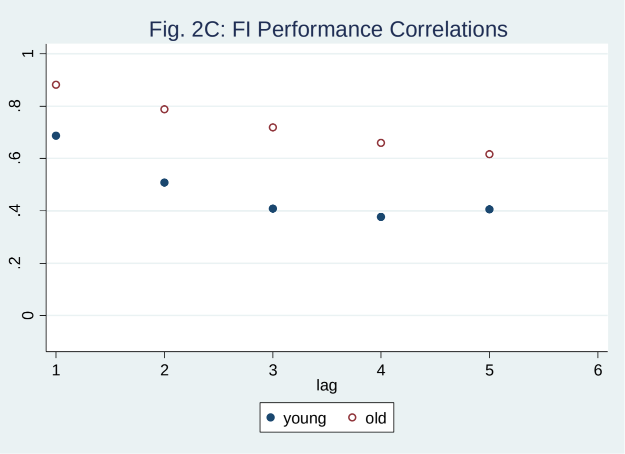


Fig. 2C: FI Performance Correlations

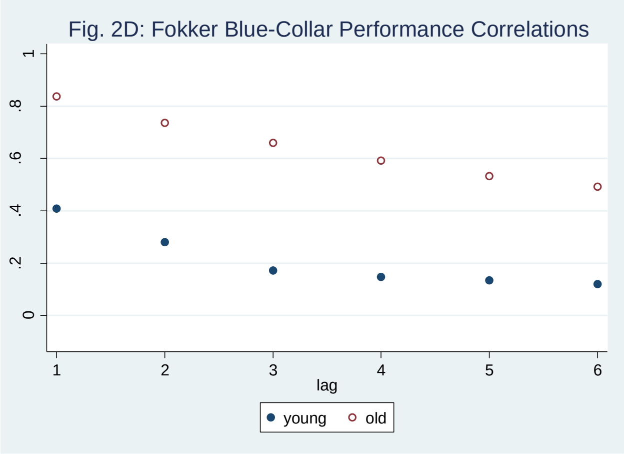


Fig. 2D: Fokker Blue-Collar Performance Correlations

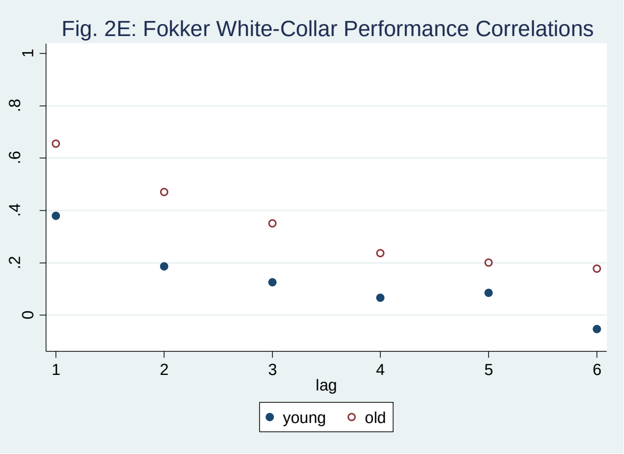


Fig. 2E: Fokker White-Collar Performance Correlations

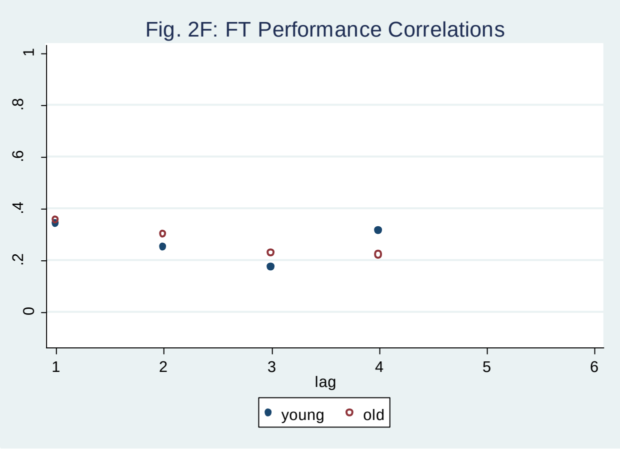


Fig. 2F: FT Performance Correlations

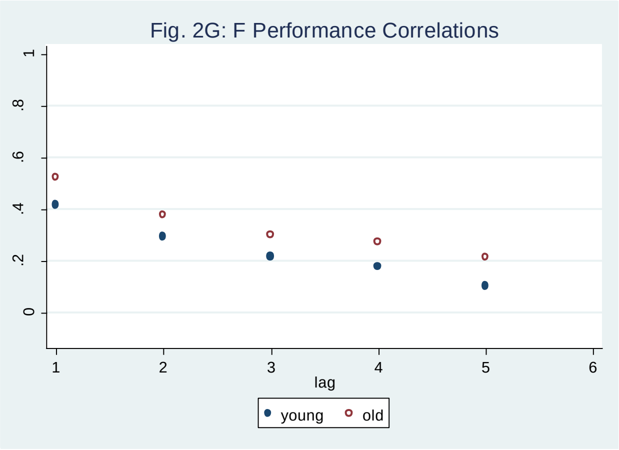


Fig. 2G: F Performance Correlations

6. Correlations Patterns in Compensation Growth

We next consider how growth in various compensation measures correlates across different lags. In their seminal papers, BGH (1994 a, b) shows that some workers experience consistently faster earnings growth and move more rapidly through the ranks of the firm; they seem to be proceeding as if along a “fast-track.” We revisit this question here – both for total compensation and, where possible for base pay and bonuses.

Table 6 shows how the growth in a given residualized compensation measure between *t*-1 and *t* correlates with growth in the same measure between t-k-1 and t-k for k=1,..,5. The autocorrelation patterns in log total compensation growth vary considerably across companies. In BGH, we see only weak autocorrelations. In other firms, the correlations are negative (GH, FI and F), positive (Fokker), and mixed (FT). A clear tendency, however, is that correlations become weaker with distance.

Autocorrelation patterns in base pay and bonus payments are very different. In all firms, the first autocorrelation in log bonus growth is strongly negative, which suggests that periods of high bonus growth are followed by periods of low growth. The evidence on the autocorrelations in log base pay growth is more mixed. For two firms (FI and F), we find negative autocorrelations in log base pay growth. For GH, autocorrelations are mixed. For the remaining firms, log base pay growth is positively autocorrelated.

The mixed findings on the autocorrelation structures presented in this section show that the earnings process is to a large degree firm specific. In addition, when variable pay components such as bonuses are part of the compensation package, the dynamics of base pay and total compensation may be very different.

Table 6. Growth Correlations for Different Compensation Measures

Panel A: BGH, GH, Fokker								
	BGH (1981-1988)	BGH (1981-1988)	BGH (1981-1988)	GH			Fokker Blue Collar	Fokker White Collar
Compensation Measure	Log Base Pay	Log Total Compen-sation	Log Bonus	Log Base Pay	Log Total Compen-sation	Log Bonus	Log Total Compen-sation	Log Total Compen-sation
Correlation of growth between t and t+1 with growth separated by:								
1 lag	0.24	-0.05	-0.27	0.03	-0.15	-0.33	0.10	0.27
2 lags	0.18	-0.04	-0.24	-0.01	-0.08	-0.10	0.14	0.23
3 lags	0.12	-0.04	-0.17	0.07	-0.15	-0.27	0.08	0.19
4 lags	0.07	0.03	0.02	Na	Na	Na	0.04	0.12
5 lags	0.01	-0.02	0.15	Na	Na	Na	0.05	0.11

Panel B: FI, FT, F									
		FI			FT			F	
Compensation Measure	Log Base Pay	Log Total Compen-sation	Log Bonus	Log Base Pay	Log Total Compen-sation	Log Bonus	Log Base Pay	Log Total Compen-sation	Log Bonus
Correlation of growth in t with growth separated by:									
1 lag	-0.25	-0.24	-0.45	0.12	-0.14	-0.46	-0.53	-0.30	-0.45
2 lags	-0.03	-0.02	-0.03	0.07	0.36	0.05	-0.04	-0.16	-0.05
3 lags	-0.03	-0.04	0.01	0.49	0.41	0.02	-0.05	-0.01	0.00
4 lags	0.00	0.00	-0.03	Na	Na	Na	-0.01	-0.01	0.00
5 lags	Na	Na	Na	Na	Na	Na	Na	Na	Na

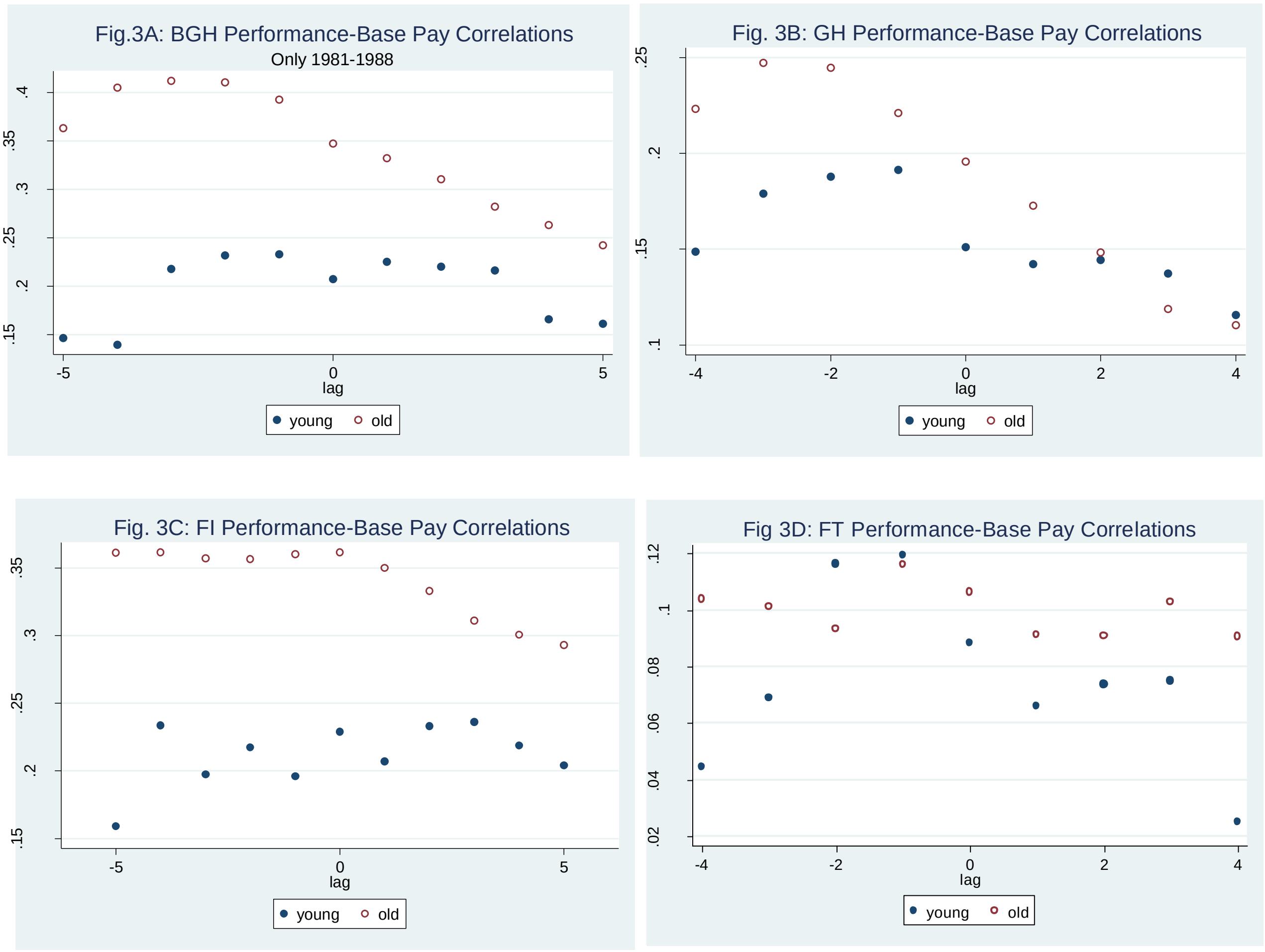
Note: We show correlations in growth for the various residualized compensation measures across up to five lags. These are pair-wise correlations and therefore the number of observations going into each cell is not common across rows within columns. We show correlations from the 1981-1988 period for BGH, because this is the period for which we have base-pay and bonus information.

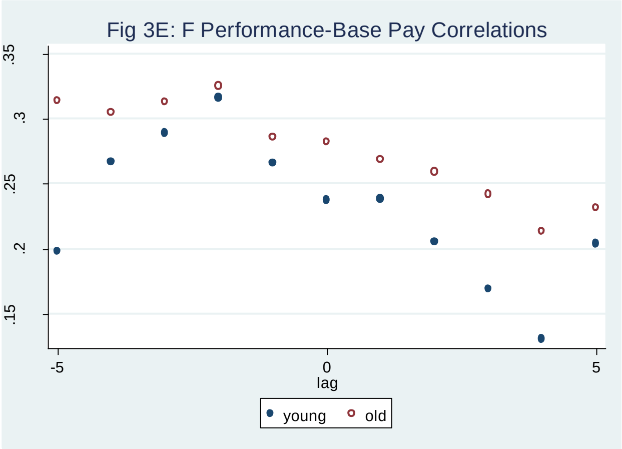
7. Correlations of Performance Ratings with Earnings Components

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 base pay. We do not simply consider the contemporaneous correlations, but also consider how earnings and performance ratings correlate when they are separated by various leads and lags.

To remain consistent, we residualize performance and the three log earnings measures in the same manner we did in Sections 4 and 5. For all earnings measures, we consider the correlation of 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. We obtain these correlations for two groups: individuals with 0-15 years of experience and individuals with 16-30 years of experience.

Figure 3. Performance-Base Pay Correlations

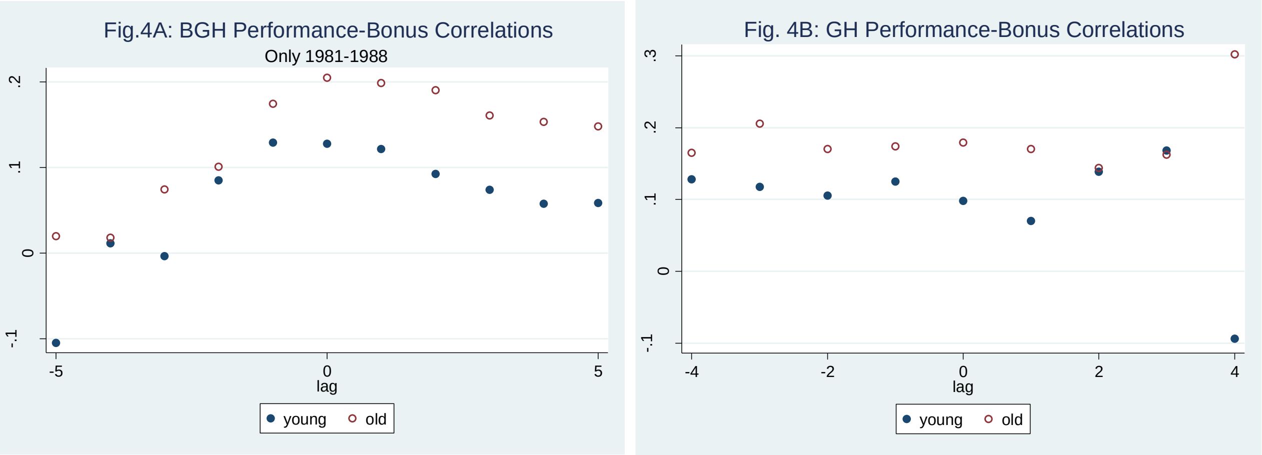




In Figure 3A-E, we show the correlations between performance and base pay for the five data sets where we can break down total compensation into base pay and bonuses. A consistent finding across firms, and in particular for more experienced workers, is that the correlations of base pay with contemporaneous ratings or ratings obtained in the near past exceed the correlations of base pay with future performance ratings. Kahn and Lange (2011) first noticed this discontinuity in the correlation patterns around the present time in their analysis of the BGH data. Here, we find the same asymmetry in GH, FT and F and among older workers for FI.

Kahn and Lange also emphasize that in BGH, the correlations of log base pay with performance ratings are higher for older workers than they are for younger workers. We find the same patterns in the other firms (with a few exceptions for GH and FT). A final finding common to all data sets that is not easily explained within the Kahn-Lange framework is that base pay correlates less with performance ratings obtained far in the future than with those obtained in the immediate future.

Figure 4. Performance-Bonus Correlations



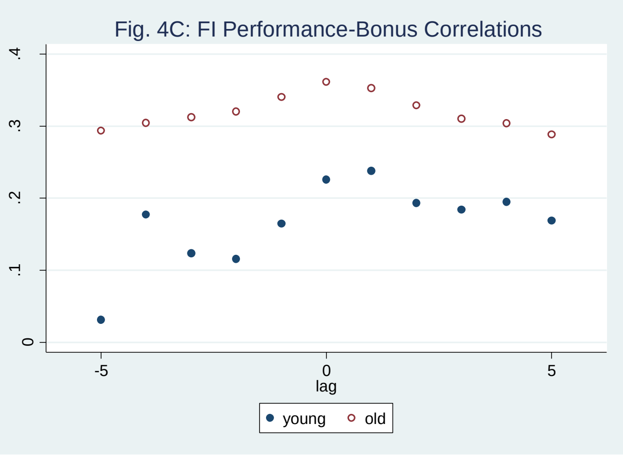


Fig. 4C: FI Performance-Bonus Correlations

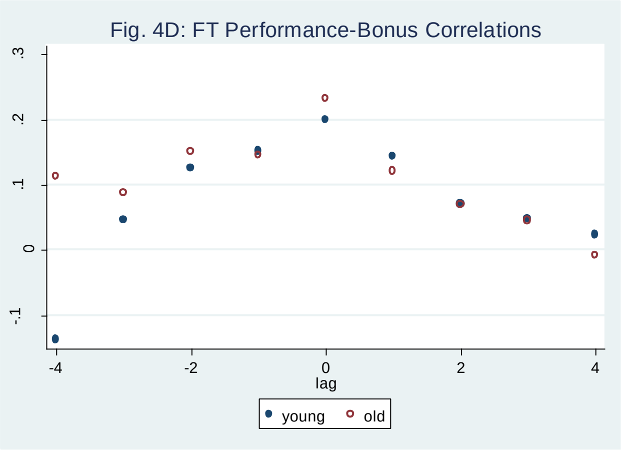


Fig. 4D: FT Performance-Bonus Correlations

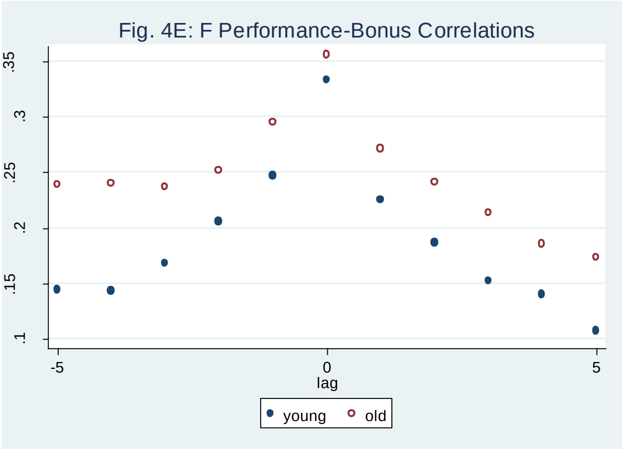


Fig. 4E: F Performance-Bonus Correlations

We next turn to the correlations between performance ratings and log bonuses (see Figures 4A-E). The observed patterns are quite different from those established for base pay. In GH, we find no discernible pattern, regardless of the age of the workers. In FT, and F, performance pay and bonuses are highly correlated with current ratings. This pattern is less pronounced but still discernible in BGH and FI. Only in GH is there no evidence that contemporaneous performance and bonuses are more highly correlated than bonuses and performance in other periods. Overall, the findings provide some support for the hypothesis that bonuses are being used to provide direct incentives.

Figure 5. Performance-Total Compensation Correlations

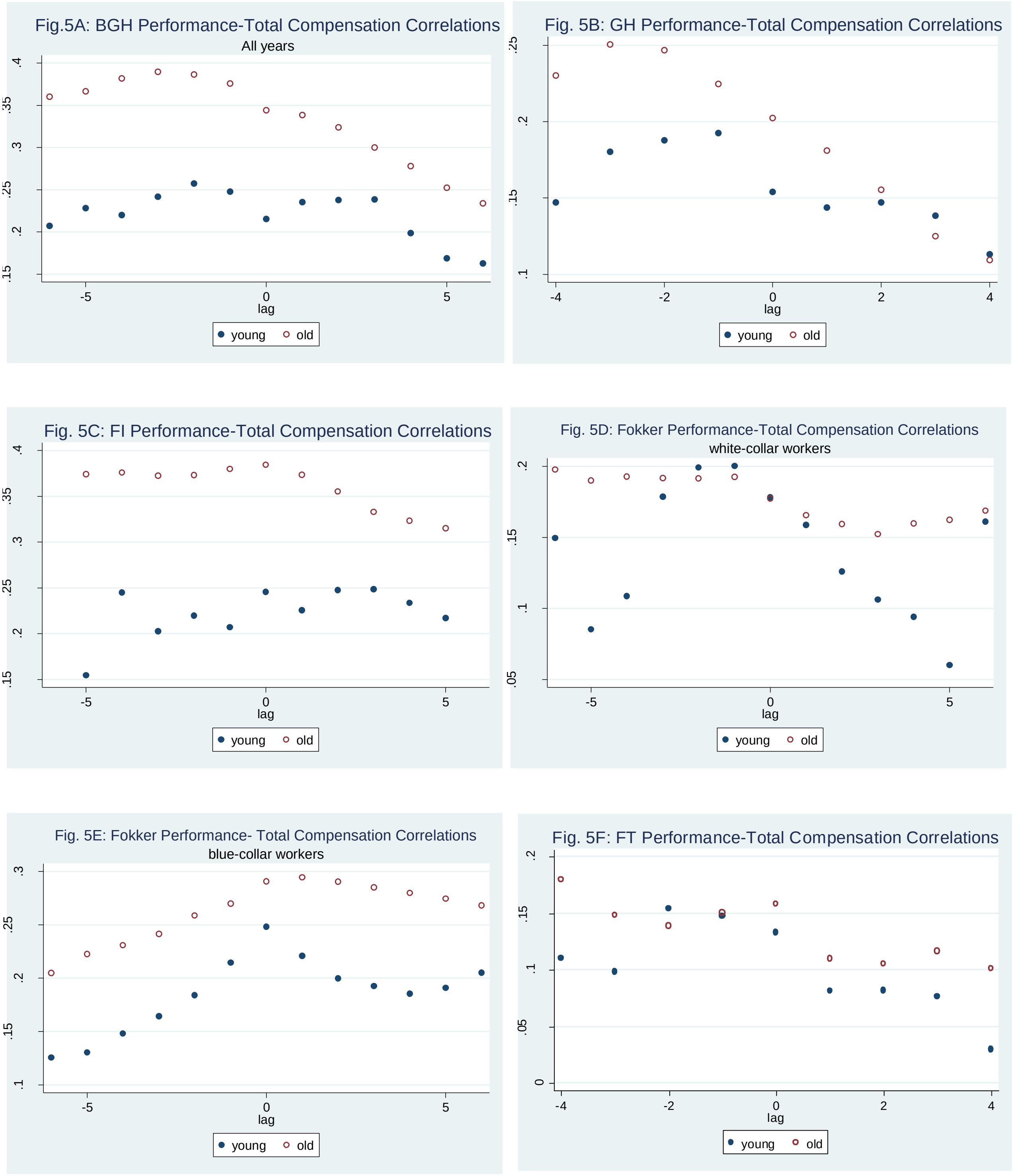


Fig.5A: BGH Performance-Total Compensation Correlations

All years

Fig. 5B: GH Performance-Total Compensation Correlations

Fig. 5C: FI Performance-Total Compensation Correlations

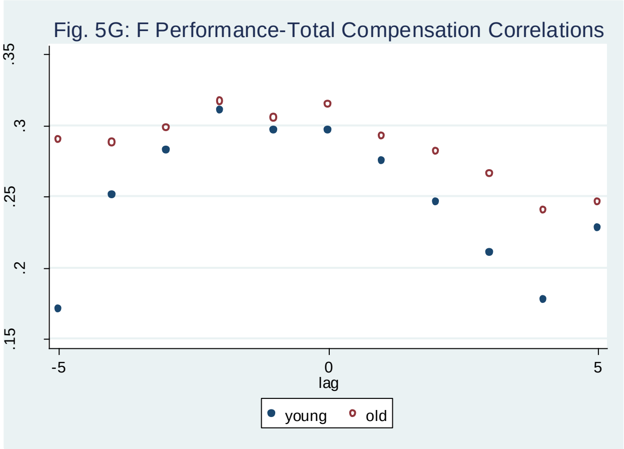
Fig. 5D: Fokker Performance-Total Compensation Correlations

white-collar workers

Fig. 5E: Fokker Performance- Total Compensation Correlations

blue-collar workers

Fig. 5F: FT Performance-Total Compensation Correlations



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 not possible, we find large differences between blue-collar 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 be in part be explained by the very strict administrative rules governing pay, as described in Dohmen (2004).

The results presented in this section show some common patterns. 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 contemporaneous correlations between log total compensation and performance ratings are high and between 0.15 and 0.40 for more experienced workers and relatively low and between 0.10 and 0.30 for less experienced workers. Second, we find a step pattern in the correlations of total compensation and base pay across leading and lagging performance ratings in many, but not, all firms. 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.

8. 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, where the two periods are separated by one year, 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 7 present statistics on the frequency of promotions and demotions in the different firms.[14] The data in the first three rows describe promotion and demotion for all individuals in the firm. The promotions frequencies vary substantially across firms and range from 2.4 percent to 16 percent. Demotions are less frequent but not uncommon. The ratio of promotions to demotions is between 3 and 80, but three firms have ratios below 4.6.

In Table 7 (lower part), we show the time to first promotion. We restrict the sample to individuals who are both recruited and who stay with the firm for at least six consecutive years within the sample period. Again, there is a lot of variation across firms. Almost 80 percent of employees in BGH, but only 14 percent of blue-collar workers at Fokker, are promoted within the first five years. For the other firms, the probability of being promoted during the first five years varies between 45 percent and 65 percent. Conditional on being promoted within the first five years, promotions are typically much more common within the first two or three years. The main exception is FI, where a very large fraction of employees is promoted during the fifth year of employment.[15] Thus, we find that in most firms, a substantial fraction of new employees are promoted relatively soon after they are recruited. However, it is also noteworthy that a large fraction is passed over for promotion in the first five years of employment and this group may never receive a promotion.

[14] We use job levels to construct promotions and demotions in BGH, Fokker, FI, FT, and F. Job levels in BGH are those generated by BGH (1994a,b). GH provides direct measures of promotions and demotions.

[15] 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 agreed.

Table 7. Promotion and Demotion Probabilities

	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%	12.1%
Prob. of Demotion	0.2%	0.4%	1.1%	2.0%	0.0%	0.8%	1.1%

Time to first promotion (if promoted within the first five years)

	BGH	GH	Fokker Blue Collar	Fokker White Collar	FI	FT	F
Year 1	31.8%	21.1%	16.1%	25.9%	12.9%	25.9%	14.0%
Year 2	35.1%	27.6%	21.2%	21.9%	10.6%	34.6%	21.1%
Year 3	17.6%	30.2%	22.3%	23.9%	9.0%	14.8%	29.8%
Year 4	9.5%	12.6%	25.9%	22.7%	9.4%	24.7%	14.9%
Year 5	5.9%	8.6%	14.5%	5.7%	58.0%	-	20.2%
Never or later	21.3%	55.4%	77.9%	43.6%	40.5%	87.9%	34.9%

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 the blue-collar jobs.

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 that they may be a consequence of differences in administrative and reporting practices across firms.[16] Nevertheless, two findings are consistent across firms: there are many more promotions than demotions, and there are more promotions among recent hires.

High performance ratings are associated with an increased promotion probability. Table 8 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

[16] 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 when we observe substantially more demotions. Some of these demotions are arbitrary reclassifications of departments within the firm hierarchy without obviously entailing changes in the job responsibilities or classification according to the union wage contracts.

(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 all negative and very similar: they all fall in the interval -0.012 to -0.033.

Table 8. 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	0.053
Performance at t and demotion between t and t+1	-0.024	-0.016	-0.016	-0.030	Na	-0.012	-0.033

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

In Table 9, we explore the relation between performance and promotions further and present odds ratios from regressions relating promotions to the performance of employees (relative to their job level) during the last 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 GH, lagged performance has a negative effect on the promotion probability, whereas in FT and F, the effect is positive.

Table 9. 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)	1.54 (0.06)
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)	1.22 (0.04)
Pseudo R-squared	0.220	0.082	0.039	0.046	0.103	0.121	0.117
N	13,167	12,417	48,857	17,671	33,339	6,510	24,911

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 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.

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

We now turn to an examination of the correlations between employee turnover and performance. Although most research on employee turnover is restricted to addressing job separations (when an employee leaves a company), two of the firms have provided information on the reason for job separation. This allows us to examine the relation between performance and, respectively, quits (employee initiated separation) and dismissals (employer initiated separation).

In Table 10 we present job separation probabilities for the six firms. The separation rates in the American firms (10.7 percent and 12.5 percent) exceed those in the 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 7.5 percent. The separation rates in these companies thus line up with the stereotypical view that European labor markets are characterized by less mobility than the U.S. labor market, and in particular the perception that there is very little labor mobility in Italy.

Table 10. 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%	5.91%
Quit rate						4.77%	5.31%
Dismissal rate						1.70%	0.60%
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	-0.046
Performance at t and quit between t and t+1	Na	Na	Na	Na	Na	-0.029	-0.037
Performance at t and dismissal between t and t+1	Na	Na	Na	Na	Na	-0.083	-0.040

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.

For FT and F, we examine the distinction between quits and dismissals. In both of these firms, the majority of separations are classified as quits. Dismissals are more frequent in FT, where they occur at a rate of 1.7 percent annually. On average, only 0.6 percent of workers at F are dismissed each year. Table 10 also shows that the correlation between separations and job performance is uniformly negative. The correlations are particularly strong in BGH, GH, and FT 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.

The relation between performance and separations is explored in more detail in Table 11. We use the same specification as in Table 9 (except for the change of the independent variable) and establish the result 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 F and Fokker (blue-collar only) does lagged performance reduce the exit rate.

Table 11. Separations and Performance (Logit)

Endogenous variable: Separation between t and t+1	BGH	GH	Fokker Blue Collar	Fokker White Collar	FI	FT	F
Performance at t	0.63 (0.03)	0.86 (0.02)	0.83 (0.03)	0.74 (0.07)	0.74 (0.14)	0.58 (0.05)	0.66 (0.04)
Performance at t-1	0.98 (0.04)	0.98 (0.02)	0.80 (0.03)	0.90 (0.09)	0.95 (0.18)	1.07 (0.10)	0.89 (0.07)
Pseudo R-squared	0.080	0.032	0.135	0.144	0.150	0.057	0.111
N	22,041	6,729	34,443	12,957	50,136	6,510	35,060

Note: Separation between *t* and *t*+1 is regressed on residualized performance from time *t* and *t*-1. All regressions control for quadratics in experience and orthogonal tenure, gender, and year dummies, and race when appropriate. Each specification furthermore includes dummy variables for the job levels in *t* and *t*-1.

10. Conclusion

In most employment relationships, objective performance measures are unavailable. For this reason, supervisors are often asked to subjectively evaluate workers' performances. In turn, the subjective performance ratings become part of the information employers use when they sort, select, and create incentives for their employees. Because personnel data including performance ratings are still rare, very little is known about how these ratings are used and what consequence they have for employees’ careers. The purpose of this paper has been to uncover any empirical regularities in the use of performance ratings across 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 the way performance scales are structured and used and in how performance ratings correlate with pay and other career outcomes. For instance, the correlation between total compensation and contemporaneous performance never exceeds 0.4, is typically above 0.2, and is generally higher for more-experienced workers. Less robust, but still notable, is our finding that, in many firms, base pay and total compensation correlate more highly with past performance ratings than future performance ratings. We also find many similarities in how performance and employee mobility is related. For example, promotions are always positively correlated with recent performance, whereas demotions and transitions out of the firm are negatively correlated with performance.

There are a number of exceptions and idiosyncrasies that likely stem from specific circumstances in the studied 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 because of 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 these correlations to a large extent are similar across firms—even if there are exceptions. For this reason, we believe that, despite prior concerns, 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 examine 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.

10. References

Baker G. P., M. Gibbs, and B. Holmstrom. 1993. "Hierarchies and compensation: A case study." *European Economic Review* Vol. 37, No. 2-3: 366-378.

__________. 1994a. "The internal economics of the firm: Evidence from personnel data." *Quarterly Journal of Economics* Vol. 109, No. 4: 881-919.

__________. 1994b. "The wage policy of the firm." *Quarterly Journal of Economics* Vol. 109, No. 4: 921-955.

Bandiera, O., I. Barankay, and I. Rasul. 2005, "Social preferences and response to incentives: Evidence from personnel data." *Quarterly Journal of Economics* Vol. 120, No. 3: 917-962.

__________. 2007. "Incentives for managers and inequality among workers: Evidence from a firm level experiment." *Quarterly Journal of Economics* Vol. 122, No. 2: 729-773.

Barlevy, Gadi, and Derek Neal. 2012, "Pay for Percentile." *American Economic Review* (forthcoming).

DeVaro, J., and M. Waldman. 2012. "The signaling role of promotions: Further theory and empirical evidence." *Journal of Labor Economics* Vol. 30, No.1: 91-147.

Dohmen, T. 2004. "Performance, seniority, and wages: Formal salary systems and individual earnings profiles." *Labour Economics* Vol. 11, No. 6: 741-763.

Dohmen, T., B. Kriechel, and G. A. Pfann. 2004. "Monkey bars and ladders: The importance of lateral and vertical movements in internal labor market careers." *Journal of Population Economics* Vol. 17, No. 2: 193-228.

Flabbi, L., and A. Ichino. 2001. "Productivity, seniority and wages: New evidence from personnel data." *Labour Economics* Vol. 8, No. 3: 359-387.

Frederiksen, A. 2010. "Earnings progression, human capital and incentives: Theory and evidence." IZA Discussion Paper 4863, Bonn.

Frederiksen, A., and E. Takáts. 2011. "Promotions, dismissals and employee selection: Theory and evidence." *Journal of Law, Economics and Organization* Vol. 27, No. 1: 159-179.

Gibbons, R., and M. Waldman. 1999. "A theory of wage and promotion dynamics inside firms." *Quarterly Journal of Economics* Vol. 114, No. 4: 1321-1358.

__________. 2006. "Enriching a theory of wage and promotion dynamics inside firms." *Journal of Labor Economics* Vol. 27, No. 1: 59-107.

Gibbs, M. 1995. "Incentive compensation in a corporate hierarchy." *Journal of Accounting and Economics* Vol. 19, No. 2–3: 247–277.

Gibbs, M., and W. Hendricks. 2004. “Do formal salary systems really matter?” *Industrial and Labor Relations Review* Vol. 58, No.1: 71-93.

Goldhaber, D., and M. Hansen. 2010. “Using performance on the job to inform teacher tenure decisions.” *American Economic Review: Papers & Proceedings* 100 (May): 250–255.

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

Hamilton, B. H., J. A. Nickerson and H. Owan, 2003. “Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation.” *Journal of Political Economy* Vol. 111, No.3: 465-497.

Kahn, L., and F. Lange. 2011. “Learning about employee and employer learning: Dynamics of performance and wage measures.” Working paper. New Haven, CT: Yale University.

Lazear, E. P. 2000. “Performance pay and productivity.” *American Economic Review* Vol. 90, No.5: 1346-1361.

Mas, A. and E. Moretti. 2009. “Peers at Work.” *American Economic Review* Vol. 99, No.1:112-145.

MacLeod, W. B. 2003. “Optimal contracting with subjective evaluation.” *American Economic Review* Vol. 93, No.1: 216-240.

Medoff, J., and K. Abraham. 1980. “Experience, performance, and earnings.” *Quarterly Journal of Economics* Vol. 85, No. 4: 703-736.

__________. 1981. “Are those paid more really more productive?”, *Journal of Human Resources* Vol. 16, No. 2: 186-216.

Milgrom, P. R. 1988. “Employment contracts, influence activities, and efficient organization design.” *Journal of Political Economy* Vol. 96, No. 1: 42-60.

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

Oyer, P., and S. Schaefer. 2010. “Personnel economics: hiring and incentives.” In *The handbook of labor economics*, ed. O. Ashenfelter and D. Card. Elsevier, Amsterdam

Prendergast, C., and R. H. Topel. 1993, “Discretion and bias in performance evaluation.” *European Economic Review* Vol. 37: 355-365.

__________. 1996. “Favoritism in organizations.” *Journal of Political Economy* Vol. 104, No. 5: 958-978.

Shearer, B. 2004. “Piece rates, fixed wages and incentives: Evidence from a field experiment.” *Review of Economic Studies* Vol. 71, No. 2: 513-534.

Tirole, J. 1986. “Hierarchies and bureaucracies: On the role of collusion in organizations.” *Journal of Law, Economics, and Organizations* Vol. 2, No. 2: 181-214.