AARHUS UNIVERSITY

Coversheet

This is the accepted manuscript (post-print version) of the article.

Contentwise, the post-print version is identical to the final published version, but there may be differences in typography and layout.

How to cite this publication

Please cite the final published version:

Andersen, L. B., Heinesen, E., & Pedersen, L. H. (2016). Individual Performance: From Common Source Bias to Institutionalized Assessment. *Journal of Public Administration Research and Theory*, 26(1), 63-78. DOI: 10.1093/jopart/muv010

Publication metadata

Title:	Individual Performance: From Common Source Bias to Institutionalized Assessment
Author(s):	Andersen, L. B., Heinesen, E., & Pedersen, L. H.
Journal:	<i>Journal of Public Administration Research and Theory</i> , 26(1), 63-78
DOI/Link:	https://doi.org/10.1093/jopart/muv010
Document version:	Accepted manuscript (post-print)

General Rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Individual performance: From common-source bias to institutionalized assessment

Abstract

Performance is perhaps the most central concept in public administration research, and this article discusses theoretically and investigates empirically how we can obtain more consistent performance measures. Theoretically, we combine existing arguments in public administration with institutional theory and the sociology of professions. Empirically, we ask whether different measures of individual performance produce different results. The investigated performance measures vary with regard to risk of common data source bias, standardization of assessment criteria and external verification of the assessment. Our investigated explanatory variables are intrinsic motivation, public service motivation and job satisfaction. Combining survey and administrative data for 747 lower secondary school teachers (teaching 5,679 students in 85 schools), we analyze four different measures of the same performance dimension for the same teachers: the teachers' self-reported contributions to students' academic skills, the students' marks for the year's work given by the teacher, marks in oral exams with one external examiner and the teacher, and marks in written exams with at least one external examiner. The associations are systematically stronger when the performance measure comes from the same data source as the explanatory variables, but when separate data sources are used and the measurement scale is institutionalized, the level of external verification does not matter much. Based on institutional theory and the sociology of professions, we develop a theoretical argument that can explain this.

1. Introduction

Finding ways to improve performance in the public sector is a very important goal in public administration research (Boyne 2003a; 2003b; Brewer and Selden 2000; Meier et al. 2007), and we therefore need performance measures that give consistent estimates of the associations between performance and other constructs. Consistency here means that analyses with different measures of the same performance dimension give similar results. Increased consistency is thus an important step towards a better understanding of good performance and its causes in the public sector, but it demands that we empirically investigate and theoretically discuss both assessment processes and data challenges involved in performance measurement.

Right now, the focus in the public administration literature is very much on the data challenges. Common data source bias has been shown to produce false positives when perceptual measures from the same source as the explanatory variables are employed to measure organizational performance (Meier and O'Toole 2013a; 2013b; Favero and Bullock 2014). This is especially the case when survey

items are based on vague and ill-defined criteria for performance (Meier and O'Toole 2013a), and when the explanatory variables are also perceptual. However, whether common data source bias is also a serious problem for the consistency of associations between key explanatory variables and *individual* performance still remains unanswered.

Common data source bias is not the only challenge in our efforts to obtain consistent performance measures. Andrews et al. (2010), for example, highlight key aspects of the assessment process when they convincingly show that internal perceptions of performance tend to be higher than external assessments of performance. But do the estimates of the associations between performance and other constructs depend on the level of external assessment? Andrews, Boyne and Walker (2006, 16) expect performance measures to be more consistent if there are standardized performance criteria and an external process for verifying their accuracy, but we need more empirical knowledge about how the assessment process is important for the results before we can further theorize on the ways to best obtain consistent performance measures.

The ambition of this article is to empirically investigate and theoretically discuss how we obtain consistent performance measures. We do this by analyzing whether the associations between perceptual explanatory variables and individual performance depend on whether the performance measure (1) comes from the same data source as the explanatory variables, (2) includes verification that is external to the actor whose performance is directly or indirectly being evaluated and (3) is standardized. Based on the results from our empirical tests, we discuss theoretically what can be done to provide more consistent performance measures.

This article adds to existing knowledge in two ways. First, the above-mentioned studies (Meier and O'Toole 2013a; 2013b; Favero and Bullock 2014) look only at organizational performance, whereas we have access to individual performance data. Analyzing individual performance contributes

to our empirical understanding of biases in performance measures, and individual performance is an important concept in itself. Second, and perhaps most importantly, we contribute to the theoretical understanding of why some performance measures are more consistent than others, and how we can obtain more consistent performance measures in the future. Drawing on the sociology of professions and institutional theory in general, we argue that the institutionalization of assessment by strengthening professional norms may be a beneficial way to move forward.

Empirically, we analyze four different measures of the same performance dimension for the same lower secondary school teachers in Denmark: How much do the teachers contribute to their students' academic skills? The four measures are teachers' self-reported contribution to their students' academic skills, marks for the year's work for a given teacher's students in the relevant subject, oral exam marks for the teacher's students in the relevant subject, and written exam marks for the teacher's students in the relevant subject. As we explain in detail below, these measures vary systematically in terms of potential for common data source bias, involvement of external examiners and standardization. We have chosen this empirical case because it allows us to compare different measures of the same performance concept with variation in characteristics that the literature sees as important.

In order to evaluate the consistency of the measures, we analyze how they are associated with the teachers' public service motivation, intrinsic motivation, and job satisfaction, respectively. We use these explanatory variables for three reasons. Firstly, and most importantly, they are theoretically expected to affect performance in public organizations and have been shown empirically to be associated with individual performance (Andersen, Heinesen, and Pedersen 2014; Bellé 2013; Brewer 2008; Callahan et al. 2003; Judge et al. 2001; Vandenabeele 2009; Williams and Anderson 1991). The theoretical backing increases the relevance of the tests, and we avoid knocking down straw-man relationships. Secondly, the measures of the concepts are perceptual, and the potential for bias in their

association with performance is therefore higher. Thirdly, they are key concepts in public administration theory in the sense that knowledge on how they relate to performance is relevant to the development of a better public sector.

The analysis combines administrative data on students' marks and backgrounds with school information about teacher-student combinations and a teacher survey. We have valid data for 5,679 students, 85 schools, and 747 teachers. We also have information about the performance measures mentioned above for the same teachers, which renders it possible to examine whether their public service motivation, intrinsic motivation and job satisfaction relate differently to the four different measures of the same performance dimension.

The paper proceeds as follows: First, we discuss the existing public administration theory on performance measurement. Then we elaborate on our choice of public service motivation, intrinsic motivation and job satisfaction as explanatory variables by clarifying the reasons for expecting positive associations with individual performance. This leads to a presentation of our expectations, methods and data. We then present and discuss the results, followed by a theoretical discussion (based on our empirical findings) of how public administration research can obtain consistent performance measures in the future.

2. Measuring Individual Performance

Much of the discussion about performance measurement implicitly or explicitly differentiates between objective and subjective measures (Bommer et al. 1995; Hoffman, Nathan, and Holden 1991). Objective measures are often seen as superior to subjective measures, and as more impartial and independent given that they are detached from the actors whose performance is being evaluated (Andrews, Boyne, and Walker 2006; Meier and O'Toole 2013a; 2013b). On the other hand, there can also be bias in objective measures; for instance, administrative data often rely on information reported

by the organizations to the register, and this information can also be manipulated. The usefulness of the distinction between objective and subjective measures has, however, been questioned (Schachter 2010; see also Brudney and England 1982; Parks 1984), making it necessary to theorize more explicitly about what contributes to measurement validity for performance measures.

Performance measures are subject to a number of measurement errors that can introduce biases in both descriptive analyses and in estimates of construct-level relationships. Focusing on the latter, it is relevant to distinguish between the assessment process and the methods we use to obtain information about the results of this assessment process. Collecting information about both performance and the explanatory variable in the same questionnaire or interview can mean that measurement errors in the two variables co-vary, inducing a bias in the estimate of their correlation (Sharma, Yetton, and Crawford 2009, 474). Below, we first discuss the assessment process and then the problems linked to using a common data source.

2.1 The Assessment Process

Measurement validity is expected to be higher when a measure is an indicator of a conceptually well-defined dimension of performance, because it is then known what should be measured and what should not be measured. There can, however, be a trade-off between consistency and comprehensiveness of the definition of performance. Students' exam marks can, for example, be a very consistent measure of their academic skills, but a less valid measure of total school performance. Accordingly, measures of performance based on administrative data can have low measurement validity if they fail to capture the performance dimensions that they are intended to measure (Andrews, Boyne, and Walker 2006). Given the technical limitations with regard to which indicators it is possible to collect (Boyne 2002), quantitative performance measures in particular tend to focus more specifically on one or a few dimensions of performance (Lazear and Gibbs 2009, 238–42). This may be a problem if it is not made

clear in the discussion of the performance assessment what has been left out. On the other extreme, performance measures in surveys are sometimes very vague and ill-defined (Meier and O'Toole 2013b) because they are too broad.

Measurement validity of performance measures is expected to depend on whether there is a process for verifying the accuracy of the assessment and whether there is external involvement in this process (Andrews, Boyne, and Walker 2006). The external involvement is particularly expected to attenuate biases in internal assessment, because people tend to be overly positive in their assessment of their own performance. There may also be leniency bias for managers, because giving poor ratings is uncomfortable (Lazear and Gibbs 2009, 247). It is not, however, self-evident that external assessments of performance are always more accurate than internal assessments. Internal stakeholders might have a better all-around understanding of the challenges facing their organization, and external stakeholders might have an incentive to underestimate achievements of public organizations or their judgments might be based on a particular stakeholder's perspective. Still, internal perceptions may also be based on partial views or even totally different understandings of what high performance is compared to criteria specified by legitimately elected politicians. On average, it is also plausible that external assessments are based on a higher level of awareness of relative performance. There are, in other words, several reasons for internal assessment of performance to diverge from external assessments. The ideal situation would be to have performance measures that give similar results both descriptively and in analyses of relationships between performance and other key concepts.

There is hardly any process of verification for self-reported performance in surveys apart from situations where outliers and protest answers are omitted. In contrast, regarding student grades for written exams there are well-established procedures for assessing the academic skills of students according to standardized criteria. The actor whose performance is being evaluated (directly or

indirectly) has a strong interest in having performance appear to be higher than it is. Podsakoff, MacKenzie and Podsakoff (2012, 561) argue that the more serious the social consequences of a particular response, the stronger the respondent's desire to provide a socially acceptable response will be. High performance is normally the most socially acceptable response, resulting in the most positive consequences for the respondents. This results in overestimation of performance, and if there is also common source bias (as discussed below), it can also give biased estimates of relationships between variables.

2.2 Common (Data) Source Bias as a Potential Problem

Common source bias is related to, but not overlapping with, the aspects of measurement validity discussed above. Common source bias can arise when information about dependent and explanatory variables comes from the same source, and can be defined as “systematic error variance shared among variables measured with and introduced as a function of the same method and/or source” (Richardson, Simmering, and Sturman 2009, 763). The same source can, for instance, mean the same data source, the same individual or the same organizational context. Below, we focus on the bias arising from measuring dependent and explanatory variables using the same data source (common data source bias), but we will return to the fact that assessments from the same individual (in different data sources) can also lead to bias. While common data source bias could be seen as a binary characteristic (same data source/not same data source), the general potential for common source bias can be seen as a continuum going from maximum risk of common source bias, with perceptual variables measured for the same individual in the same data source, to a lower risk of common source bias if, for example, the same individual answers questions about the dependent and independent variables at different times and in different types of data sources. If information about the two variables comes from the same organization (for example analyzing the association between motivation reported in a questionnaire to

each employee and sickness absence registered by an administrative employee), common source bias is still possible (for example because the same organizational culture makes them answer in a way that makes the organization look good), but we would expect the risk of common source bias to be lower.

If a given survey provides information about both explanatory variables (e.g. motivation) and respondents' assessment of their own performance, these measures may share the same bias, which can pose a serious problem (Podsakoff and Organ 1986; Podsakoff et al. 2003). Richardson, Simmering, and Sturman (2009) argue that common source bias can either inflate or attenuate relationships, but that it is most commonly expected to cause inflation. The potential for bias is especially high when both the dependent and explanatory variables concern socially sensitive issues, because some respondents will then answer in a socially desirable manner instead of expressing their true feelings. In cases where the same social desirability bias exists in both the dependent and explanatory variables, this alone can produce a significant correlation. Methods exist to deal with this (Kim and Kim 2013), but they are contested (Podsakoff and Organ 1986; Favero and Bullock 2014). Social desirability bias may also generate an attenuation bias if variables that are not subject to social desirability bias – for instance gender – are associated with variables that are subject to social desirability bias (Spector 2009). Another type of common source bias is caused by informants' predilection towards giving similar responses to separate survey questions, using only the higher or lower points on a response scale (Andrews, Boyne, and Walker 2006). This is called response set (Couch and Keniston 1960).

Meier and O'Toole (2013a) have shown that the risk of generating false results due to common data source bias is considerable for organizational performance. The difference in the t-values between analyses using either subjective or objective performance measures indicates that 27% of the correlations using measures with potential common data source bias are spurious. However, the objective measure used was the respondents' perception of the organization's TAKS score.ⁱ This is a

measure of perception of the position of the organization on a performance measurement scale rather than a performance measure in itself, and consistency between findings with different performance measures can only be expected if the same criteria of performance are being evaluated. Furthermore, there is no theoretical argument for the causal relationships investigated, and especially if many potential associations are analyzed, spuriousness (including common source bias) may not be the reason behind the variation in associations between analyses using different performance measures. Random variation will inevitably result in some type I errors (incorrect rejection of a true null hypothesis). These issues do not alter the fact that there is a difference between measures of performance applied, but they may inflate the number of spurious associations.

Thus, it is relevant to compare different performance measures for relationships with a theoretically informed expectation regarding causality between the explanatory and dependent variables. The next section discusses three variables for which there are theoretical arguments for expecting an effect on individual performance.

3. Theoretical Arguments for Associations between Public Service Motivation, Intrinsic Motivation, Job Satisfaction and Individual Performance

The criteria behind our choice to investigate public service motivation, intrinsic motivation, and job satisfaction are 1) data availability, 2) that the constructs are relevant to public administration as a discipline and 3) that there are theoretical arguments for a possible causal effect on performance.

Public service motivation is an altruistic type of motivation, as it can be seen as the orientation to do good for others and society through service provision (Perry, Hondeghem, and Wise 2010). One of the central and earliest statements in the public service motivation literature is that public service motivation is expected to affect individual performance positively (Perry and Wise 1990). The causal argument is that more public service-motivated individuals will exert a higher work effort, and that this

ultimately leads to higher performance (Perry, Hondeghem, and Wise 2010). Recent theoretical developments suggest that the person-job fit and person-environment fit mediate the association between public service motivation and performance (Bright 2007; Christensen and Wright 2011; Kristoff-Brown, Zimmerman, and Johnson 2005). A systematic review of the empirical studies reports positive associations between public service motivation and subjective performance indicators, while none report negative associations (Brewer and Selden 2000; Petrovsky and Ritz 2014; Ritz 2009). Positive associations are also found between public service motivation and self-reported performance ratings by supervisors (Camilleri and van der Heijden 2007; Naff and Crum 1999), and positive associations are found for different behavioral outcomes, which could mediate a positive association between public service motivation and performance. Examples include whistleblowing (Brewer and Selden 2000), organizational citizenship (Pandey, Wright, and Moynihan 2008), performance information use (Moynihan and Pandey 2010, 859), and work effort (Frank and Lewis 2004; Pedersen 2014). Additionally, an experimental study made in a real-life setting documents a positive relationship between public service motivation and objective performance indicators (Bellé 2013). Hence, there is good reason to expect a causal public service motivation-performance relationship.

Intrinsic motivation is defined as the interest in or enjoyment of work for its own sake (Le Grand 2003, 53). A key difference between intrinsic motivation and public service motivation is that while the latter is directed towards contributing to others and society, the potential energy in intrinsic motivation only relates to the individual in question. Self Determination Theory (e.g. Deci and Ryan 1985; Gagné and Deci 2005) argues that intrinsic motivation maximizes performance, and several studies recommend that jobs be expanded to enhance intrinsic motivation (e.g. Lawler and Hall 1970; Lazear and Gibbs 2009, 192–3). The intrinsic motivation-performance relationship is also an integral part of Motivational Crowding Theory, which argues that intrinsic motivation can be “crowded out”

when command/incentive systems undermine employees' perceptions of competence, autonomy, and relatedness, while "crowding in" denotes cases where intrinsic motivation is increased by sustaining these basic needs (Jacobsen, Hvitved, and Andersen 2014). Numerous empirical studies have tested the Motivation Crowding Theory. Most studies find that payments and command-and-control regulations, which are perceived as controlling, reduce intrinsic motivation, work effort and performance (e.g. Andersen and Pallesen 2008; Bertelli 2006; Bohnet, Frey, and Huck 2001; Dickinson and Villeval 2008; Falk and Kosfeld 2006; Frey and Jegen 2001; Jacobsen, Hvitfeld, and Andersen 2013). In sum, there is reason to expect an association between intrinsic motivation and performance.

Job satisfaction is defined as a pleasant emotional state resulting from outcomes from a job that meets or exceeds the employees' expectations. Job satisfaction is a function of the perceived relationship between what one wants from one's job and what one perceives it as offering or entailing (Locke 1969, 316). The proposition of a causal relationship between job satisfaction and job performance goes back to the Hawthorne studies, but the nature of the relationship and the empirical findings remain contested. The basic argument is that there is an association between attitudes and behavior (Fishbein and Ajzen 1975), and that attitudes toward the job affect behaviors on the job. However, causal models that propose inverse, bidirectional or spurious relationships, or relationships mediated by intervening variables are present in the literature (see Judge et al. 2001 for a review). A meta-analysis of the empirical studies in the field, however, suggests that there is a direct, moderate relationship between job satisfaction and job performance (Judge et al. 2001).

In sum, there are theoretical arguments for expecting associations between public service motivation, intrinsic motivation, job satisfaction and individual performance. Based on this, we are ready to propose the expectations to be tested.

4. Expectations

We compare four measures of the same performance dimension (Danish teachers' contributions to their students' academic skills). The first measure is teachers' self-reported contribution to students' learning as reported by the teacher in the questionnaire, which also measures the explanatory variables. The second is marks for the year's work for the teacher's students in the relevant subject (given by the teacher alone based on application of standardized criteria on the students' total performance during the year). The third is oral exam marks for the teacher's students in the relevant subject (questions asked by the teacher, assessment done by one external examiner and the teacher), and the fourth is written exam marks for the teacher's students in the relevant subject (answers to national questions assessed by one external examiner and the teacher).

The same teacher participates in all performance assessments, and the differences thus relate only to potential common data source bias, degree of external involvement in the assessment and standardization of measurement criteria. All four performance measures can have common source bias related to the fact that the same individual participated in the performance assessment: The same teacher filled out the survey questions measuring the explanatory variables on the one hand, and on the other hand answered questions about self-reported contributions to student academic skills, gave marks for the year's work and participated in grading during oral and written exams.

As mentioned, some individuals might tend to provide socially desirable answers, implying that the same respondents might exaggerate the levels of performance, public service motivation, intrinsic motivation and job satisfaction. This suggests that a potential bias will be towards a too-strong positive association if the level of common source bias is high.

The common source literature implies that self-reported contribution to student learning is especially vulnerable to common source bias (due to the common data source). This is expected to

mean higher correlations for self-reported contribution. The arguments presented by Andrews et al. (2006) imply that stronger processes of verification and standardizations reduce errors in the performance measurement (and thus reduce problems with correlated errors). When teachers report the extent to which they perceive that they contribute to improving the students' academic skills, the criteria for the assessment are not standardized. In contrast, the same teachers' evaluations of the students' skills through marks for the year's work are based on more standardized criteria and use the same grading scale.

Marks for the year's work and marks in oral exams differ by virtue of the participation of an external examiner (in oral exams), and oral and written exams further differ in relation to teacher participation in the exam situation. Written exams have standardized questions (formulated on the national level), while the questions in oral exams are not standardized. It can therefore be argued that written exams have a higher level of standardization than both oral exams and marks for the year's work, although the same measurement scale (grades) is used (meaning the performance measure itself is equally institutionalized).

The systematic differences between the performance measures allow us to analyze what affects the associations between performance measures and perceptual explanatory variables. Hypothesis 1 below concerns the use of the same data source to measure dependent and explanatory variables by comparing self-reported contributions and marks for the year's work. Hypothesis 2 concerns the effect of including an external examiner by comparing marks for the year's work and oral exams, and Hypothesis 3 concerns standardization of criteria (the difference between oral and written exams and also a more general difference between self-reported contribution and the three types of marks). Given that we expect common source bias to increase the correlations, we expect stronger correlations the more vulnerable the performance measure is to common source bias.

H1: Associations between perceptual explanatory variables and measures of individual performance are stronger if explanatory variables and dependent variables are measured in the same survey.

H2: Associations between perceptual explanatory variables and measures of individual performance are weaker if the process for verifying the accuracy of the performance assessment is external to the actors whose performance is being evaluated.

H3: Associations between perceptual explanatory variables and measures of individual performance are weaker if the process for assessing performance is standardized.

5. Methods

Investigating the associations between public service motivation, intrinsic motivation and job satisfaction on the one hand and individual performance, defined as the teacher's contribution to their students' academic skills, on the other hand involves important methodological challenges. One is selection. For example, teachers may be assigned to classes in a non-random manner. If teachers with high motivation and/or strong qualifications select into teaching more able students, this would create an upward bias in the results produced when looking at exam marks. If these teachers were assigned to teach classes with disadvantaged students, it could create a downward bias. The problem of non-random sorting on both observable and unobservable characteristics is addressed by including student fixed effects in the empirical models. This is possible as each student has examination marks in multiple subjects and different teachers in different subjects. The student fixed effects control for unobservable student characteristics, which are constant across the subjects, as it is the variation between subjects for each student that is used in the estimations. Thus, the effect of the explanatory variables – teacher public service motivation, teacher intrinsic motivation and teacher job satisfaction – on student marks is estimated by comparing the relative marks in different subjects for each student with the relative public service motivation, intrinsic motivation and job satisfaction of the student's

teachers in these subjects. Thus, no between-student variation is used. In this way, the non-random sorting of students and teachers into classes is taken into account. It also means that we analyze variations in teachers' in-role performance, as their extra-role behavior is not linked to a specific subject. Extra-role performance, therefore, does not vary between subjects for the same students. The same identification strategy is used in Clotfelter, Ladd and Vigdor (2010). We estimate models of the form:

$$y_{ijsk} = \alpha + T_{ijsk}\beta + \gamma_s + X_{ijsk}\delta + \mu_i + u_{ijsk} , \quad (1)$$

where y_{ijsk} is the mark of student i in subject s taught by teacher j in school k , T_{ijsk} is a vector of teacher characteristics, γ_s are subject fixed effects, X_{ijsk} is a vector of interaction terms between student characteristics, teacher characteristics and subject dummy variables, μ_i are student fixed effects, u_{ijsk} is the error term, α is the constant term, and β and δ are vectors of parameters. All subjects are taught in the same basic classes, and student fixed effects will therefore also take into account class-specific characteristics, such as class size and peer-group characteristics. Student fixed effects also control for variables at higher levels (e.g. schools and municipalities).

We have data for teachers in grades 7–9. In principle, we could include teacher characteristics for each grade as separate variables, but high correlation in these variables would be present because many students have the same teacher in a given subject for all three years. Instead, we only use the characteristics of the 9th grade teachers but include a dummy variable for change of teachers from grades 8 to 9 and interaction terms between this variable and all teacher characteristics. In robustness checks, we have replaced 9th grade teacher characteristics with average characteristics of teachers in grades 7–9, and this gives basically the same results.

In some of the estimations, the dependent variable is teacher self-reported contribution to student academic skills, which is measured at teacher level (through self-reporting). In such cases, we estimate models of the form:

$$y_{jk} = \alpha + T_{jk}\beta + Z_{jk}\theta + \mu_k + u_{jk} , \quad (2)$$

where y_{jk} is self-reported contribution to student academic skills of teacher j at school k , T_{jk} is a vector of other teacher characteristics, μ_k are school fixed effects, and u_{jk} is the error term. Z_{jk} is a vector of variables (one for each subject except for a reference) measuring the extent to which the teacher has classes in the different subjects. For instance, for a teacher who only teaches math classes, the variable for math will be equal to unity and the variables for other subjects will all be equal to zero, whereas for a teacher who teaches an equal number of students in math and physics classes, but no other subjects, the variables for math and physics will both be equal to 0.5, and the variables for other subjects will be zero. The variables in Z_{jk} are constructed by “collapsing” student-subject observations of subject dummies on teachers.

Similarly, we have constructed average student characteristics for each teacher by “collapsing” variables for individual student characteristics on teachers. Such variables may be relevant, since we are not able to control for student fixed effects in the self-reported contribution to student academic skills models. However, these variables are not included in the self-reported performance models we present for several reasons. Firstly, the teacher’s self-reported performance is not specifically related to the students we observe in our data, but to other cohorts as well. Secondly, we take account of school fixed effects, and the variation in average student characteristics between classes within schools is rather small. When included in the model, the variables for average student characteristics of the teacher’s classes are largely insignificant, and the estimated effects of the perceptual explanatory

variables (teacher motivation and job satisfaction) do not depend significantly on whether we control for average student characteristics. In both Model (1) and Model (2) we estimate robust standard errors, taking account of clustering on schools. This is a highly aggregate level of clustering, which produces conservative standard errors.

6. Data and Institutional Setting

Compulsory education in Denmark runs to the 9th grade (about age 15). Public schools are free. Expenditure on public schools is financed by municipal taxes, primarily income tax, but extensive grants and equalization schemes eliminate the greater part of financial inequalities between municipalities. Private schools (including boarding schools) are heavily subsidized; on average, parents pay a tuition fee of about 15% of the costs. In the early grades, almost 90% of pupils are enrolled in public schools. In the 9th grade, this percentage is lower, about 75%, primarily because many choose to shift to boarding school in the 9th grade. At the end of the 9th grade, there is an examination in different subjects (see below for more details). Academic skills are measured at the end of compulsory education, and the written exams use national standards that are not influenced by interest organizations, users, school managers, etc. Evaluations of the year's work and oral exams are also based on national criteria, but these are nowhere near as standardized as those for the written exams, and the process for verifying the accuracy of the assessment is more strongly standardized for exams (especially written exams). We focus only on the in-role performance of each teacher, analyzing the students' marks in a given subject controlled statistically for the same students' marks in other subjects, as explained in more detail below.

The investigation is based on three sources of data. The first source is a survey of all teachers in 85 schools conducted from December 2010 through June 2011, which includes information on teacher qualifications, motivation and job satisfaction. Secondly, to link each individual teacher with each

individual student in each subject, we obtained information from these 85 schools on the distribution of students and teachers for classes in grades 7–9 for students who completed 9th grade in 2011. In this dataset, teachers are identified by name and initials, providing a link to the teacher questionnaires containing the same information, and each student is identified by a unique personal identification number. This enables us to link to administrative register information on students, which is the third source of data. The register data contains each individual student's 9th grade marks, personal characteristics (e.g. gender, age and ethnic background), and socioeconomic variables for parents (including education, income, family structure, labor market status and working experience). The teacher survey and data collection at schools were conducted at relatively large schools to maximize the number of teachers investigated for the given amount of resources available for data collection. The 221 largest public schools in Denmark were contacted, and 85 agreed to participate; non-participation decisions were mainly due to practical issues. Differences between participating and non-participating schools in terms of students' average marks and socioeconomic characteristics are very small. However, the 85 participating schools are of course not representative of all Danish schools, because they are larger. However, if the effects of teacher characteristics vary according to school size, the fact that we have selected large schools may be an advantage in terms of generalizability to other countries, given that Danish schools are generally smaller than schools in most other countries.

6.1 Dependent Variables

All four measures of performance concern the academic skills of the students.

Self-reported contribution to students' academic skills is a self-assessed measure of performance in the sense that it is the teachers themselves who make the assessment of the extent to which they contribute to student learning. However, this measure is indirect, as the wording of all the included items is directed towards the students' academic skills, i.e. the extent to which the teachers

feel they make an educational difference in the lives of their students. Hence, the teachers assess their performance indirectly through the assessment of their students' progress. The measurement statistics in this study suggest that the measurement validity and reliability are satisfactory. The Appendix (Table A1) presents the wording, the CFA and Cronbach's alpha.

It should be noted that we have tried to maximize measurement validity in the standard ways. The questions address behavior that is fairly observable and concrete, and we use several items to measure the concept. Validity and reliability are also tested in the standard ways (see Table A1 in the Appendix for details). The fact that we focus on individual performance can also increase measurement validity, because teachers can give a more precise assessment of their individual performance compared to performance at an aggregated level. For instance, a teacher probably knows whether they can get through to unmotivated students, whereas they most likely have less knowledge about what is going on in the other classrooms in their school. In line with this, the student fixed effect design (discussed in detail below) ensures that we only analyze the variation in performance linked to the teachers' in-role behavior in a given subject.

Student marks are assessed using three different measures: *marks for the year's work*, *oral exam marks* and *written exam marks*, which vary in their degree of external involvement in the assessment, but all of which are indirect in the sense that the teachers' performance is evaluated indirectly via assessments of their students' academic skills in the teacher's subject compared to the same students' performance in other subjects. Marks for the year's work are given once annually by the teacher, based on the student's classroom performance. Accordingly, there is no external censorship. In contrast, oral examinations are conducted by the teacher and an external examiner; hence, oral exam marks have a higher degree of objectivity (because the process for verifying their accuracy is more external to the teacher, whose performance is indirectly evaluated). Most written exams are graded by

the teacher and an external examiner, but in some subjects they are graded by two external examiners (and not by the teacher). External examiners are appointed by the Ministry of Children and Education. Written exam marks have a higher degree of external assessment than oral exam marks, since exam questions are determined centrally (by the Ministry), and the teacher is not part of the examination.

The analysis is based on marks in nine different subjects: Danish, Math, English, History, Science, Biology, Geography, Religion and Social Studies. The marks in written Danish used in the analysis are an average of three individual marks (reading, spelling and essays) and math marks are an average of two individual written marks. The exams in Danish and Math and the oral examinations in Science and English are mandatory. All students must take two additional exams: a written exam in Biology or Geography and a written exam in English, or an oral exam in History, Religion, or Social Studies. The Ministry of Children and Education decides the subjects in which students in different classes are examined. The reason why we use average marks in the case of multiple individual marks in the same subject (which is the case for written Danish and Math) is that we focus on teacher effects, and teacher characteristics are obviously constant within subjects for a given class. A 7-point scale is used for grading. In order to simplify the interpretation of the results, however, we use standardized marks, which have mean zero and standard deviation unity for each individual mark in a given subject in a given year (calculated for all of the students in the 85 schools, not just the estimation sample, which is restricted to observations with teacher information). Similarly, self-reported performance and the three perceptual explanatory variables discussed below are also standardized, rendering estimated coefficients comparable across models.

6.2 The Explanatory Variables

All perceptual explanatory variables were measured using survey questions validated in prior surveys. Intrinsic motivation is measured as factor scores from a confirmatory factor analysis. Both the

confirmatory factor analysis and Cronbach's alpha indicate satisfactory measurement validity and reliability (see Appendix, Table A2, which also shows the wording of the items). The survey questions regarding public service motivation were based on international and Danish surveys (Andersen and Pedersen 2012; Perry 1996), and the scores on the public service motivation dimensions were calculated using a confirmatory factor analysis, which together with the Cronbach's alphas for the dimensions indicate acceptable measurement validity and reliability, except for "attraction to public policy making" where the Cronbach's alpha is only 0.57. This is, however, a general problem in the public service motivation literature (e.g. Ritz 2011). Measurement statistics for public service motivation are shown in the Appendix, Tables A3 and A4. The measure of job satisfaction is based on a single item. The respondents are asked to assess the extent to which they are satisfied with their jobⁱⁱ on a scale from 0–10 (Hackman and Oldham 1975, 282; Jung 2014, 964). Summary statistics of dependent and explanatory variables used in the estimates are shown in Table A5 in the Appendix.

It should be noted that the survey data on public service motivation, intrinsic motivation and job satisfaction were collected prior to measuring the exam marks. Hence, the exam marks cannot influence the explanatory variables, which might otherwise be the case if teachers reported higher satisfaction with their jobs because their students achieved better results in their exams. Self-reported performance is measured in the same survey – and thus at the same time – as the explanatory variables.

7. Results

This section presents the simple correlations between the different performance measures, followed by estimates in student fixed effects models of the associations between the performance measures and the three explanatory variables. These results are discussed in the following section.

Table 1 shows the correlations between the four measures of students' academic skills at the teacher level, i.e. the marks are the average marks of the teacher's students. The correlations provide an

indication of the degree of consistency between the teacher's perceived positive influence on students' academic skills and the same students' more objective results in terms of marks for the year's work, oral exam marks and written exam marks. Self-reported performance has low correlations with the three grading variables and is not significantly correlated with the students' written exam scores. In contrast, there are much higher correlations between a teacher's students' average results, measured as the marks for the year's work, oral exam marks and written exam marks. Of these three correlations, the smallest is the one between marks for the year's work and oral exam marks. Reasons for this may be that results of oral examinations are more random, and that it is more difficult for teachers to evaluate students' class work in oral subjects (where students typically do not hand in many school papers during the year).

[TABLE 1 HERE]

Tables 2, 3, and 4 show the results of student fixed effects regressions of the four performance measures as dependent variables combined with the three explanatory variables. The presentation of results focuses on the estimated conditional correlations between the performance measures and the perceptual explanatory variables (teacher public service motivation, intrinsic motivation, and job satisfaction). The parameters of interest in Tables 2–4 are those shown in the first row, which are the conditional associations between the dependent variables and the perceptual explanatory variables if there was no change of teachers between grades 8 and 9. As discussed above, the explanatory variables regarding teacher characteristics and the self-reported performance measure are for 9th grade teachers. In the estimations, we control for whether there was a change of teachers from the 8th to the 9th grade in the subject. We also include interaction terms between this dummy variable and all teacher characteristics, since the effects of teacher characteristics may be expected to be greater if students did not experience teacher change. If the teachers change, there will be less time for their public service

motivation, intrinsic motivation and job satisfaction (along with other characteristics) to influence student performance. Only the “change of teacher” variable, the perceptual explanatory variable and their interaction term are shown in Tables 2–4, but the full models include a number of important control variables (see Tables A6–A8 in the Appendix).

Table 2 shows the associations between public service motivation and the four performance measures. The associations are positive and significant across all four measures. The public service motivation-self-reported contribution relationship (0.194) is notably stronger than the associations between public service motivation and the three measures of student marks, and it is also more significant (the t value is about twice as high, at 5.0 compared to 2.3–2.6 in the marks regressions). The size of the coefficient is smaller for written exams (comparisons between coefficient sizes are only meaningful because the performance measures are standardized).

[TABLE 2 HERE]

Table 3 shows the results of the estimation of the associations between intrinsic motivation and the four measures of performance. The associations are significant for self-reported contributions to students’ academic skills and marks for the year’s work, and again there is a marked difference between the self-reported measure (where the coefficient is 0.419; see Model 1 in Table 3) and the three measures using student marks. The significance of the associations is lower the higher the degree of external involvement in the assessment; i.e. lowest for Model 4 (written exam marks where the point estimate is even negative) followed by Models 3, 2 and 1. The association is insignificant for both oral and written exams, and the size of the coefficient is smaller for written exams.

[TABLE 3 HERE]

Table 4 shows the results when job satisfaction is the key explanatory variable. The associations are statistically significant for self-reported performance and marks for the year’s work, and there is a

marked difference in the size and significance of the associations between Models 1 and 2. Compared to the estimate of intrinsic motivation in the regression for marks for the year's work, the point estimate in the regression for oral exam marks is a little higher, but not significantly different. The estimate in the oral exam marks regression is not significantly different from zero at the 5% level, but is so at the 10% level. Again, the estimate in the model for written marks is very low and clearly not significant. Since each student does not receive exam marks in all subjects, the sample size is smaller in the oral and written exam marks regressions than in teacher self-reports and marks for the year's work regressions. Some of the differences in significance in Tables 2-4 between Models 1 and 2 on the one hand and Models 3 and 4 on the other hand could be thought, therefore, to be due to the different samples rather than the different dependent variables. However, we show that this is not the case in the robustness checks reported in Tables A9-A11 in the Appendix. Restricting the sample to the oral or written exam samples does not significantly affect the estimates for the parameters of interest for the regression with self-reported contribution and marks for the year's work. For instance, the estimated effect of public service motivation on self-reported performance is 0.23 and 0.21 for the oral and written exam samples, respectively, and thus very close to the estimate for the full sample reported in Table 2 (which is 0.19); the estimates of the effect of public service motivation on marks for the year's work are 0.030 and 0.043 for the oral and written exam samples, and these estimates are also close to the estimate for the full sample (0.031) in Table 2.

[TABLE 4 HERE]

As discussed in the Methods section, the self-reported performance models are estimated using teacher-level data and school fixed effects, whereas the models for the three types of marks are estimated using student-subject observations and student fixed effects. The considerable differences between the estimates in Model 1 and Models 2-4 in Tables 2-4 are not due to differences in estimation techniques.

When we “collapse” the student-subject observations of marks and explanatory variables on teachers and estimate school fixed effects models using these teacher-level data, we obtain parameter estimates of teacher motivation and job satisfaction that are slightly lower than the corresponding estimates in Tables 2-4.

Although the consistency of the associations varies a little between the four performance measures and public service motivation, intrinsic motivation and job satisfaction, the similarities are more interesting than the differences. The key similarities are (1) that all associations drop greatly between Model 1 (self-reported contribution) and Model 2 (year’s work), (2) that the associations are slightly stronger in Model 3 (oral exams) than in Model 4 (written exams), and (3) that the associations are quite similar for all performance measures using the same grading scale, with the exception of written exams in relation to intrinsic motivation and job satisfaction. This exception is the one key difference: There seem to be no statistically significant associations between intrinsic motivation and written exam marks (Table 3, Model 4) and between job satisfaction and written exam marks (Table 4, Model 4), while the conditional public service motivation-performance correlation remains statistically significant for all four performance measures (Table 2, Models 1–4). This may imply that only public service motivation has a true association with performance (in this sample), but it is also possible that written exams are a less complete measure of student academic skills than the other measures. Regardless, our results illustrate that the highest associations in analyses using information from the same data source may not be the most robust associations when compared with results based on performance data from other sources. In relation to the hypotheses, the similarities are, however, more interesting.

The strongest finding is that the results support Hypothesis 1, since all three analyses (Tables 2, 3, and 4) show much higher associations in Model 1 than in Models 2–4. This major drop in the size of

the coefficients (which are directly comparable, because both dependent and perceptual explanatory variables are standardized) happens when we move from a model with potential common data source bias to a model using separate data sources to measure the variables.

Hypothesis 2 expects weaker associations if the process for verifying the accuracy of the performance assessment is external to the actors whose performance is being evaluated. This implies that we would expect higher associations for marks for the year's work (Model 2) than for marks in oral exams (Model 3). However, this is not the case. Across Tables 2-4, no systematic drop in the size of the coefficient can be seen between Models 2 and 3. Instead, the size of the coefficient remains at a similar level. Thus, the degree of external verification in itself does not seem to be very important.

Hypothesis 3 addresses the standardization of the criteria in the assessment process, expecting associations between perceptual explanatory variables and individual performance measures to be weaker if the assessment process is based on more standardized criteria for what constitutes high performance. The analyses lend some support to this expectation. As we will discuss in depth below, once the grading scale is institutionalized, it does not seem to matter much for the associations what type of evaluation/exam we are talking about. The associations tend to be a little stronger for the year's work and oral exams than for written exams, and this is consistent with Hypothesis 2 in the sense that the assessment process for yearly marks and oral exams is less standardized than for written exams, where the questions are also standardized and where the teacher does not participate in the exam situation.

In sum, the analysis shows that common data source bias seems just as problematic for measures of individual performance as it is for measures of organizational performance. Additionally, our findings suggest that institutionalization of the assessment can be a way to ensure consistent performance measures. The next section discusses this theoretically.

8. Professional Norms as a Way to Institutionalize Performance Assessment

If we want to understand institutionalization of performance measures in professionalized public organizations, it is useful to draw upon both institutional theory and the sociology of professions (see Salomonsen and Andersen 2014). More specifically, we argue that the institutionalization of the grading scale is tied to the teacher profession and its professional norms, but that other types of institutions are also relevant.

Institutionalization can be regulative, normative, and/or cognitive (Scott 2007; 2008a). Normative institutionalization takes place if there are norms and informal rules that concern the use of the performance measure, and if morally authorized prescriptions for how to behave are the basis of legitimacy (Scott 2007, 52). These prescriptions are often determined and sanctioned by professions (DiMaggio and Powell 1991, 70), and Scott (2008a, 219) argues that “the professions in modern society have assumed leading roles in the creation and tending of institutions.” Professional institutionalization has an element of reflexivity and agency (Muzio, Brock, and Suddaby 2013, 709), and therefore performance assessment can be institutionalized through professional norms. The sociology of professions argues that a profession and the public can make an implicit bargain in which the profession guarantees high-standard services in exchange for status and income (Burau and Andersen 2014). The profession can guarantee delivery of these “high-standard” services by institutionalizing a number of professional norms that are sanctioned within the profession and often also internalized by the professionals. This is exactly what we see with the grading scale. When this performance measure tends to give consistent results, we argue that it is due to the institutionalization of the scale through professional norms. Outsiders are often unable to make evaluations in a professionalized field. For instance, it can be difficult for outsiders to the teaching profession to determine what mark a specific student should have in a given subject. Professional norms are

institutionalized as a combination of formal education and peer socialization on the job. The professional norms on grading are largely developed in practice on the job, and the elaborate censorship institutions—described in detail below—sustain detailed formal rules and procedures that back up peer-review processes, which ensure that the grading scales are standardized and institutionalized across the teaching community.

This mechanism might even be stronger for professions with more firm professional norms than Danish teachers, and it might also be possible for decision-makers to promote this institutionalization. Danish teachers have a medium level of professionalism and semi-firm professional norms (Andersen and Pedersen 2012), and the normative institutionalization of grades exists, but could be stronger. Scott (2008b, 225–227) has shown that professionals can also be the co-creators of regulative institutions, and there are both formal rules and professional norms governing the grading scale. Especially through external examiners, decision-makers can facilitate the spread and sanctioning of these rules and norms. Selected on the basis of their professional competence, external examiners are trained teachers and are thus both professionals and representatives for the national government. This illustrates how an institutionalization (of the Danish grading scale) can also include regulative institutionalization (compliance with existing formal rules ensured by the external examiners), although the dominant logic is normative (through professional institutionalization of norms). Combining these logics, external examiners may ensure uniform grading across schools, thereby ensuring consistent performance assessments, even for marks where they do not participate in the evaluation. They do participate in the oral exams (being present during the exam and agreeing with the teacher about the mark), and teacher and examiner discuss the marks (often during a phone call) for the written exams. The point is that this can contribute to institutionalizing the grading scale as it is an iterative process taking place in grading

situation after grading situation, making the teachers follow the rules and norms even when they do their grading alone.

In addition, most citizens have been students at one point and hence have received grades. The scale is therefore taken for granted and has obtained a hegemonic status, where actors internal as well as external to the educational sector will have a hard time coming up with alternatives. There is also an understanding of what different grades mean, and this understanding is reflected in common beliefs and grounded in cultural support (Scott 2008a, 429). Thus, the grading scale as a performance measure can be seen as being institutionalized through regulative, normative and cognitive institutions (Scott 2007; 2008a), but we would still argue that it is primarily the normative institutionalization based on the teachers' profession that secures consistency in the actual, individual assessment of students.

Some of this logic may be transferable to performance measurement in other public organizations: If we could base our performance measures more on professional norms, our assessment might be more consistent, but it demands that the professions are willing to incorporate the legitimately decided goal of public organizations into their professional norms. Gaus (1936, 39–40) expressed this very precisely when he argued that “the responsibility of the civil servant to the standards of his profession, in so far as those standards make for the public interest, may be given official recognition.”

9. Conclusion

The ambition of this article was to empirically investigate and theoretically discuss how we obtain consistent performance measures, and our conclusions include both specific empirical findings and more general theoretical arguments. In line with other recent studies (Meier and O'Toole 2013a; 2013b), our results show that the associations between the same performance dimension and three perceptual explanatory variables are very different, depending on whether or not performance is measured using the same data source as the explanatory variables. This implies that future studies of

the relationship between perceptual variables and performance should try to avoid research designs with a high risk of common source bias, by using data from independent data sources to measure the dependent variable and the explanatory variables (see also Favero and Bullock 2014).

The other key empirical result is that performance measures with varying degrees of external assessment are surprisingly consistent – and we argue that the reason for this is that the scale employed is heavily institutionalized according to professional norms. The teachers assess their own performance indirectly through assessments of their students' performance according to a grading scale that is institutionalized by professional norms backed by regulative institutions. The consistency of the institutionalized measures indicates that there is potential for using professional norms more generally to ensure the consistency of performance assessments, and we argue that if professions are willing to incorporate the legitimately decided goals of public organizations into their professional norms, an institutionalization of performance assessment through professional norms could be highly beneficial. In Gaus' (1936, 39-40) words, "the standards and ideals of individual civil servants" can increase the consistency of performance assessment, because they serve as an "inner check" on the performance assessment. If this can find empirical support, it may be time to bring professionalism back to the core of public administration theory. Furthermore, autonomous professional norms can serve to insulate some public employees from efficiency-undermining political pressures (Miller 2000, 307).

In sum, our results suggest that the consistency of performance measures is not only a methodological problem. In addition to avoiding the use of a common data source to measure both dependent and independent variables, we argue that how institutionalized the assessment is can also be important. The investigated performance measures are institutionalized through professional norms and supported by external examiners (who have both professional and regulative authority), and this seems to make the teachers use the grading scale in a similar way. The fact that the performance measures

tend also to be consistent when no external verification is present suggests that it might be beneficial to base more performance assessments on professional norms.

References

- Andersen, Lotte Bøgh, and Lene Holm Pedersen. 2012. "Public Service Motivation and Professionalism." *International Journal of Public Administration* 35:46–57.
- Andersen, Lotte Bøgh, and Thomas Pallesen. 2008. "'Not just for the money?' How Financial Incentives Affect the Number of Publications at Danish Research Institutions." *International Public Management Journal* 11:28-47.
- Andersen, Lotte Bøgh, Eskil Heinesen, and Lene Holm Pedersen. 2014. "How does Public Service Motivation among Teachers Affect Student Performance in Schools?" *Journal of Public Administration Research and Theory* 24:651-671.
- Andrews, Rhys, George A. Boyne, and Richard M. Walker. 2006. "Subjective and Objective Measures of Organizational Performance: An Empirical Exploration." In *Public Service Performance: Perspectives on Measurement and Management*, edited by George A. Boyne, Kenneth J. Meier, Laurence J. O'Toole, Jr., and Richard M. Walker, 14–34. New York, NY: Cambridge University Press.
- Andrews, Rhys, George A. Boyne, M. Jae Moon, and Richard M. Walker. 2010. "Assessing Organizational Performance: Exploring Differences between Internal and External Measures." *International Public Management Journal* 13:105–29.
- Bellé, Nicola. 2013. "Experimental Evidence on the Relationship between Public Service Motivation and Job Performance." *Public Administration Review* 73:143–53.

- Bertelli, A.M. 2006. "Motivation Crowding and the Federal Civil Servant: Evidence from the U.S. Internal Revenue Service." *International Public Management Journal* 9:3–23.
- Bohnet, Iris, Bruno S. Frey, and Steffen Huck. 2001. "More Order with Less Law: On Contract Enforcement, Trust, and Crowding." *The American Political Science Review* 95:131–44.
- Bommer, William H., Jonathan L. Johnson, Gregory A. Rich, Philip M. Podsakoff, and Scott B. Mackenzie. 1995. "On the Interchangeability of Objective and Subjective Measures of Employee Performance: A Meta-Analysis." *Personnel Psychology* 48:587–605.
- Boyne, George A. 2002. "Concepts and Indicators of Local Authority Performance: An Evaluation of the Statutory Framework in England and Wales." *Public Money and Management* 22:17–24.
- Boyne, George A. 2003a. "Sources of Public Service Improvement: A Critical Review and Research Agenda." *Journal of Public Administration Research and Theory* 13:367–94.
- Boyne, George A. 2003b. "What is Public Service Improvement?" *Public Administration* 81:211–27.
- Brewer, Gene A. 2008. "Employee and Organizational Performance." In *Motivation in Public Management: The Call of Public Service*, edited by James L. Perry and Annie Hondeghem, 136–56. Oxford: Oxford University Press.
- Brewer, Gene A., and Sally Coleman Selden. 2000. "Why Elephants Gallop: Assessing and Predicting Organizational Performance in Federal Agencies." *Journal of Public Administration Research and Theory* 4:685–711.
- Bright, Leonard. 2007. "Does Person–Organization Fit Mediate the Relationship between Public Service Motivation and the Job Performance of Public Employees?" *Review of Public Personnel Administration* 27:361–79.
- Brudney, Jeffrey L., and Robert E. England. 1982. "Urban Policy Making and Subjective Service Evaluations: Are They Compatible?" *Public Administration Review* 42: 127-135.

- Burau, Viola & Lotte Bøgh Andersen. 2014. "Professions and professionals. capturing the changing role of expertise through theoretical triangulation" *American Journal of Economics and Sociology* 73 (1) 264-293.
- Callahan, J.S., A. L. Brownlee, M.D. Brtek, and H.L. Tosi. 2003. Examining the unique effects of multiple motivational sources on task performance. *Journal of Applied Social Psychology* 33:2515–35.
- Camilleri, E., and B.I.J.M. van der Heijden. 2007. Organizational commitment, public service motivation, and performance within the public sector. *Public Performance & Management Review* 31:241–74.
- Christensen, Robert K., and Bradley E. Wright. 2011. The effects of public service motivation on job choice decisions: Disentangling the contributions of person–organization fit and person–job fit. *Journal of Public Administration Research and Theory* 21:723–43.
- Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2010. Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources* 45:655–81.
- Couch, Arthur, and Kenneth Keniston. 1960. Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology* 60:151–74.
- Deci, Edward L., and Richard M. Ryan. 1985. *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.
- Dickinson, David, and Marie-Claire Villeval. 2008. Does monitoring decrease work effort? The complementarity between agency and crowding-out theories. *Games and Economic Behavior* 63:56–76.

- DiMaggio, P.J. and W.W. Powell. 1991. The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. In W.W. Powell, & P.J. DiMaggio (Eds.), *The new institutionalism in organizational analysis* (pp. 63-82). Chicago: University of Chicago Press.
- Falk, Armin, and Michael Kosfeld. 2006. The hidden costs of control. *American Economic Review* 96:1611–30.
- Favero, Nathan, and Justin B. Bullock. 2014. How (Not) to Solve the Problem: An Evaluation of Scholarly Responses to Common Source Bias. *Journal of Public Administration Research and Theory*. Advance Accesss. Published May 24, 2014.
- Fishbein, Martin, and Icek Ajzen. 1975. Belief, attitude, intention, and behavior: An introduction to theory and research. Reading, MA: Addison-Wesley.
- Frank, Sue A., and Gregory B. Lewis. 2004. Government employees: Working hard or hardly working? *The American Review of Public Administration* 34:36–51.
- Frey, Bruno, and Reto Jegen. 2001. Motivation crowding theory. *Journal of Economic Surveys* 15:589–611.
- Gagné, Marylène, and Edward L. Deci. 2005. Self-determination theory and work motivation. *Journal of Organizational Behavior* 26:331–62.
- Gaus, J.M. 1936. The responsibility of public administration. In J.M. Gaus, L.D. White, & M.E. Dimock (Eds.), *The Frontiers of Public Administration* (pp. 39-40). Chicago: University of Chicago Press.
- Hackman, Richard J., and Greg R. Oldham. 1975. Development of the job diagnostic survey. *Journal of Applied Psychology* 60:159–70.

- Hoffman, Calvin C., Barry R. Nathan, and Lisa M. Holden. 1991. A comparison of Validation criteria: Objective versus subjective performance measures and self- versus supervisor ratings. *Personnel Psychology* 44:601–19.
- Jacobsen, Christian Bøtcher, Johan Hvitved, and Lotte Bøgh Andersen. 2014. Command and motivation: How the perception of external interventions relates to intrinsic motivation and public service motivation. *Public Administration* 92 (4): 790–806.
- Judge, Timothy A., Carl J. Thoresen, Joyce E. Bono, and Gregory K. Patton. 2001. The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological Bulletin* 127:376–407.
- Jung, Chan Su. 2014. Organizational goal ambiguity and job satisfaction in the public sector. *Journal of Public Administration Research and Theory*. 24 (4): 955-981
- Kim, Seung Hyun, and Sangmook Kim. 2013 National Culture and Social Desirability Bias in Measuring Public Service Motivation *Administration & Society* Published online before print September 4, 2013, doi: 10.1177/0095399713498749.
- Kristof-Brown, Amy L., Ryan D. Zimmerman, and Erin C. Johnson. 2005. Consequences of individuals' fit at work: A meta-analysis of person–job, person–organization, person–group, and person–supervisor fit. *Personnel Psychology* 58:281–342.
- Lawler, Edward E., and Douglas T. Hall. 1970. Relationship of job characteristics to job involvement, satisfaction, and intrinsic motivation. *Journal of Applied Psychology* 54:305–12.
- Lazear, Edward P., and Michael Gibbs. 2009. *Personnel economics in practice*, 2nd ed., Hoboken, NJ: John Wiley & Sons.
- Le Grand, Julian. 2003. *Motivation, agency, and public policy: Of knights & knaves, pawns & queens*, Oxford: Oxford University Press.

- Locke, Edwin A. 1969. What is job satisfaction?, *Organizational Behavior and Human Performance* 4: 309–36.
- Meier, Kenneth J., and Laurence J. O'Toole, Jr., George A. Boyne, and Richard M. Walker. 2007. Strategic management and the performance of public organizations: Testing venerable ideas against recent theories. *Journal of Public Administration Research and Theory* 17:357–77.
- Meier, Kenneth J., and Laurence J. O'Toole. 2013a. Subjective organizational performance and measurement error: Common source bias and spurious relationships. *Journal of Public Administration Research and Theory* 23: 429-456.
- Meier, Kenneth J., and Laurence J. O'Toole. 2013b. I Think (I Am Doing Well), Therefore I Am: Assessing the Validity of Administrators' Self-assessments of Performance. *International Public Management Journal* 16:1–27.
- Miller, G. 2000. Above politics: Credible commitment and efficiency in the design of public agencies. *Journal of Public Administration Research and Theory-PART, 10, 2*, 289–327.
- Moynihan, Donald P., and Sanjay K. Pandey. 2010. The big question for performance management: Why do managers use performance information? *Journal of Public Administration Research and Theory* 20:849–66.
- Muzio, D, D.M Brock and R. Suddaby. 2013. Professions and Institutional Change: Towards and Institutional Sociology of the Professions. *Journal of Management Studies*, 50, 5, 699-721.
- Naff, Katherine C., and John Crum. 1999. Working for America: Does public service motivation make a difference? *Review of Public Personnel Administration* 19:5–16.
- Pandey, Sanjay K., Bradley E. Wright, and Donald P. Moynihan. 2008. Public Service Motivation and inter-organizational citizenship behavior: Testing a preliminary model. *International Public Management Journal* 11:89–108.

- Parks, Roger B. 1984. Linking Objective and Subjective Measures of Performance. *Public Administration Review*. 44: 118-127
- Pedersen, Lene Holm. 2014. Committed to the public interest? Motivation and behavioural outcomes among local councilors. *Public Administration*. 92 (4): 886–901.
- Perry, James L. 1996. Measuring public service motivation: An assessment of construct reliability and validity. *Journal of Public Administration Research and Theory*. 6:5–24.
- Perry, James L., and Lois R. Wise. 1990. The motivational bases of public service. *Public Administration Review* 50:367–73.
- Perry, James L., Annie Hondeghem, and Lois R. Wise. 2010. Revisiting the motivational bases of public service: Twenty years of research and an agenda for the future. *Public Administration Review* 70:681–90.
- Petrovsky, Nicolai and Adrian Ritz. 2014. Public service motivation and performance: a critical perspective. *Evidence-based HRM: a Global Forum for Empirical Scholarship* 2: 57 – 79.
- Podsakoff, Philip M., and Dennis W. Organ. 1986. Self-reports in organizational research: Problems and prospects. *Journal of Management* 12:531–44.
- Podsakoff, Philip M., Scott B. MacKenzie, and Nathan P. Podsakoff. 2012. Sources of Method Bias in Social Science Research and Recommendations on How to Control It. *Annual Review of Psychology* 63: 539-569.
- Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff. 2003. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology* 88:879–903.

- Richardson, Hettie A., Marcia J. Simmering, and Michael C. Sturman. 2009. A tale of three perspectives: Examining post hoc statistical techniques for detection and correction of common method variance. *Organizational Research Methods* 12:762–800.
- Ritz, Adrian. 2009. Public service motivation and organizational performance in Swiss federal government. *International Review of Administrative Sciences* 75:53–78.
- Ritz, Adrian. 2011. Attraction to public policy-making: A qualitative inquiry into improvements in PSM measurement. *Public Administration* 89:1128–47.
- Salomonsen, Heidi & Lotte Bøgh Andersen. 2014. Institutional change and professional norms: A study of grading behaviour in higher education. *International Journal of Public Administration* 37 (10): 611-624.
- Schachter, Hindy Lauer. 2010. Objective and Subjective Performance Measures: A Note on Terminology. *Administration & Society* 42:550–567.
- Scott, R.W. 2007. *Institutions and organizations* (3rd ed.). London: Sage.
- Scott, R.W. 2008a. Approaching adulthood: The maturing of institutional theory. *Theory and Society*, 37, 5, 427–442.
- Scott, R.W. 2008b. Lords of the dance: Professionals as institutional agents. *Organization Studies*, 29, 2, 219–238.
- Sharma, Rajeev, Philip Yetton and Jeff Crawford. 2009. Estimating the Effect of Common Method Variance: The Method-Method Pair Technique With An Illustration From TAM Research, in *MIS Quarterly* 33 (3): 473-490.
- Spector, Paul E. 2009. Method variance in organizational research: Truth or urban legend. *Organizational Research Methods* 9:221–32.

Vandenabeele, Wouter. 2009. The mediating effect of job satisfaction and organizational commitment on self-reported performance: More robust evidence of the PSM–performance relationship.

International Review of Administrative Sciences 75:11–34.

Williams, Larry J., and Stella E. Anderson. 1991. Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management* 17:601–17.

Table 1**Correlations between the different performance measures**

	Self-reported performance	Marks for the year's work	Oral exam marks
Marks for the year's work	0.2005 p < .00005 n = 747		
Oral exam marks	0.2421 p < .00005 n = 554	0.5595 p < .00005 n = 554	
Written exam marks	0.0524 p = .2387 n = 508	0.7400 p < .00005 n = 508	0.6884 p < .00005 n = 245

Note: Units of analysis are those of the teachers. The number of valid observations varies, because there are oral exams in some subjects, written exams in other subjects and both oral and written exams in a third category of subjects. The correlations between the three types of marks are based on averaging the relevant student-subject observations over teachers; for instance, when calculating the correlation between teachers' average marks for the year's work and average written exam marks, we only use observations for marks for the year's work in subjects with written exams and only for the students who actually took the written exams.

Table 2**Public Service Motivation and Performance. School and Student Fixed Effects Regressions**

	(1)	(2)	(3)	(4)
	Self-reported contribution	Year's work	Oral exams	Written exams
Teacher public service motivation	0.194*** (0.039)	0.031** (0.012)	0.043* (0.019)	0.029* (0.012)
Teacher public service motivation × change of teachers		−0.005 (0.019)	−0.003 (0.027)	−0.041 (0.022)
Change of teachers		−0.026 (0.028)	0.006 (0.063)	−0.026 (0.053)
Fixed effects	Schools	Students	Students	Students
Observations				
Students × subjects		39056	15858	13600
Students		5679	5546	5412
Teachers	747	747	554	508
Classes	281	281	277	269
Schools	85	85	85	85
Adj. R^2	0.093	0.673	0.505	0.648
R^2 (within)	0.095	0.026	0.023	0.068

Note: The full models can be seen in Table A6 in the Appendix. Robust standard errors clustered on schools in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$. The units of analysis are teachers (Model 1) or individual students in a given subject taught by a given teacher (Models 2-4). The adjusted R^2 statistic is from OLS regressions with dummies for schools (Model 1) or students (Models 2-4). The R^2 (within) statistic is R^2 from regressions on the within-school (Model 1) or within-student (Models 2-4) transformed variables.

Table 3**Intrinsic Motivation and Performance. School and Student Fixed Effect Regressions**

	(1)	(2)	(3)	(4)
	Self-reported contribution	Year's work	Oral exams	Written exams
Teacher intrinsic motivation	0.419*** (0.045)	0.027* (0.012)	0.025 (0.015)	−0.008 (0.013)
Teacher intrinsic motivation × change		0.003 (0.020)	−0.031 (0.026)	−0.010 (0.020)
Change of teachers		−0.023 (0.028)	−0.005 (0.060)	−0.030 (0.052)
Fixed effects	Schools	Students	Students	Students
Observations	747	39056	15858	13600
Adj. R^2	0.231	0.673	0.503	0.631
R^2 (within)	0.232	0.025	0.021	0.067

Note: The full models can be seen in Table A7 in the Appendix. Robust standard errors clustered on schools in parentheses.

* $p < .05$, ** $p < .01$, *** $p < .001$. The units of analysis are teachers (Model 1) or individual students in a given subject taught by a given teacher (Models 2-4). The adjusted R^2 statistic is from OLS regressions with dummies for schools (Model 1) or students (Models 2-4). The R^2 (within) statistic is R^2 from regressions on the within-school (Model 1) or within-student (Models 2-4) transformed variables. See Table 2 for numbers of students, classes, teachers and schools.

Table 4**Job Satisfaction and Performance. School and Student Fixed Effect Regressions**

	(1)	(2)	(3)	(4)
	Self-reported contribution	Year's work	Oral exams	Written exams
Teacher job satisfaction	0.310*** (0.044)	0.030* (0.014)	0.038 (0.019)	0.007 (0.013)
Teacher job satisfaction × change		0.013 (0.021)	−0.033 (0.035)	−0.022 (0.022)
Change of teachers		−0.025 (0.028)	−0.011 (0.062)	−0.031 (0.052)
Fixed effects	Schools	Students	Students	Students
Observations	743	38855	15761	13584
Adj. R^2	0.143	0.673	0.504	0.631
R^2 (within)	0.145	0.026	0.022	0.066

Note: The full model can be seen in Table A8 in the Appendix. Robust standard errors clustered on schools in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$. The units of analysis are teachers (Model 1) or individual students in a given subject taught by a given teacher (Models 2-4). The adjusted R^2 statistic is from OLS regressions with dummies for schools (Model 1) or students (Models 2-4). The R^2 (within) statistic is R^2 from regressions on the within-school (Model 1) or within-student (Models 2-4) transformed variables. The numbers of students, classes, teachers and schools are the same as in Table 2, except that the number of teachers is 743 in Models (1) and (2), 551 in (3) and 507 in (4).

ONLINE APPENDIX

Table A1: Measurement Statistics for Self-reported Contributions to Students' Academic Skills. CFA

Questions (English translation in italics)	SFL	SMC (R ²)
Jeg føler, at jeg gør en betydelig uddannelsesmæssig forskel i mine elevers liv. <i>I feel that I am making a significant educational difference in the lives of my students.</i>	0.543	0.295
Hvis jeg virkelig anstrenger mig, kan jeg få selv den mest vanskelige og umotiverede elev til at gøre fremskridt. <i>If I try really hard, I can make progress with even the most difficult and unmotivated students.</i>	0.474	0.224
Jeg har succes med eleverne i min klasse. <i>I am successful with the students in my class.</i>	0.710	0.505

Note: All standardized factor loadings (SFL) and correlations are significant at $p < .001$. Given that the model is “just identified” (Chin 1998), we do not report fit statistics. We also have data on the following items: *I am able to get silence in the class when necessary* and *I usually know how to get through to students*. Models including one or both of these items together with the three items shown above provided highly acceptable fit statistics. We use the three-item model because these three items are seen as the most valid measures of the teachers' perceived contribution to their students' academic skills, and because the AIC statistics indicate that this is the best model. The substantial results are the same regardless of the number of items included in the measure. Cronbach's alpha = 0.55

Table A2: Measurement Statistics for Intrinsic Motivation. CFA

Questions	SFL	SMC (R ²)
Jeg nyder i høj grad mit daglige arbejde. <i>I really enjoy my daily work.</i>	.793	.629
En ret stor del af mine arbejdsopgaver er kedelige. <i>Many of my responsibilities at work are boring. (reversed)</i>	.401	.160
Jeg glæder mig altid til at tage på arbejde om morgenen. <i>I always look forward to going to work in the mornings.</i>	.711	.506
Mit arbejde er meget spændende. <i>My work is very interesting.</i>	.782	.611
Jeg kan godt lide at udføre de fleste af mine arbejdsopgaver. <i>I like performing most of my responsibilities at work.</i>	.703	.494
Jeg føler en stor personlig tilfredsstillelse ved at udføre mit arbejde. <i>I feel a great sense of personal satisfaction when going about my work.</i>	.807	.650
Hvis jeg vandt 5 mio. kr. i lotto, ville jeg stadig blive ved med at udføre mit nuværende arbejde. <i>If I won DKK 5 million in the lottery, I would still keep my present job.</i>	.584	.341

Note: All standardized factor loadings (SFL) and correlations are significant at $p < .001$. RMSEA is 0.078, CFI is 0.971, TLI is 0.956, and SRMR indicator is 0.027. Cronbach's alpha = 0.83.

Table A3: Cronbach's Alphas and Correlations between Public Service Motivation Dimensions

	Cronbach's alpha	2	3	4
1. Self-sacrifice	0.84	0.31	0.12	0.56
2. Compassion	0.80		0.01	0.38
3. Attraction to public policy making	0.57			0.10
4. Commitment to the public interest	0.70			

Table A4: Measurement Statistics for Public Service Motivation Dimensions. CFA

Dimensions and items (English translation in italics)	SFL	SMC (R ²)
Self-sacrifice		
Det er vigtigere for mig at gøre en forskel i forhold til samfundet end at opnå personlig vinding. <i>Making a difference in society means more to me than personal achievements.</i>	0.584	0.341
Jeg mener, at man skal bidrage med mere til samfundet, end man modtager. <i>I feel people should give more back to society than they get from it.</i>	0.546	0.298
Jeg er villig til at risikere at skulle tilsidesætte mine personlige behov for samfundets skyld. <i>I am willing to risk personal loss to help society.</i>	0.858	0.736
Jeg er klar til at lide afsavn for samfundets skyld. <i>I am prepared to make sacrifices for the good of society.</i>	0.834	0.695
Jeg sætter samfundsmæssige forpligtigelser over hensynet til mig selv. <i>I believe in putting duty before self.</i>	0.785	0.616
Compassion		
Jeg bliver følelsesmæssigt berørt, når jeg ser mennesker i nød. <i>It is difficult for me to contain my feelings when I see people in distress.</i>	0.724	0.524
For mig er hensyntagen til andres velfærd meget vigtig. <i>For me, considering the welfare of others is one of the most important values.</i>	0.709	0.502
Jeg bliver meget berørt, når jeg ser andre mennesker blive behandlet uretfærdigt. <i>I get very upset when I see other people being treated unfairly.</i>	0.717	0.514
Jeg føler sympati overfor mindre privilegerede mennesker med problemer. <i>I feel sympathy towards underprivileged persons dealing with problems.</i>	0.674	0.454
Attraction to public policy making		
Jeg forbinder generelt politik med noget positivt. <i>I generally associate politics with something positive.</i>	0.503	0.253
Jeg bryder mig ikke om politiske studehandler. <i>The give and take of public policy making does not appeal to me. (reversed)</i>	0.381	0.146
Jeg har ikke særligt høje tanker om politikere. <i>I do not care much for politicians. (reversed)</i>	0.841	0.707
Commitment to the public interest		
Det er vigtigt for mig, at offentlige ydelser gavner samfundet som helhed. <i>It is important for me that public services contribute to the common good.</i>	0.434	0.189
Jeg så helst, at offentligt ansatte gør det, der er bedst for hele samfundet, selvom det skulle gå ud over mine egne interesser. <i>I would prefer seeing public officials do what is best for the whole community even if it harmed my interests.</i>	0.567	0.321
Det er vigtigt for mig at bidrage til det fælles bedste. <i>It is important for me to contribute to the common good.</i>	0.734	0.539
Det er min borgerpligt at gøre noget, der tjener samfundets bedste. <i>I consider public service my civic duty.</i>	0.752	0.566

Note: SFL = standardized factor loading; SMC = squared multiple correlation coefficients. All standardized factor loadings and correlations are significant at $p < .001$. RMSEA is 0.027 (90% C.I. 0.021–0.032), CFI is 0.954, TLI is 0.944, and SRMR indicator is 0.04. Cronbach's alphas and correlations can be seen in Table A3 above.

Table A5. Summary Statistics (student $\times$ subject observations)

	count	mean	Sd	min	max
<i>Dependent variables</i>					
Self-reported performance	39056	0.021	0.999	−4.000	1.465
Marks for the year's work	39056	−0.002	0.965	−3.464	1.812
Oral exam marks	15858	−0.001	0.998	−3.009	1.559
Written exam marks	13600	0.007	0.940	−3.149	1.955
<i>Perceptual explanatory variables</i>					
Public service motivation	39056	0.027	1.056	−3.832	2.544
Intrinsic motivation	39056	0.028	0.983	−5.538	1.213
Job satisfaction	38855	0.033	0.956	−5.376	1.249
<i>Other teacher characteristics</i>					
Female teacher	39056	0.499	0.500	0.000	1.000
Female student $\times$ female teacher	39056	0.248	0.432	0.000	1.000
Qualifications in subject	39056	0.609	0.488	0.000	1.000
Special teacher education	39056	0.080	0.271	0.000	1.000
No teacher education	39056	0.022	0.148	0.000	1.000
Experience 0-4 years	39056	0.192	0.394	0.000	1.000
Experience 5-9 years	39056	0.267	0.442	0.000	1.000
Change of teachers	39056	0.257	0.437	0.000	1.000
<i>Subject dummies</i>					
Danish	39056	0.115	0.319	0.000	1.000
Math	39056	0.115	0.318	0.000	1.000
English	39056	0.118	0.322	0.000	1.000
History	39056	0.112	0.315	0.000	1.000
Physics/Chemistry	39056	0.105	0.306	0.000	1.000
Biology	39056	0.105	0.306	0.000	1.000
Geography	39056	0.106	0.308	0.000	1.000
Religion	39056	0.109	0.312	0.000	1.000
Social Studies	39056	0.116	0.320	0.000	1.000
<i>Student characteristics</i>					
Female student	39056	0.494	0.500	0.000	1.000
Student is immigrant or child of immigrants	39056	0.075	0.263	0.000	1.000

Table A6: Full Fixed Effects Regression Results with Public Service Motivation as the Key Explanatory

Variable (corresponding to Table 2)

	(1) Self-reported		(2) Marks, year's work		(3) Oral exam marks		(4) Written marks	
Public service motivation	0.194***	(0.039)	0.031**	(0.012)	0.043*	(0.019)	0.029*	(0.012)
Public service m. $\times$ change of teachers			-0.005	(0.019)	-0.003	(0.027)	-0.041	(0.022)
Female teacher	0.158	(0.081)	-0.028	(0.026)	-0.054	(0.037)	-0.006	(0.031)
Female student $\times$ female teacher			0.034	(0.022)	0.104**	(0.034)	0.031	(0.033)
Qualifications in subject			-0.026	(0.019)	0.003	(0.035)	0.005	(0.025)
Special teacher education	-0.185	(0.185)	0.001	(0.049)	0.000	(0.071)	0.040	(0.067)
No teacher education	-0.229	(0.177)	-0.217**	(0.064)	-0.146	(0.124)	-0.159	(0.189)
Experience 0-4 years	0.077	(0.107)	-0.034	(0.027)	-0.069	(0.049)	-0.056	(0.034)
Experience 5-9 years	0.279**	(0.087)	-0.005	(0.030)	0.049	(0.037)	0.028	(0.029)
Change of teachers			-0.026	(0.028)	0.006	(0.063)	-0.024	(0.052)
Female teacher $\times$ change			0.004	(0.038)	0.036	(0.085)	0.036	(0.052)
Fem. student $\times$ female teacher $\times$ change			-0.029	(0.027)	-0.030	(0.068)	-0.089	(0.050)
Qualifications in subject $\times$ change			0.012	(0.031)	-0.083	(0.066)	0.038	(0.051)
Special teacher education $\times$ change			-0.093	(0.065)	0.020	(0.108)	0.092	(0.089)
No teacher education $\times$ change			0.019	(0.078)	0.037	(0.150)	-0.058	(0.226)
Experience 0-4 yrs. $\times$ change			0.019	(0.053)	0.050	(0.102)	-0.008	(0.066)
Experience 5-9 yrs. $\times$ change			0.029	(0.046)	-0.054	(0.070)	-0.151**	(0.057)
Math	-0.085	(0.156)	0.208***	(0.029)			0.203***	(0.032)
English	-0.226	(0.163)	0.115***	(0.028)	0.148***	(0.032)	0.267***	(0.055)
History	-0.097	(0.250)	0.132***	(0.028)	0.120*	(0.053)		
Physics/Chemistry	-0.505*	(0.205)	0.149***	(0.030)	0.129**	(0.039)		
Biology	-0.122	(0.192)	0.124**	(0.037)			0.254***	(0.037)
Geography	-0.412	(0.211)	0.154***	(0.042)			0.286***	(0.033)
Religion	0.025	(0.205)	0.000	(0.031)	0.025	(0.052)		
Social Studies	-0.295	(0.229)	0.113***	(0.029)	0.103*	(0.050)		
Math $\times$ female student			-0.442***	(0.026)			-0.441***	(0.025)
English $\times$ female student			-0.254***	(0.023)	-0.299***	(0.033)	-0.456***	(0.057)
History $\times$ female student			-0.250***	(0.027)	-0.287***	(0.057)		
Physics $\times$ female student			-0.334***	(0.029)	-0.248***	(0.039)		
Biology $\times$ female student			-0.243***	(0.033)			-0.424***	(0.042)
Geography $\times$ female student			-0.389***	(0.034)			-0.599***	(0.035)
Religion $\times$ female student			-0.084**	(0.027)	-0.074	(0.065)		
Social studies $\times$ female student			-0.265***	(0.023)	-0.343***	(0.069)		
Math $\times$ immigrant			0.048	(0.054)			0.057	(0.056)
English $\times$ immigrant			0.151***	(0.037)	0.128*	(0.060)	0.171	(0.130)
History $\times$ immigrant			0.186***	(0.054)	0.112	(0.116)		
Physics $\times$ immigrant			0.151**	(0.045)	-0.048	(0.066)		
Biology $\times$ immigrant			0.069	(0.045)			-0.099	(0.071)
Geography $\times$ immigrant			0.108	(0.059)			-0.056	(0.081)
Religion $\times$ immigrant			0.279***	(0.050)	0.084	(0.081)		
Social Studies $\times$ immigrant			0.176**	(0.052)	-0.178	(0.108)		
Constant	0.044	(0.136)	0.039	(0.024)	0.008	(0.039)	0.007	(0.033)
Fixed effects	Schools		Students		Students		Students	
Observations	747		39056		15858		13600	
R ² (within)	0.095		0.026		0.023		0.068	

Robust standard errors clustered on schools in parentheses: * $p < .05$, ** $p < .01$, *** $p < .001$.

Table A7: Full Fixed Effects Regression Results with Intrinsic Motivation as the Key Explanatory Variable
(corresponding to Table 3)

	(1) Self-reported		(2) Marks, year's work		(3) Oral exam marks		(4) Written marks	
Intrinsic motivation	0.419***	(0.045)	0.027*	(0.012)	0.025	(0.015)	-0.008	(0.013)
Intrinsic motivation × change			0.003	(0.020)	-0.031	(0.026)	-0.010	(0.020)
Female teacher	0.043	(0.080)	-0.033	(0.026)	-0.066	(0.038)	0.001	(0.032)
Female student × female teacher			0.034	(0.022)	0.104**	(0.034)	0.032	(0.033)
Qualifications in subject			-0.024	(0.019)	-0.005	(0.036)	0.009	(0.025)
Special teacher education	-0.161	(0.173)	0.007	(0.049)	0.007	(0.069)	0.026	(0.068)
No teacher education	-0.322	(0.163)	-0.244***	(0.065)	-0.174	(0.129)	-0.176	(0.189)
Experience 0-4 years	0.040	(0.093)	-0.036	(0.027)	-0.074	(0.049)	-0.057	(0.034)
Experience 5-9 years	0.215*	(0.086)	-0.008	(0.031)	0.043	(0.039)	0.032	(0.029)
Change of teachers			-0.023	(0.028)	-0.005	(0.060)	-0.030	(0.052)
Female teacher × change of teachers			0.008	(0.040)	0.049	(0.086)	0.031	(0.049)
Female student × female teacher × change			-0.029	(0.027)	-0.030	(0.069)	-0.090	(0.049)
Qualifications in subject × change			0.006	(0.031)	-0.093	(0.067)	0.038	(0.050)
Special teacher education × change			-0.108	(0.070)	0.016	(0.106)	0.085	(0.091)
No teacher education × change			0.048	(0.082)	0.046	(0.152)	-0.042	(0.207)
Experience 0-4 yrs. × change			0.017	(0.053)	0.073	(0.104)	-0.005	(0.067)
Experience 5-9 yrs. × change			0.032	(0.046)	-0.039	(0.071)	-0.141*	(0.055)
Math	-0.135	(0.148)	0.206***	(0.030)			0.202***	(0.033)
English	-0.307*	(0.151)	0.109***	(0.028)	0.144***	(0.032)	0.265***	(0.055)
History	-0.062	(0.220)	0.131***	(0.027)	0.123*	(0.052)		
Physics/Chemistry	-0.503**	(0.184)	0.144***	(0.030)	0.112**	(0.041)		
Biology	-0.140	(0.179)	0.119**	(0.035)			0.246***	(0.038)
Geography	-0.554**	(0.208)	0.151***	(0.041)			0.283***	(0.032)
Religion	-0.055	(0.185)	-0.005	(0.031)	0.017	(0.053)		
Social Studies	-0.321	(0.200)	0.114***	(0.029)	0.098	(0.053)		
Math × female student			-0.443***	(0.025)			-0.443***	(0.025)
English × female student			-0.254***	(0.023)	-0.300***	(0.033)	-0.460***	(0.057)
History × female student			-0.250***	(0.027)	-0.285***	(0.057)		
Physics × female student			-0.334***	(0.029)	-0.249***	(0.039)		
Biology × female student			-0.244***	(0.033)			-0.423***	(0.042)
Geography × female student			-0.389***	(0.034)			-0.599***	(0.036)
Religion × female student			-0.084**	(0.027)	-0.078	(0.065)		
Social Studies × female student			-0.265***	(0.023)	-0.343***	(0.069)		
Math × immigrant			0.042	(0.053)			0.049	(0.057)
English × immigrant			0.150***	(0.037)	0.125*	(0.060)	0.163	(0.130)
History × immigrant			0.180***	(0.051)	0.104	(0.117)		
Physics × immigrant			0.152**	(0.046)	-0.043	(0.067)		
Biology × immigrant			0.074	(0.046)			-0.109	(0.072)
Geography × immigrant			0.113	(0.058)			-0.060	(0.083)
Religion × immigrant			0.276***	(0.048)	0.069	(0.081)		
Social Studies × immigrant			0.173***	(0.048)	-0.180	(0.110)		
Constant	0.162	(0.131)	0.045	(0.024)	0.031	(0.042)	0.004	(0.033)
Fixed effects	Schools		Students		Students		Students	
Observations	747		39056		15858		13600	
R ² (within)	0.232		0.025		0.021		0.067	

Robust standard errors clustered on schools in parentheses: * $p < .05$, ** $p < .01$, *** $p < .001$.

Table A8: Full Fixed Effects Regression Results with Job Satisfaction as the Key Explanatory Variable
(corresponding to Table 4)

	(1) Self-reported		(2) Marks, year's work		(3) Oral exam marks		(4) Written marks	
Job satisfaction	0.310***	(0.044)	0.030*	(0.014)	0.038	(0.019)	0.007	(0.013)
Job satisfaction × change			0.013	(0.021)	-0.033	(0.035)	-0.022	(0.022)
Female teacher	0.109	(0.088)	-0.031	(0.027)	-0.078*	(0.039)	-0.003	(0.032)
Female student × female teacher			0.034	(0.022)	0.101**	(0.034)	0.033	(0.033)
Qualifications in subject			-0.023	(0.019)	-0.008	(0.036)	0.008	(0.025)
Special teacher education	-0.161	(0.176)	0.007	(0.050)	0.005	(0.069)	0.024	(0.069)
No teacher education	-0.413*	(0.157)	-0.239***	(0.065)	-0.183	(0.118)	-0.174	(0.219)
Experience 0-4 years	0.094	(0.097)	-0.033	(0.027)	-0.056	(0.048)	-0.056	(0.034)
Experience 5-9 years	0.250*	(0.095)	-0.008	(0.032)	0.052	(0.040)	0.029	(0.030)
Change of teachers			-0.025	(0.028)	-0.011	(0.062)	-0.031	(0.052)
Female teacher × change			0.006	(0.038)	0.062	(0.090)	0.032	(0.049)
Female student × female teacher × change			-0.029	(0.027)	-0.020	(0.070)	-0.090	(0.050)
Qualifications in subject × change			0.003	(0.031)	-0.091	(0.067)	0.040	(0.050)
Special teacher education × change			-0.102	(0.068)	0.011	(0.104)	0.087	(0.092)
No teacher education × change			0.022	(0.081)	0.034	(0.145)	-0.040	(0.240)
Experience 0-4 yrs. × change			0.019	(0.053)	0.055	(0.100)	-0.004	(0.067)
Experience 5-9 yrs. × change			0.032	(0.045)	-0.044	(0.071)	-0.137*	(0.056)
Math	-0.108	(0.153)	0.206***	(0.030)			0.202***	(0.033)
English	-0.289	(0.155)	0.111***	(0.028)	0.141***	(0.031)	0.263***	(0.054)
History	-0.139	(0.247)	0.130***	(0.027)	0.116*	(0.052)		
Physics/Chemistry	-0.536*	(0.206)	0.144***	(0.030)	0.113**	(0.041)		
Biology	-0.180	(0.186)	0.120**	(0.036)			0.249***	(0.038)
Geography	-0.552*	(0.215)	0.149***	(0.041)			0.280***	(0.033)
Religion	-0.098	(0.196)	-0.008	(0.031)	0.011	(0.053)		
Social Studies	-0.383	(0.206)	0.111***	(0.029)	0.098	(0.053)		
Math × female student			-0.443***	(0.025)			-0.442***	(0.025)
English × female student			-0.254***	(0.024)	-0.301***	(0.034)	-0.458***	(0.057)
History × female student			-0.251***	(0.027)	-0.281***	(0.057)		
Physics × female student			-0.332***	(0.029)	-0.248***	(0.038)		
Biology × female student			-0.244***	(0.033)			-0.423***	(0.042)
Geography × female student			-0.390***	(0.034)			-0.598***	(0.036)
Religion × female student			-0.084**	(0.027)	-0.077	(0.065)		
Social Studies × female student			-0.266***	(0.023)	-0.336***	(0.069)		
Math × immigrant			0.037	(0.053)			0.049	(0.056)
English × immigrant			0.148***	(0.037)	0.126*	(0.060)	0.164	(0.130)
History × immigrant			0.178***	(0.052)	0.102	(0.115)		
Physics × immigrant			0.144**	(0.045)	-0.047	(0.066)		
Biology × immigrant			0.069	(0.046)			-0.105	(0.072)
Geography × immigrant			0.113	(0.060)			-0.057	(0.082)
Religion × immigrant			0.277***	(0.049)	0.069	(0.082)		
Social Studies × immigrant			0.170***	(0.050)	-0.202	(0.109)		
Constant	0.134	(0.138)	0.044	(0.024)	0.037	(0.042)	0.007	(0.033)
Fixed effects	Schools		Students		Students		Students	
Observations	743		38855		15761		13584	
R ² (within)	0.145		0.026		0.022		0.066	

Robust standard errors clustered on schools in parentheses: * $p < .05$, ** $p < .01$, *** $p < .001$.

Table A9. Robustness check: Public service motivation and performance. Self-reported performance and marks for the year's work regressions using observations from the two examination samples

	(1) Self-reported Oral exam sample	(2) Year's work Oral exam sample	(3) Self-reported Written exam sample	(4) Year's work Written exam sample
Public service motivation	0.228*** (0.048)	0.030 (0.017)	0.209*** (0.052)	0.043** (0.015)
Public service motivation $\times$ change		-0.007 (0.027)		-0.073* (0.034)
Change of teachers		-0.001 (0.053)		-0.152 (0.078)
Fixed effects	Schools	Students	School	Students
Observations	554	15858	508	13600
R^2 (within)	0.123	0.026	0.114	0.062

Robust standard errors clustered on schools in parentheses: * $p < .05$, ** $p < .01$, *** $p < .001$. Models 1 and 3 correspond to Model 1 in Table 2 (and Table A6) except for the restricted samples. Similarly, Models 2 and 4 correspond to Model 2 in Table 2 (and Table A6).

Table A10. Robustness check: Intrinsic motivation and performance. Self-reported performance and marks for the year's work regressions using observations from the two examination samples

	(1) Self-reported Oral exam sample	(2) Year's work Oral exam sample	(3) Self-reported Written exam sample	(4) Year's work Written exam sample
Intrinsic motivation	0.443*** (0.058)	0.040* (0.015)	0.390*** (0.045)	0.017 (0.014)
Intrinsic motivation $\times$ change		-0.022 (0.027)		-0.055 (0.032)
Change of teachers		-0.011 (0.053)		-0.175* (0.076)
Fixed effects	Schools	Students	Schools	Students
Observations	554	15858	508	13600
R^2	0.256	0.027	0.222	0.058

Robust standard errors clustered on schools in parentheses: * $p < .05$, ** $p < .01$, *** $p < .001$. Models 1 and 3 correspond to Model 1 in Table 3 (and Table A7) except for the restricted samples. Similarly, Models 2 and 4 correspond to Model 2 in Table 3 (and Table A7).

Table A11. Robustness check: Job satisfaction and performance. Self-reported performance and marks for the year's work regressions using observations from the two examination samples

	(1)	(2)	(3)	(4)
	Self-reported Oral exam sample	Year's work	Self-reported Written exam sample	Year's work
Job satisfaction	0.344*** (0.058)	0.034* (0.017)	0.268*** (0.043)	0.018 (0.014)
Job satisfaction $\times$ change		0.011 (0.032)		-0.049 (0.034)
Change of teachers		-0.002 (0.053)		-0.169* (0.077)
Fixed effects	Schools	Students	Schools	Students
Observations	551	15761	507	13584
R^2	0.169	0.026	0.140	0.058

Robust standard errors clustered on schools in parentheses: * $p < .05$, ** $p < .01$, *** $p < .001$. Models 1 and 3 correspond to Model 1 in Table 4 (and Table A8) except for the restricted samples. Similarly, Models 2 and 4 correspond to Model 2 in Table 4 (and Table A8).

ⁱ The Texas education system uses a standardized test called the Texas Assessment of Knowledge and Skills (TAKS), a criterion-based test given in elementary grades and as an exit exam at high school level. TAKS is a high-stakes test, and students must pass this test to receive a diploma from a Texas high school. The specific performance indicator is the percentage of students who pass all TAKS tests (math, reading, writing, social sciences, etc.) using what is termed the accountability sample. This is the official criterion used to evaluate schools and school districts in the state and constitutes a major part of the annual grades that the state assigns to schools (Meier and O'Toole 2012, 8).

ⁱⁱ The translated wording was: "All in all, how satisfied or dissatisfied are you with your current job on a scale from 0 (very dissatisfied) to 10 (very satisfied)?" The Danish version was: "*Alt taget i betragtning, hvor tilfreds eller utilfreds er du med dit nuværende job på en skala fra 0 (meget utilfreds) til 10 (meget tilfreds)?*"